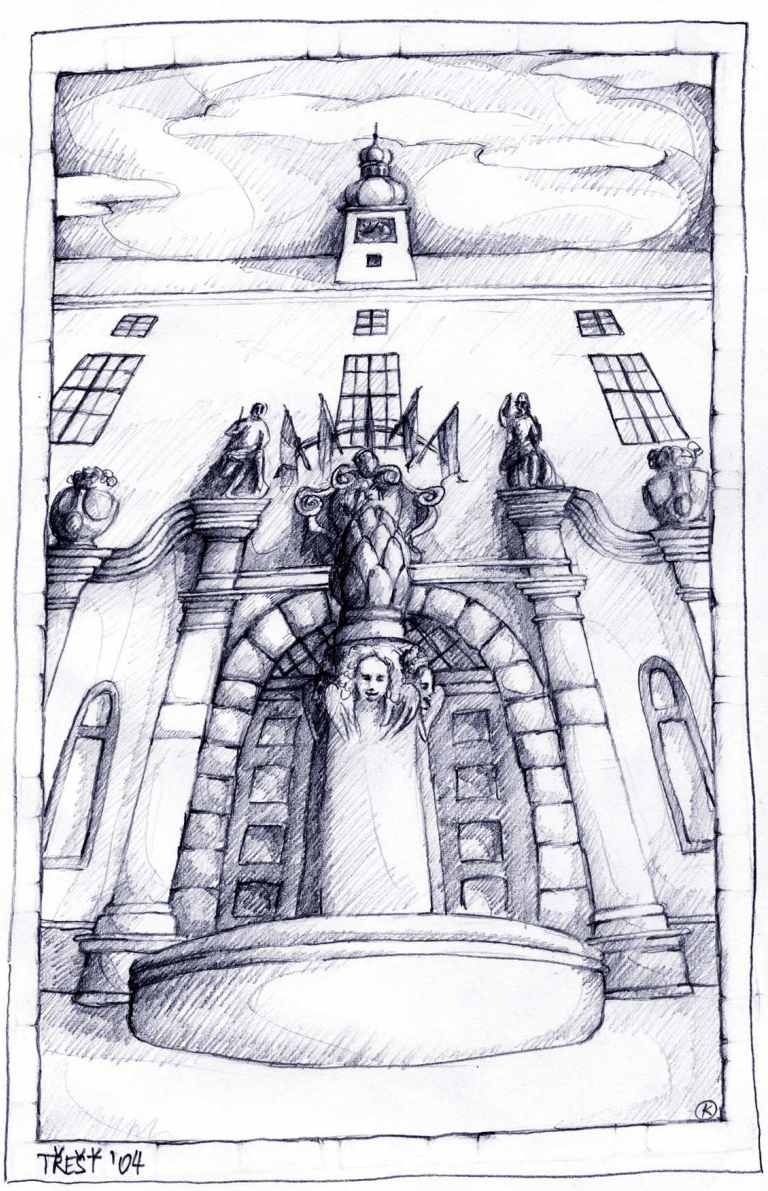
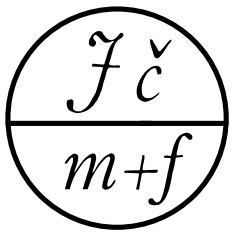

TŘEST '04

ROBUST 2004

Sborník prací 13. letní školy JČMF ROBUST 2004
uspořádané Jednotou českých matematiků a fyziků
za podpory KPMS MFF UK
a České statistické společnosti
ve dnech 7. – 11. června 2004 v Třešti



Praha 2004

ISBN 80-7015-972-3
JČMF 57-552-04

ROBUST 2004 – PÁR SLOV ÚVODEM

Ve dnech 7. – 11. června 2004 se v prostorách školícího střediska AV ČR v Třešti uskutečnila již třináctá letní škola JČMF ROBUST 2004. Tato akce byla připravema skupinou pro výpočetní statistiku MVS JČMF za podpory ČStS, KPMS MFF UK, MÚ AV ČR a ÚTM FSI ČVUT. Akce se zúčastnilo 84 účastníků z ČR, Slovenska, Belgie a USA.

Tak jako v minulosti, i ROBUST 2004 byl věnován vybraným trendům matematické i aplikované statistiky, teorie pravděpodobnosti a analýzy dat. Pozvání k přednesení přehledných přednášek přijali:

- Prof. RNDr. Tomáš Cipra, DrSc., MFF UK Praha, *Zajištění v pojišťovnictví a jeho matematické aspekty*.
- Doc. RNDr. Daniela Jarušková, CSc., FSV ČVUT Praha, *Extrémy gaussovských posloupností a procesů*.
- Doc. RNDr. Jan Picek, CSc, TUL Liberec *Testy a odhady Paretova indexu*.
- Doc. RNDr. Zuzana Prášková, CSc, MFF UK Praha, *Metoda bootstrap – 25 let*.

Celkem bylo předneseno 62 přednášek, z toho 28 přednesli postgraduální studenti. Již poněkolikáté jich bylo, k naší velké radosti, tolik, že jsme mohli z jejich vystoupení nejenom vytvořit dva samostatné půldenní bloky, ale ocenit též nejlepší vystoupení/dosažené výsledky. Komise pod předsednictvím prof. G. Wimmera z MFF UKo v Bratislavě vyhodnotila vystoupení přednášejících a navrhla firmě *Elkan* k ocenění za nejlepší prezentovanou práci Ing. M. Omelky.

Mnoho času též bylo věnováno diskusím. Pondělní večer byl věnován volné diskusi o výuce statistiky a pravděpodobnosti, především pro informatiky a informatiku. Během úterního večera vystoupili zástupci firem *Elkan* a *TriloByte*, kteří předvedli nejnovější verze programů *MATHEMATICA* a *S+*. Vedle odborných diskusí se též konaly diskuse volnější, a to ať již během středečního výletu do Telče a Brnice či návštěvy památníku Franze Kafky a muzea třešťských betlémů. I počasí nám tentokrát vyšlo, snad jenom s vyjímkou deště během čtvrtečního večerního opekání buřtů v zámecké zahradě.

Z přednesených 62 přednášek naleznete 51 v přiloženém sborníku. Jinými slovy to znamená, že připěli prakticky všichni přednášející. Chtěli bychom tímto poděkovat nejenom jim, ale též všem těm, kteří články recenzovali. Budete se možná divit, že čtyři články nejsou zařazený podle abecedy nýbrž

na konci sborníku. Vemte prosím na vědomí, že se nejedná o „chybu“ editorů sborníku, nýbrž o **neoprávněnou víru autorů** v to, že pokud „něco“ pošlou do „sítě“, tak jejich „zásilka“ dorazí k příjemci. Dovolujeme si touto cestou všem čtenářům připomenout, že odeslání mailu v době spamů a antispamových ochran nikterak neazaručuje, že zpráva skutečně dorazí...

V Praze 24. prosince 2004 Jaromír Antoch a Gejza Dohnal

OBSAH

KONFIDENČNÉ INTERVALY PRE VARIANČNÝ KOMPONENT V MODELI S DVOMI VARIANČNÝMI KOMPONENTAMI

Barbora Arendacká

Kľúčové slová: Konfidenčné intervaly, variančný komponent, zovšeobecnené testovacie premenné, zovšeobecnené p-hodnoty.

Abstrakt: Konštrukcia konfidenčného intervalu, resp. testovanie hypotézy o neznámom variančnom komponente σ_1^2 v triede rozdelení $N_t(0, \sigma_1^2 W + \sigma^2 I_t)$ sú úlohy s rušivým parametrom σ^2, ktorých riešenie klasickými metódami je známe len v špeciálnych prípadoch. V článku uvedieme ako je možné tieto úlohy riešiť pomocou zovšeobecnených p-hodnôt (pojem zaviedli Tsui a Weerahandi [9]) a ukážeme ako môžme voliť zovšeobecnené testovacie premenné v prípade, keď matica W má viac ako dve rôzne vlastné hodnoty. (Na nejednoznačnosť použiteľnej testovacej premennej v tejto situácii upozornili Zhou a Mathew [11].)

1 Úvod

Uvažujme normálne rozdelený n-rozmerný vektor pozorovaní y,

$$y \sim N_n(X\beta, \sigma_1^2 ZZ^T + \sigma^2 I_n), \tag{1}$$

kde β a $(\sigma_1^2, \sigma^2)^T$ sú vektory neznámych parametrov, $\sigma_1^2 \geq 0$, $\sigma^2 > 0$ a $\mathcal{R}(Z) \not\subseteq \mathcal{R}(X)$, kde $\mathcal{R}(A)$ označuje priestor generovaný stĺpcami matice A. Úloha testovať hypotézu

$$H_0 : \sigma_1^2 \leq \sigma_{10}^2 \text{ vs. } H_1 : \sigma_1^2 > \sigma_{10}^2 \tag{2}$$

pre ľubovoľné nezáporné σ_{10}^2 je potom invariantná vzhľadom na posunutie v strednej hodnote, t.j. na transformáciu $y \longrightarrow y + X\beta$, $\beta \in R^{p_1}$ a model (1) môžme zredukovať skonštruovaním maximálneho invariantu $z = B^T y$, kde, ak hodnosť X je p, B^T je $(n-p) \times n$ matica taká, že $B^T B = I_{n-p}$ a $BB^T = M = I - X(X^T X)^- X^T$. Vektor z má potom normálne rozdelenie $N_{n-p}(0, \sigma_1^2 W + \sigma^2 I_{n-p})$, kde $W = B^T ZZ^T B$.

Olsen, Seely a Birkes [6] ukázali, že minimálne postačujúce štatistiky pre triedu rozdelení vektora z majú tvar $U_i = z^T E_i E_i^T z$, $i = 1, \ldots, r$, sú navzájom nezávislé a platí $U_i \sim (\lambda_i \sigma_1^2 + \sigma^2) \chi_{\nu_i}^2$, $i = 1, \ldots, r$, kde $\lambda_1 > \lambda_2 > \cdots > \lambda_r \geq 0$ sú navzájom rôzne vlastné hodnoty matice W, $\nu_1, \nu_2, \ldots, \nu_r$ ich násobnosti a $E_i, i = 1, \ldots, r$ je $(n-p) \times \nu_i$ matica, ktorej stĺpce tvoria ortonormálne vlastné vektory prislúchajúce k vlastnej hodnote λ_i, teda $W = \sum_{i=1}^r \lambda_i E_i E_i^T$.

V ďalšom budeme predpokladať, že matica W je singulárna, t.j. $\lambda_r = 0$. Táto podmienka nie je splnená len v situáciách, keď máme k dipozícii iba minimálny počet pozorovaní.

Prehľad presných testov na hladine významnosti α pre (2) možno nájsť v článku [1], pozri tiež [8]. Testy sú známe len v prípade, keď $\sigma_{10}^2 = 0$, vtedy je nulová hypotéza v (2) ekvivalentná s hypotézou $H_0 : \theta = 0$, $\theta = \frac{\sigma_1^2}{\sigma^2}$. Testovacie štatistiky je potom možné použiť na konštrukciu konfidenčných intervalov pre θ. Dve z týchto štatistík sú:

- $T_W = \frac{\sum_{i=1}^{r-1} U_i}{U_r}$ – Waldov test [7]
- $T_{GM} = \frac{\sum_{i=1}^{r-1} \lambda_i U_i}{U_r}$ - modifikovaný Waldov test [2]

Postup pre konštrukciu konfidenčných intervalov priamo pre variančný komponent σ_1^2 navrhol Michalski [4] založiť na bayesovskom invariantnom kvadratickom nevychýlenom (BIQU) odhade $\hat{\sigma}_1^2(u, v)$ pre σ_1^2 vzhľadom na apriórne rozdelenie také, že $E(\sigma_1^2) = u$, $E(\sigma^2) = 1$, $Var(\sigma_1^2) = v$, $Var(\sigma^2) = 0$, $u, v \geq 0$, resp. na odhade $\hat{\sigma}_1^2(\infty)$, ktorý dostaneme ako limitu BIQU odhadov pre $v \to \infty$. V oboch prípadoch sa dá príslušný odhad vyjadriť ako lineárna kombinácia postačujúcich štatistík U_i: $\hat{\sigma}_1^2 = \sum_{i=1}^{r} c_i^B U_i$, kde konštanty c_i^B majú známy tvar. Pri tomto postupe sa však počítajú konfidenčné intervaly pre každú hodnotu θ zvlášť a ako výsledný interval navrhuje Michalski zvoliť interval s najväčšou dĺžkou ako určitú ochranu pred najhoršou možnou situáciou.

Iný prístup, ktorý umožňuje konštruovať konfidenčné intervaly priamo pre σ_1^2, resp. testovať (2) aj pre nenulové σ_{10}^2, sa zakladá na zovšeobecnených p-hodnotách a zovšeobecnených testovacích premenných. Tieto pojmy zaviedli Tsui a Weerahandi [9] a ďalej rozpracoval Weerahandi [10]. V nasledujúcej časti aplikujeme tento prístup na náš problém.

2 Riešenie pomocou zovšeobecnených testovacích premenných

V úlohách s rušivými parametrami pri testovaní hypotéz pomocou p-hodnôt je často nemožné alebo nie práve jednoduché nájsť testovaciu štatistiku vhodnú na definovanie extrémnej oblasti, ktorej distribúcia nezávisí od rušivých parametrov. Tento problém je ale riešiteľný, ak namiesto jednej "globálnej" testovacej štatistiky pre daný problém, budeme hľadať vhodnú testovaciu štatistiku pre každú realizáciu dát osobitne. Ide teda o to, nájsť funkciu závislú okrem náhodného výberu a parametrov (ako testovacia štatistika) aj na napozorovaných dátach, takú, že pre každé pevné napozorované hodnoty dát je nájdená funkcia vhodnou testovacou štatistikou na definovanie extrémnej oblasti. V našom prípade pôjde o funkciu $T(U, u, \sigma_1^2, \sigma^2)$, kde $U = (U_1, \ldots, U_r)$ je vektor postačujúcich štatistík a u je jeho napozorovaná hodnota, s vlastnosťami:

1. napozorovaná hodnota $t_{obs} = T(u, u, \sigma_1^2, \sigma^2)$ nezávisí na neznámych parametroch,
2. pre pevné σ_1^2, distribúcia T nezávisí na σ^2 pre každé u,
3. pre pevné u a σ^2 je $P(T \leq t, \sigma_1^2)$ nerastúca funkcia σ_1^2 pre každé t.

T potom nazývame *zovšeobecnenou testovacou premennou stochasticky rastúcou v* σ_1^2 (ďalej len zovšeobecnená testovacia premenná). Vďaka vlastnosti 3 T usporadúva výberový priestor a zovšeobecnená extrémna oblasť pre testovanie (2) založená na T má tvar

$$C(u, \sigma_1^2, \sigma^2) = \{v; T(v, u, \sigma_1^2, \sigma^2) \geq T(u, u, \sigma_1^2, \sigma^2)\},$$

k nej prislúchajúca zovšeobecnená p-hodnota je

$$p(u) = \sup_{\sigma_1^2 \leq \sigma_{10}^2} P(U \in C(u, \sigma_1^2, \sigma^2) |\ \sigma_1^2) =$$

$$= \sup_{\sigma_1^2 \leq \sigma_{10}^2} P(T(U, u, \sigma_1^2, \sigma^2) \geq t_{obs} |\ \sigma_1^2) = P(T(U, u, \sigma_{10}^2, \sigma^2) \geq t_{obs} |\ \sigma_{10}^2).$$

Vďaka vlastnosti 2 testovacej premennej T je p-hodnota vyčísliteľná.
Podobne ako klasická p-hodnota, aj zovšeobecnená p-hodnota slúži ako miera súhlasu, resp. nesúhlasu dát s nulovou hypotézou. Jej vysoké hodnoty hovoria v prospech nulovej, nízke v prospech alternatívnej hypotézy, a teda testy založené na zovšeobecnených p-hodnotách zamietajú nulovú hypotézu pre malé hodnoty $p(u)$.

K zovšeobecnenej extrémnej oblasti môžme zadefinovať na *dátach založenú silofunkciu*

$$\pi(u, \sigma_1^2) = P(U \in C(u, \sigma_1^2, \sigma^2) | \sigma_1^2),$$

pre ktorú platí:

a) $\pi(u, \sigma_{10}^2) = p(u)$,
b) ak je $T(U, u, \sigma_1^2, \sigma^2)$ spojitá náhodná premenná pre každé u, $\pi(u, \sigma_1^2) = 1 - F_{T(U,u,\sigma_1^2,\sigma^2)}(t_{obs} |\ \sigma_1^2)$, a teda pre pevné u je $\pi(U, u, \sigma_1^2) = 1 - F_T(T(U, u, \sigma_1^2, \sigma^2) |\ \sigma_1^2)$ náhodná premenná rovnomerne rozdelená na $(0, 1)$,
c) pre každé pevné u je $\pi(u, \sigma_1^2)$ neklesajúcou funkciou σ_1^2.

Vďaka b) a c) môžme pomocou silofunkcie π skonštruovať konfidenčný interval pre σ_1^2. Pre $\gamma_1, \gamma_2 \in (0, 1)$ také, že $\gamma_2 - \gamma_1 = 1 - \alpha$ a dané napozorované u platí:

$$P(\gamma_1 \leq \pi(U, u, \sigma_1^2) \leq \gamma_2) = 1 - \alpha \tag{3}$$

a hranice obojstranného $(1 - \alpha)100\%$-ného zovšeobecneného konfidenčného intervalu $(\underline{\sigma_1^2},\ \overline{\sigma_1^2})$ pre σ_1^2 získame riešením rovníc

$$\pi(u, \underline{\sigma_1^2}) = \gamma_1 \text{ a } \pi(u, \overline{\sigma_1^2}) = \gamma_2. \tag{4}$$

Keďže pravdepodobnostné tvrdenie (3) platí podmienene pri danom, pevnom u, zovšeobecnený konfidenčný interval si zachováva frekvenčné vlastnosti konfidenčného intervalu podmienene, pri danom u. Priamo nevyplýva, že jeho nepodmienená pravdepodobnosť pokrytia je tiež $1-\alpha$. Zachovanie konfidenčnej úrovne sa overuje simulačne, keďže priamy výpočet skutočnej pravdepodobnosti pokrytia zovšeobecneného konfidenčného intervalu je komplikovaný.

Konkrétne testovacie premenné

Označme $V_1 = \frac{U_1}{\sigma^2+\sigma_1^2\lambda_1} \sim \chi^2_{\nu_1}, \ldots, V_{r-1} = \frac{U_{r-1}}{\sigma^2+\sigma_1^2\lambda_{r-1}} \sim \chi^2_{\nu_{r-1}}$ a $V_r = \frac{U_r}{\sigma^2} \sim \chi^2_{\nu_r}$ a uvažujme najprv prípad $r = 2$. Waldova štatistika má v tomto prípade tvar $T_W = U_1/U_2$. Napozorovaná hodnota tohto podielu, u_1/u_2, nezávisí na neznámych parametroch, ale jeho distribúcia závisí na rušivom parametri σ^2, keďže:

$$\frac{U_1}{U_2} = \frac{V_1(\lambda_1\sigma_1^2+\sigma^2)}{V_2\sigma^2} = \frac{V_1}{V_2}\left(\lambda_1\frac{\sigma_1^2}{\sigma^2}+1\right).$$

Závislosť na parametri σ^2 odstránime prenásobením jeho prevrátenej hodnoty v prvom výraze v zátvorke náhodnou premennou U_2 a predelením napozorovanou hodnotou u_2. Je zrejmé, že napozorovaná hodnota celého výrazu sa nezmení, a teda zostane nezávislá na neznámych parametroch, a zároveň distribúcia vzniknutého výrazu bude spĺňať vlastnosť 2 zovšeobecnenej testovacej premennej. Dostaneme

$$\begin{aligned} T(U_1,U_2,u_1,u_2,\sigma_1^2,\sigma^2) &= \frac{V_1}{V_2}\left(\lambda_1\frac{\sigma_1^2U_2}{\sigma^2u_2}+1\right) = \\ &= \frac{V_1}{V_2}\left(\lambda_1\frac{\sigma_1^2V_2}{u_2}+1\right) = V_1\left(\frac{1}{V_2}+\frac{\lambda_1\sigma_1^2}{u_2}\right). \end{aligned}$$

Je ľahké overiť, že T spĺňa aj vlastnosť 3 zovšeobecnenej testovacej premennej. Podobným postupom dostaneme v prípade, keď $r > 2$, testovaciu premennú

$$T_{c_1,\ldots,c_{r-1}} = \sum_{i=1}^{r-1} c_iV_i\left(\frac{1}{V_r}+\frac{\lambda_i\sigma_1^2}{u_r}\right),$$

kde c_i, $i = 1,\ldots,r-1$ sú ľubovoľné kladné reálne čísla. Kým v prípade $r = 2$ sa dá ukázať, že všetky zovšeobecnené testy založené na U_1, U_2, je možné založiť na T, v prípade, keď počet vlastných hodnôt matice W je vyšší ako 2, sa táto jednoznačnosť stráca. Práve na príklade testovacej premennej $T_{c_1,\ldots,c_{r-1}}$ na to poukázali Zhou a Mathew [11]. Ako voliť konštanty c_i zostáva stále nevyriešené. Istou inšpiráciou však môžu byť testovacie štatistiky na testovanie nulovosti σ_1^2 uvedené v časti 1.

Uvažujme testovacie premenné

$$\begin{aligned} T_1 &= T_{1,\ldots,1} = \sum_{i=1}^{r-1} V_i \left(\frac{1}{V_r} + \frac{\lambda_i \sigma_1^2}{u_r} \right), \\ T_\lambda &= T_{\lambda_1,\ldots,\lambda_{r-1}} = \sum_{i=1}^{r-1} \lambda_i V_i \left(\frac{1}{V_r} + \frac{\lambda_i \sigma_1^2}{u_r} \right), \\ T_{1/\lambda} &= T_{1/\lambda_1,\ldots,1/\lambda_{r-1}} = \sum_{i=1}^{r-1} \frac{1}{\lambda_i} V_i \left(\frac{1}{V_r} + \frac{\lambda_i \sigma_1^2}{u_r} \right). \end{aligned}$$

Je zrejmé, že pri testovaní (2) so $\sigma_{10}^2 = 0$ je za platnosti H_0 T_1 zhodná s Waldovou štatistikou T_W, resp. T_λ s Gnotovou - Michalského štatistikou T_{GM}. Ďalej, výjduc z odhadu $\hat{\sigma}_1^2(\infty) = \sum_{i=1}^r c_i^B U_i$, $c_i^B = \frac{1}{\lambda_i \sum_{i=1}^{r-1} \nu_i}, i = 1, \ldots, r$ dostaneme $T_{1/\lambda}$.
Totiž, $\hat{\sigma}_1^2(\infty) = \sum_{i=1}^r c_i^B V_i(\sigma^2 + \lambda_i \sigma_1^2)$ a odstránením závislosti posledného výrazu na parametri σ^2 prenásobením tohto parametra napozorovanou hodnotou u_r a jeho predelením náhodnou premennou U_r, dostaneme zovšeobecnenú testovaciu premennú

$$\sum_{i=1}^{r-1} c_i^B V_i \left(\frac{u_r}{V_r} + \lambda_i \sigma_1^2 \right) + c_r^B u_r$$

s napozorovanou hodnotou

$$\sum_{i=1}^{r-1} c_i^B u_i + c_r^B u_r,$$

od ktorej odvodená na dátach založená silofunkcia $\pi(u_1, \ldots, u_r, \sigma_1^2)$ pre testovanie (2) má tvar:

$$\begin{aligned} \pi &= P\left(\sum_{i=1}^{r-1} c_i^B V_i \left(\frac{u_r}{V_r} + \lambda_i \sigma_1^2 \right) + c_r^B u_r \geq \sum_{i=1}^{r-1} c_i^B u_i + c_r^B u_r \mid \sigma_1^2 \right) = \\ &= P\left(\sum_{i=1}^{r-1} \frac{1}{\lambda_i \sum_{i=1}^{r-1} \nu_i} V_i \left(\frac{u_r}{V_r} + \lambda_i \sigma_1^2 \right) \geq \sum_{i=1}^{r-1} \frac{1}{\lambda_i \sum_{i=1}^{r-1} \nu_i} u_i \mid \sigma_1^2 \right) = \\ &= P\left(\sum_{i=1}^{r-1} \frac{1}{\lambda_i} V_i \left(\frac{1}{V_r} + \frac{\lambda_i \sigma_1^2}{u_r} \right) \geq \sum_{i=1}^{r-1} \frac{u_i}{\lambda_i u_r} \mid \sigma_1^2 \right) = \\ &= P(T_{1/\lambda} \geq t_{1/\lambda} \mid \sigma_1^2) = \pi_{T_{1/\lambda}}(u_1, \ldots, u_r, \sigma_1^2). \end{aligned}$$

Nejednoznačnosť, ktorá sa objavuje pri testovaní (2) sa však neobmedzuje len na problém voľby konštánt c_i v $T_{c_1,..,c_{r-1}}$. Dajú sa skonštruovať ďalšie testovacie premenné, ktoré sa nedajú odvodiť z premennej $T_{c_1,..,c_{r-1}}$ pre žiadnu voľbu c_i. Napríklad,

$$T_2 = \frac{\sum_{i=1}^{r-1} V_i}{V_r} - \sum_{i=1}^{r-1} \frac{u_i}{u_r + \lambda_i \sigma_1^2 V_r}.$$

Keď testujeme (2) so $\sigma_{10}^2 = 0$ bude zovšeobecnená p-hodnota spočítaná pomocou T_2 zhodná s klasickou p-hodnotou spočítanou pomocou Waldovej štatistiky T_W. V prípade nevyváženého modelu jednoduchého triedenia s náhodným efektom sa testovacia premenná T_2 zhoduje s testovacou premennou navrhnutou Weerahandim [10] pre tento model.

Príťažlivou vlastnosťou T_2 je jednoduchosť výpočtu jej na dátach založenej silofunkcie. Tá sa dá vyjadriť ako jednorozmerný integrál, pod ktorým vystupuje len hustota a distribučná funkcia χ^2 rozdelenia.
Označme $W = \sum_{i=1}^{r-1} V_i$, $W \sim \chi_\nu^2$, $\nu = \sum_{i=1}^{r-1} \nu_i$.
Keďže napozorovaná hodnota T_2 je 0,

$$\pi_{T_2}(u_1, \ldots, u_r, \sigma_1^2) = P\left(\sum_{i=1}^{r-1} V_i/V_r - \sum_{i=1}^{r-1} \frac{u_i}{u_r + \lambda_i \sigma_1^2 V_r} \geq 0 \mid \sigma_1^2 \right) =$$

$$= P\left(W \geq \sum_{i=1}^{r-1} \frac{u_i V_r}{u + \lambda_i \sigma_1^2 V_r} \mid \sigma_1^2 \right) = 1 - \int_0^\infty F_W \left(\sum_{i=1}^{r-1} \frac{u_i v}{u + \lambda_i \sigma_1^2 v} \right) f_{V_r}(v) dv,$$

kde F_W, f_{V_r} označujú distribučnú funkciu a hustotu príslušnej náhodnej premennej.

Naproti tomu, $T_{c_1,..,c_{r-1}}$ je pre pevné σ_1^2 rozdelená ako lineárna kombinácia s náhodnými koeficientami nezávislých χ^2 rozdelených náhodných premenných. Pri výpočte hodnôt jej na dátach založenej silofunkcie možno využiť nasledujúci vzťah:

$$\pi_{T_{c_1,\ldots,c_{r-1}}}(u_1, \ldots, u_r, \sigma_1^2) = P\left(\sum_{i=1}^{r-1} c_i V_i \left(\frac{1}{V_r} + \frac{\lambda_i \sigma_1^2}{u_r} \right) \geq \sum_{i=1}^{r-1} \frac{c_i u_i}{u_r} \mid \sigma_1^2 \right) =$$

$$= \int_0^\infty \left(1 - F_v \left(\sum_{i=1}^{r-1} \frac{c_i u_i}{u_r} \right) \right) f_{V_r}(v) dv,$$

kde F_v je distribučná funkcia lineárnej kombinácie nezávislých χ^2 rozdelených náhodných premenných $\sum_{i=1}^{r-1} c_i (\frac{1}{v} + \frac{\lambda_i \sigma_1^2}{u_r}) V_i$ a jej hodnoty sa dajú spočítať Imhofovým algoritmom [3], a f_{V_r} je hustota $\chi_{\nu_r}^2$ rozdelenia.

Numerické výsledky

Podľa vyššie uvedených vzťahov je možné pre konkrétne dáta a zvolené σ_1^2 spočítať hodnoty na dátach založených silofunkcií pre jednotlivé zovšeobecnené testovacie premenné. Hranice obojstranného konfidenčného intervalu je podobne možné získať numerickým riešením (4) pre vhodne zvolené γ_1, γ_2,

napríklad, v prípade obojstranného 95%-ného intervalu môžme zvoliť $\gamma_1 = 0.025$ a $\gamma_2 = 0.975$. Ako už bolo spomenuté, skutočnú konfidenčnú úroveň takto skonštruovaných intervalov je vhodné overiť simulačne. Z tohto pohľadu sme jednotlivé testovacie premenné skúmali (s uspokojivým výsledkom) na konkrétnych prípadoch modelu (1). Tu uvedieme len jeden.
Ide o model (1) s $X = (1_{30}\ s)$, $Z = (A\ r)$, kde 1_n označuje $n \times 1$ vektor jednotiek, s je 30×1 vektor s i-tou zložkou $s_i = -3 + 6 * (i-1)/29$, $A = I_5 \otimes 1_6$ ($\otimes$ označuje Kroneckerov súčin) a r je 30×1 vektor s i-tou zložkou $r_i = (-2 + 4 * (i-1)/29)^2$.
Na základe 2000 simulácií pre skutočné hodnoty parametrov (σ_1^2, σ^2) z množiny {(0.1,10), (0.5,2), (1,1), (2,0.5), (5,0.2)} sa nasimulované pravdepodobnosti pokrytia intervalov získaných pomocou jednotlivých testovacích premenných pohybovali v rozpätí od 0.9450 do 0.9615, čo nenaznačuje že by skutočné pravdepodobnosti pokrytia týchto zovšeobecnených intervalov boli nižšie ako stanovená 95%-ná úroveň.
Isté rozdiely medzi jednotlivými testovacími premennými sa však objavujú pri porovnaní dĺžok intervalov pomocou nich skonštruovaných. Podľa našich skúseností sa v tomto smere javí ako nie príliš vhodná na použitie testovacia premenná T_λ, ktorá nielenže vedie k širším intervalom, ale navyše aj rozsah dĺžok jednotlivých intervalov je pomerne veľký. V uvedenom príklade pri $(\sigma_1^2, \sigma^2) = (5, 0.2)$ bol rozdiel medzi najkratším a najdlhším intervalom viac ako 2300. Priemerné dĺžky intervalov založené na 20 simuláciách sú uvedené v nasledujúcej tabuľke.

(σ_1^2, σ^2)	$(0.1, 10)$	$(0.5, 2)$	$(1, 1)$	$(2, 0.5)$	$(5, 0.2)$
T_2	25.43	9.53	11.45	14.50	32.02
T_1	9.06	7.06	18.17	39.22	108.01
$T_{1/\lambda}$	38.78	10.93	12.81	15.03	32.12
T_λ	23.45	22.08	83.53	177.32	483.75

Tabulka 1: Priemerné dĺžky 95%-ných konfidenčných intervalov.

3 Záver

Zovšeobecnené testovacie premenné predstavujú možnosť ako konštruovať, v istom zmysle presné, konfidenčné intervaly v úlohách s rušivými parametrami. V článku sme ukázali, ako ich je možné použiť na konštrukciu konfidenčných intervalov pre parameter σ_1^2 v modeli s dvomi variančnými komponentami. Navrhli sme ako voliť konštanty v testovacej premennej z článku [11] a uviedli sme tiež alternatívnu, výpočtovo jednoduchšiu testovaciu premennú. Okrem jednej, sa podľa simulačnej štúdie všetky nami skúmané testovacie premenné javia vhodné na použitie.

Reference

[1] Gnot S., Jankowiak-Rosłanowka M., Michalski A. (1992). *Testing for hypothesis in mixed linear models with two variance components.* Listy Biometryczne – Biometrical Letters **29**, 13 – 31.

[2] Gnot S., Michalski A. (1991). *Linear and quadratic estimation from inter- and intra-block sources of information.* Statistics **22**, 17 – 32.

[3] Imhof J.P. (1961). *Computing the distribution of quadratic forms in normal variables.* Biometrika **48**, 419 – 426.

[4] Michalski A. (2003). *On some aspects of the optimal statistical inference on variance components in mixed linear models.* Tatra Mountains Mathematical Publications **26**, 133 – 153.

[5] Michalski A., Zmyślony R. (1996). *Testing hypotheses for variance components in mixed linear models.* Statistics **27**, 297 – 310.

[6] Olsen A., Seely J., Birkes D. (1976). *Invariant quadratic unbiased estimation for two variance components.* The Annals of Statistics **5**, 878 – 890.

[7] Seely J., El-Bassiouni Y. (1983). *Applying Wald's variance component test.* The Annals of Statistics **11**, 197 – 201.

[8] Šírková L., Witkovský V. (2001). *On testing variance components in unbalanced mixed linear model.* Applications of Mathematics **46**, 191 – 213.

[9] Tsui K.W., Weerahandi S. (1989). *Generalized p-values in significance testing of hypotheses in the presence of nuisance parameters.* Journal of the American Statistical Association **84**, 602 – 607.

[10] Weerahandi S. (1995). *Exact statistical methods for data analysis* . Springer-Verlag, New York.

[11] Zhou L., Mathew T. (1994). *Some tests for variance components using generalized p-values.* Technometrics **36**, 394 – 402.

Poďakovanie: Ďakujem RNDr.V.Witkovskému CSc. za mnohé podnetné rady a pripomienky. Práca bola podporená grantami Vedeckej grantovej agentúry Slovenskej republiky VEGA 1/0264/03 a VEGA 2/4026/04.

Adresa: B. Arendacká, Ústav merania SAV, Dúbravská cesta 9,
841 04 Bratislava, SR

E-mail: barbora.arendacka@post.sk

INFERENCE ZALOŽENÁ NA SEKVENČNÍCH POŘADÍCH

Lucie Belzová

Klíčová slova: Pořadí, sekvenční pořadí, Wilcoxonův test.

Abstrakt: Tématem článku jsou „klasická" a sekvenční pořadí. Jsou zde uvedeny jejich definice, základní vlastnosti a vztah mezi nimi. Dále je ukazáno, že testové statistiky založené na pořadích resp. na sekvenčních pořadích (tj. v testové statistice nahradíme „klasické" pořadí sekvenčním) jsou za určitých předpokladů ekvivalentní.

1 Pořadí a sekvenční pořadí

Nechť $X_1, \ldots, X_n$ jsou nezávislé náhodné veličiny se spojitou distribuční funkcí F. Náhodné veličiny $X_1, \ldots, X_n$ uspořádáme podle velikosti a nejmenší z nich označíme $X_{(1)}$, druhou nejmenší $X_{(2)}$ až největší $X_{(n)}$. Platí tedy

$$X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}.$$

$X_{(i)}$ se nazývá i-tá pořádková statistika.

Jestliže náhodná veličina X_i je j-tá co do velikosti mezi veličinami $X_1, \ldots, X_n$, tj. $(X_i = X_{(j)})$, pak *pořadí* R_{in} této veličiny je rovno číslu j. Hodnota R_{in} je tedy rovna počtu těch veličin, které jsou menší nebo rovny X_i.

Dále definujeme *sekvenční pořadí* R_{ii} náhodné veličiny X_i jako pořadí X_i mezi veličinami $X_1, \ldots, X_i$. Barndorff-Nielsen [1] dokázali, že náhodné veličiny $R_{11}, R_{22}, \ldots R_{nn}$ jsou nezávislé a platí

$$\mathrm{P}(R_{ii} = r_i) = \frac{1}{i}, \qquad r_i = 1, \ldots, i; \quad i = 1, \ldots, n.$$

Uvažujme lineární pořadovou statistiku následujícího tvaru:

$$T_n = \sum_{i=1}^{n} c_{in} J_n\left(\frac{R_{in}}{n+1}\right), \tag{1}$$

kde $c_{1n}, c_{2n}, \ldots c_{nn}$ jsou známé regresní konstanty a $J_n(\frac{i}{n+1})$ pro $i = 1, \ldots, n$ jsou skóry generované následujícím způsobem:

$$J_n\left(\frac{i}{n+1}\right) = \mathrm{E}J(U_{(i)}),$$

kde $U_{(i)}$ je i-tá pořádková statistika z n nezávislých rovnoměrně rozdělených náhodných veličin na intervalu $(0, 1)$.
Dále předpokládáme, že

$$\int_0^1 J(u)du = 0, \tag{2}$$

$$0 < \int_0^1 J^2(u)du = A < \infty \tag{3}$$

a

$$\sum_{i=1}^{n} c_{in} = 0. \tag{4}$$

Nyní uvažujme statistiku založenou na sekvenčních pořadích:

$$M_n = \sum_{i=1}^{n} (c_{in} - \bar{c}_{i-1,n}) J_i \left(\frac{R_{ii}}{i+1} \right), \tag{5}$$

kde

$$\bar{c}_{i-1,n} = \frac{1}{i-1} \sum_{j=1}^{i-1} c_{jn} \qquad \text{a} \qquad \bar{c}_{0,n} = 0.$$

Tedy M_n je součtem nezávislých náhodných veličin.

Mason [3] dokázal, že pokud platí

$$\max_{1 \le i \le n} \frac{c_{in}^2}{\sum\limits_{j=1}^{n} (c_{jn} - \bar{c}_{nn})^2} = o(1)$$

jsou statistiky T_n a M_n asymptoticky ekvivaletní podle kvadratického středu, tj. platí

$$\mathrm{E} \left(\frac{T_n - M_n}{\sigma_n^2} \right) \overset{n \to \infty}{\longrightarrow} 0, \tag{6}$$

kde $\sigma_n^2 = \mathrm{var} T_n$.

2 Dvouvýběrový Wilcoxonův test

Nechť $X_1, \ldots, X_m$ resp. $Y_1, \ldots, Y_n$ je náhodný výběr z rozdělení s distribuční funkcí F resp. G.

Dvouvýběrový Wilcoxonův test testuje hypotézu, že distribuční funkce F a G jsou stejné, tj. $H_0 : F = G$, proti alternativě posunutí v poloze, tzn. $H_1 : G(x) = F(x - \Delta), \quad \Delta \neq 0$.

Veličiny $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ (tzv. sdružený výběr) uspořádáme vzestupně podle velikosti a označíme R_{iN}, $i = 1, \ldots N$, $(N = m+n)$ pořadí i-té veličiny ze sdruženého výběru. Pak Wilcoxonova statistika je rovna součtu pořadí druhého výběru, tedy

$$W_N = \sum_{i=m+1}^{N} R_{iN}. \tag{7}$$

Platí $T_N = W_N$ pro volbu $J(u) = u$ a

$$\begin{aligned} c_{iN} &= 0 \qquad i = 1, \ldots, m \\ &= 1 \qquad i = m+1, \ldots, N. \end{aligned}$$

Bohužel pro tuto skórovou funkci a tyto regresní konstanty neplatí podmínky (2) a (4), proto upravíme volbu následovně:

$$J(u) = u - \frac{1}{2}$$

$$\begin{array}{rcll} c_{iN} & = & -\frac{n}{N} & i = 1, \ldots, m \\ & = & \frac{m}{N} & i = m+1, \ldots, N. \end{array}$$

Potom pořadové statistiky T_N a M_N jsou rovny:

$$T_N = -\frac{n}{N} \sum_{i=1}^{m} \left(\frac{R_{iN}}{N+1} - \frac{1}{2} \right) + \frac{m}{N} \sum_{i=m+1}^{N} \left(\frac{R_{iN}}{N+1} - \frac{1}{2} \right) \tag{8}$$

$$M_N = -\frac{n}{N} \left(R_{11} - \frac{1}{2} \right) + \sum_{i=m+1}^{N} \frac{m}{i-1} \left(\frac{R_{ii}}{i+1} - \frac{1}{2} \right), \tag{9}$$

protože

$$\begin{array}{rcll} c_{iN} - \bar{c}_{i-1,N} & = & -\frac{n}{N} & i = 1 \\ & = & 0 & i = 2, \ldots, m \\ & = & \frac{m}{i-1} & i = m+1, \ldots, N. \end{array}$$

3 Simulace

3.1 Normální rozdělení

Uvažujme náhodné výběry $X_1, \ldots, X_m$ z N$(0, 1)$ a $Y_1, \ldots, Y_m$ z N$(\Delta, 1)$, kde $\Delta = 0, 0.25, 0.5, 0.75, 1, 1.5, 2, 2.5, 3$.

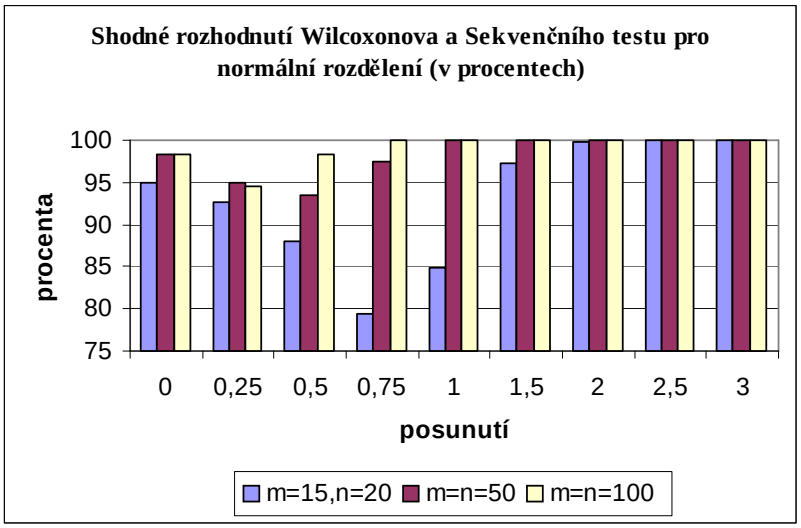
Pro rozsahy výběrů $m = 15$, $n = 20$ resp. $m = n = 50$ resp. $m = n = 100$ a různé velikosti posunutí (Δ) byly spočteny testové statistiky T_n (Wilcoxonova) a M_n („sekvenční Wilcoxonova“)a obě byly porovnány s kritickou hodnotu Wilcoxonova testu na hladině spolehlivosti $\alpha = 0.05$. Pro každou kombinaci volby rozsahu a posunutí se provedlo 1000 simulací.

procentuální zastoupení shodných rozhodnutí Wilcoxonova a „Wilcoxonova sekvenčního“ testu pro jednotlivé situace. Je vidět, že s rostoucími rozsahy výběrů jsou rozhodnutí testů ve více případech stejná. Dále, jak bychom očekávali, počet shodných rozhodnutí roste s rostoucím posunutím od určité hodnoty posunutí p (závisí na rozsazích výběrů). A naopak, pokud posunutí Δ je mezi 0 a p, počet stejných rozhodnutí klesá.

Pro $m = 15$, $n = 20$ a $\Delta = 0.75$ jsou v tabulce (Tab. 2) a v grafu (Graf 2) uvedeny počty jednotlivých možností rozhodnutí obou testů. (Tato kombinace parametrů měla nejméně shodných rozhodnutí 79,5%.) Pro ostatní kombinace parametrů je situace obdobná, tj. sekvenční test je slabší než Wilcoxonův.

Δ	**m=15,n=20**	**m=n=50**	**m=n=100**
0	95,0	98,3	98,4
0,25	92,6	94,9	94,6
0,5	88,0	93,5	98,3
0,75	79,5	97,4	100,0
1	84,8	100,0	100,0
1,5	97,3	100,0	100,0
2	99,8	100,0	100,0
2,5	100,0	100,0	100,0
3	100,0	100,0	100,0

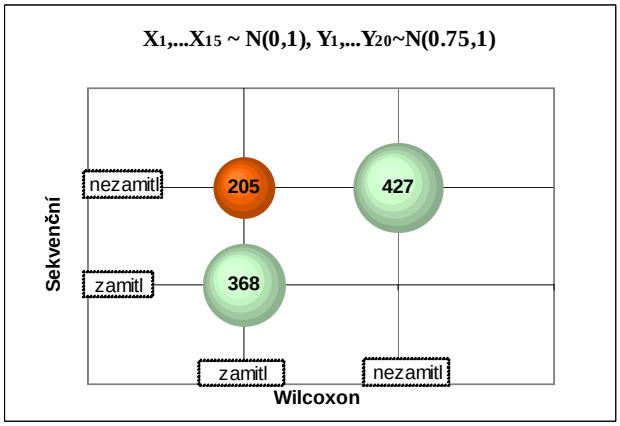
Tabulka 1: Shodné rozhodnutí Wilcoxonova a Sekvenčního testu pro normální rozdělení (v procentech).



Graf 1

	Wilcoxon	
Sekvenční	**zamítl**	**nezamítl**
zamítl	368	0
nezamítl	205	427

Tabulka 2: Normální rozdělení, $m = 15$, $n = 20$, $\Delta = 1,5$.



Graf 2

3.2 Logistické rozdělení

Vzhledem k tomu, že Wilcoxonuv test je lokálně nejsilnejší pořadový test (viz [2]) pro logistické rozdělení, byla obdobná simulace provedena i pro logistické rozdělení $\mathrm{L}(a, b)$, které má hustotu

$$f(x) = \frac{\exp\{-\frac{x-a}{b}\}}{(1+\exp\{-\frac{x-a}{b}\})^2} \qquad x, a \in R, \quad b > 0.$$

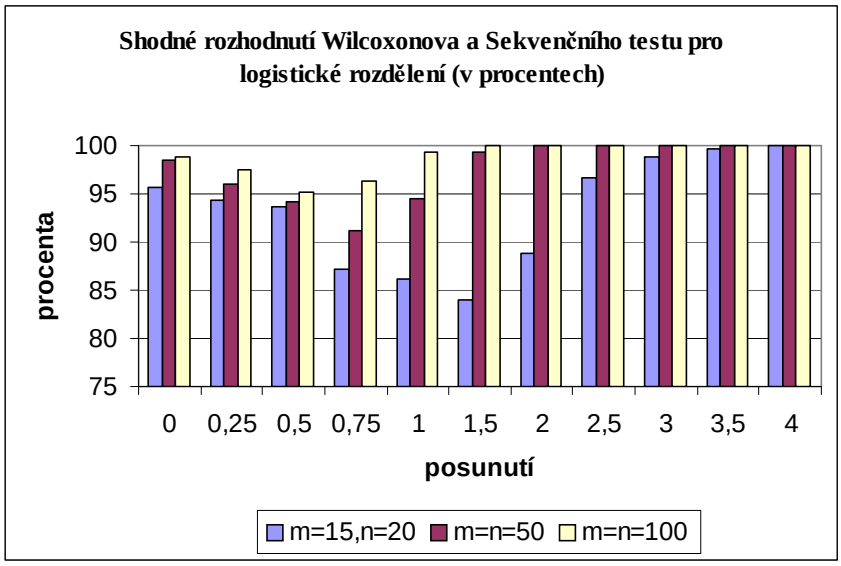
Analogicky jako u normálního rozdělení se nagerovaly náhodné výběry $X_1, \ldots, X_m$ z $\mathrm{L}(0, 1)$ a $Y_1, \ldots, Y_m$ z $\mathrm{L}(\Delta, 1)$, kde $\Delta = 0$, 0.25, 0.5, 0.75, 1, 1.5, 2, 2.5, 3, 3.5, 4 a $m = 15$, $n = 20$ resp. $m = n = 50$ resp. $m = n = 100$.

Výsledky jsou obdobné jako u normálního rozdělení. Dle tabulky (Tab. 3) a grafu (Graf 3) je patrné, že opět počet shodných rozhodnutí Wilcoxonova a sekvenčního testu roste s rozsahem výběrů a velikostí posunutí od určité hodnoty posunutí p. Při pevných rozsazích a když $\Delta \in \langle 0, p \rangle$, počet stejných rozhodnutí klesá.

Jako v případě normálního rozdělení i pro logistické rozdělení je zde uveden graf (Graf 4) a tabulka (Tab. 4) se zastoupením jednotlivých rozhodnutí obou testů pro kombinaci parametrů, u které bylo nejméně shodných rozhodnutí. Tentokrát tato situace nastala opět pro rozsahy výběrů $m = 15$, $n = 20$, ale velikost posunutí je větší a to 1,5. Pro ostatní kombinace parametrů je rozložení rozhodnutí obdobné,tedy opět můžeme prohlásit, že Wilcoxonův test je silnější než-li sekvenční.

Δ	**m=15,n=20**	**m=n=50**	**m=n=100**
0	95,6	98,5	98,8
0,25	94,4	96,0	97,5
0,5	93,6	94,1	95,2
0,75	87,1	91,1	96,3
1	86,1	94,5	99,3
1,5	84,0	99,4	100,0
2	88,8	100,0	100,0
2,5	96,7	100,0	100,0
3	98,8	100,0	100,0
3,5	99,6	100,0	100,0
4	100,0	100,0	100,0

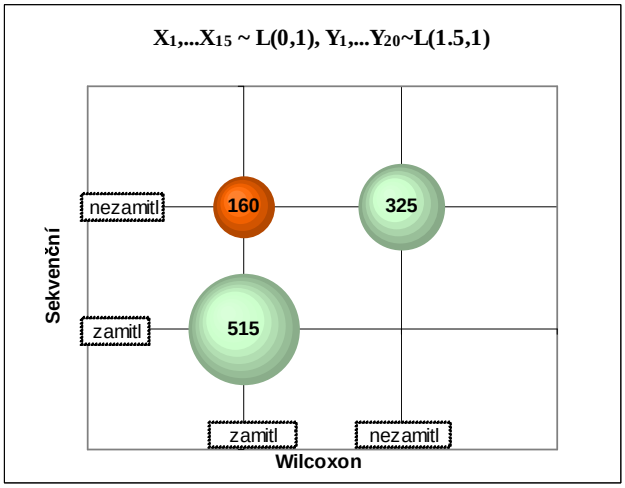
Tabulka 3: Shodné rozhodnutí Wilcoxonova a Sekvenčního testu pro logistické rozdělení (v procentech).



Graf 3

	Wilcoxon	
Sekvenční	**zamítl**	**nezamítl**
zamítl	515	0
nezamítl	160	325

Tabulka 4: Logistické rozdělení, $m = 15$, $n = 20$, $\Delta = 1,5$.



Graf 4

4 Závěr

Pro normální a logistické rozdělení jsme nevyvrátili platnost vztahu (6), tj. v testové statistiky T_n a M_n jsou asymptoticky ekvivaletní podle kvadratického středu. Dále se ukázalo, že Wilcoxonův test je silnější než jeho sekvenční analogie. A za třetí čím větší posunutí u druhého výběru uvažujeme, tím více je shodných rozhodnutí uvažovaných testů.

Reference

[1] Barndorff-Nielsen O. (1963). *On the limit behaviour of extreme order statistics.* The Annals of Matematical Statistics **34**, 992 – 1002.

[2] Jurečková J. (1981). *Pořadové testy.* SPN Praha.

[3] Mason David M. (1981). *On the use of a statistic based on sequential ranks to prove limit theorems for simple linear rank statistics.* The Annals of Statistics **9**, 424 – 436.

Poděkování: Tato práce je podporována výzkumným záměrem MSM 113200008.

Adresa: L. Belzová, KPMS, MFF UK, Sokolovská 83, Praha 8 - Karlín

E-mail: belzova@karlin.mff.cuni.cz

ČASOPROSTOROVÉ BODOVÉ PROCESY

Viktor Beneš, Michaela Prokešová

Klíčová slova: Časoprostorové procesy, kótovaný bodový proces, věrohodnost, Coxovy procesy, podmíněná intenzita, metoda minimálního kontrastu.

Abstrakt: Příspěvek uvádí základní přístupy k modelování náhodných bodových procesů v čase a prostoru. V první části se pracuje s pojmem podmíněné intenzity v kontextu kótovaných časových procesů. Druhá část se zabývá dvojně stochastickými procesy, kde se porovnávají modely s různě definovanými řídícími poli.

1 Základní pojmy časoprostorových procesů

Časoprostorové bodové procesy se užívají k modelování náhodných událostí v čase a prostoru (prostoru nejčastěji dvou- či tří- dimenzionálním). Oborů aplikací je mnoho, jmenujme např. epidemiologii – výskyty nákazy v regionu, nukleární medicínu – z radioaktivního zdroje implantovaného v orgánu se zachycují fotony na povrchu detektoru, seismologii – epicentra zemětřesení v Zemi, životní prostředí - výskyt lesních požárů, nebo zemědělství – růst plevele na poli, apod. O časoprostorových bodových procesech pojednávají monografie [12] a zejména [6], další časopisecké citace obsahuje přehledový článek [11].

V definici náhodného bodového procesu na podmnožině S Eukleidovského prostoru $\mathbb{R}^d$ se označí N systém lokálně konečných podmnožin S a vybaví se σ-algebrou $\mathcal{N} = \{B \in N;\ \mathrm{card}(B \cap A) = m,\ m = 0, 1, 2, \ldots,\ A \in \mathcal{B}(S)\}$, kde $\mathcal{B}(.)$ značí borelovskou σ-algebru na příslušné množině, $\mathcal{B}(\mathbb{R}^k) = \mathcal{B}^k$. Náhodný bodový proces je potom náhodný element $X : \ (\Omega, \mathcal{A}, P) \longrightarrow (N, \mathcal{N})$, kde zobrazení je definováno na obecném pravděpodobnostním prostoru. Současně $X(A)$ značí počet bodů X v $A \in \mathcal{B}(S)$. Zabývejme se pojmem věrohodnost realizace $(x_1, \ldots, x_n)$ bodového procesu X v omezené množině $A \subset S$.

Janossyho míra J_n (restrikce na A) má tvar

$$J_n(\mathrm{d}x_1 \times \cdots \times \mathrm{d}x_n \mid A) \approx P(\text{právě } n \text{ bodů v } A \text{ po jednom v } \mathrm{d}x_1, \ldots, \mathrm{d}x_n)$$

(v tomto přehledovém článku kvůli stručnosti stavíme na heuristických definicích, rigorozní postup lze najít např. v [6]. Bodový proces se nazývá regulární, existuje-li hustota j_n míry J_n vzhledem k μ^n, kde μ je daná referenční míra na S. Potom věrohodnost L_A realizace $(x_1, \ldots, x_n)$ na A je $L_A(x_1, \ldots, x_n) = j_n(x_1, \ldots, x_n|A)$.

Zajímá nás analytické vyjádření věrohodnosti. Uvažme nejprve Poissonův bodový proces v $A \subset \mathbb{R}^d$ s funkcí intenzity $\lambda(x)$. Zde realizace $(x_1, \ldots, x_n)$, má věrohodnost

$$L_A = \prod_{i=1}^{n} \lambda(x_i) \exp\left(-\int_A \lambda(x)\mathrm{d}x\right).$$

Pro většinu jiných prostorových (v $\mathbb{R}^d$) procesů je vyjádření věrohodnosti obtížné.

Pro časový bodový proces X v $\mathbb{R}_+$ vhodným podmíněním σ-algebrou $\mathcal{H}_{t_-}$ událostí do času t definujeme podmíněnou intenzitu $\lambda^\star$ předpisem

$$\lambda^\star(t)\mathrm{d}t \approx \mathbb{E}[X(\mathrm{d}t)|\mathcal{H}_{t_-}].$$

Položme $A = [0, T]$ a uvažme realizaci $t_1 < t_2 < \cdots < t_{X(T)}$ procesu X na A, píšeme $X(T) = X(A)$. Zde obecně má věrohodnost tvar

$$L_A(t_1, \ldots, t_{X(T)}) = \prod_{i=1}^{X(T)} \lambda^\star(t_i) \exp\left(-\int_0^T \lambda^\star(x)\mathrm{d}x\right). \tag{1}$$

Proto časová poloosa díky svému uspořádání podpoří modelování časoprostorového bodového procesu, technickým nástrojem je kótovaný bodový proces.

Nechť $(K, \mathcal{B}_K)$ je separabilní úplný metrický prostor kót s referenční mírou λ_K. Kótovaný bodový proces vzniká přiřazením kót $k_i \in K$ bodům $x_i \in S$. X je kótovaný bodový proces $\{(x_i, k_i)\}$ na $S \times K$ je-li $X_g = \{(x_i)\}$ bodový proces na S. Časoprostorový proces je potom kótovaný bodový proces na $\mathbb{R}_+ \times \mathbb{R}^d$. Kótovaný bodový proces X na $S \times K$ se nazývá regulární, jestliže existuje Janossyho hustota

$$\begin{aligned} j_n(t_1, \ldots, t_n, k_1, \ldots, k_n)\ \mathrm{d}t_1 \ldots \mathrm{d}t_n\ \lambda_K(\mathrm{d}k_1) \ldots \lambda_K(\mathrm{d}k_n) \quad &\approx \\ \approx\ P(\text{“body v } \mathrm{d}t_i \text{ s kótami v } \mathrm{d}k_i\text{”}). \end{aligned}$$

Pro regulární kótovaný bodový proces X na $\mathbb{R}_+ \times K$ se definuje podmíněná intenzita jako náhodná funkce $\lambda^\star(t, k) \approx \mathbb{E}[X(\mathrm{d}t \times \mathrm{d}k)|\mathcal{H}_{t_-}]$. Realizace X na $[0, T] \times K$ tvaru $(t_1, k_1), \ldots, (t_{X_g(T)}, k_{X_g(T)})$ má věrohodnost

$$L_T = \prod_{i=1}^{X_g(T)} \lambda^\star(t_i, k_i) \exp\left(-\int_0^T \int_K \lambda^\star(u, v)\mathrm{d}u\lambda_K(\mathrm{d}v)\right). \tag{2}$$

Jedním ze základních modelů časoprostorových bodových procesů jsou samobudící se procesy. Dospějeme k nim tak, že nejprve popíšeme časový Hawkesův proces na $S \subset \mathbb{R}_+$ ([7]). V tomto modelu se uvažují dva typy bodů a) stacionární Poissonův bodový proces imigrantů s intenzitou μ_c, b) pro existující body t_i následníci tvoří nezávislé Poissonovy bodové procesy s mírou intenzity $\mu(A - t_i)$, kde $\mu(S) < 1$ a $\operatorname{supp}\mu \subset \mathbb{R}_+$. Hustotu μ vzhledem k Lebesgueově míře značíme též μ.

Podmíněná intenzita Hawkesova procesu je lineární:

$$\lambda^\star(t) = \mu_c + \sum_{0 < t_i < t} \mu(t - t_i).$$

Tedy pro parametrický tvar hustoty μ se odhad parametrů modelu realizuje metodou maximální věrohodnosti užitím (1).

Ověření shody modelu s daty (viz [9]) je založeno na jiném základním principu nazývaném náhodná změna času (viz [14]). Volně řečeno, procházíme-li $\mathbb{R}_+$ od 0 tak, že v čase t je rychlost $\frac{1}{\lambda^\star(t)}$, potom okamžiky, kdy dosahujeme body procesu, tvoří stacionární Poissonův proces s jednotkovou intenzitou. Tedy po odhadu $\lambda^\star$ následuje posouzení transformované realizace známými metodami pro stacionární Poissonův proces.

Integrál z podmíněné intenzity $\Lambda^\star(t) = \int_0^t \lambda^\star(u)\mathrm{d}u$ se nazývá kompenzátor bodového procesu X na $\mathbb{R}_+$ a známý je rozklad formulovaný v následující větě. X je adaptovaný na filtraci $\mathcal{F} = \{\mathcal{F}_t, t \in \mathbb{R}_+\}$ (rostoucí systém σ-algeber), jestliže $X(t)$ je $\mathcal{F}_t$-měřitelné pro každé t.

Věta 1.1. *Nechť X je adaptovaný na filtraci $\mathcal{F}$ a má spojitou podmíněnou intenzitu $\lambda^\star$, pak proces*

$$M(t) = X(t) - \Lambda^\star(t)$$

je $\mathcal{F}$-martingal, t.j. pro každé $s > t > 0$

$$\mathbb{E}[M(s) \mid \mathcal{F}_t] = M(t).$$

Pro časoprostorové procesy formulujeme princip náhodné změny času přesně.

Věta 1.2. *Nechť X je kótovaný bodový proces na $\mathbb{R}_+ \times K$ s podmíněnou intenzitou $\lambda^\star(t, \kappa)$ kladnou na $[0, \infty) \times K$ a zleva spojitou v t λ_K-s.j., s kompenzátorem*

$$\Lambda_k^\star(t) = \int_0^t \lambda^\star(s, k)\mathrm{d}s,$$

splňujícím $\Lambda_k^\star(t) \to \infty$ při $t \to \infty$, λ_K-s.j. Potom při náhodných změnách času

$$(t, k) \mapsto (\Lambda_k^\star(t), k),$$

je X transformován na kótovaný Poissonův proces s jednotkovou časovou intenzitou a stacionárním rozdělením kóty $\lambda_K(.)$.

Obecně platí $\lambda^\star(t, k) = \lambda_g^\star(t) f^\star(k \mid t)$, kde $f^\star(k|t)$ je podmíněná hustota kóty v čase t při daném $\mathcal{H}_{t-}$ a $\lambda_g^\star$ je podmíněná intenzita X_g. U procesu s nepredikovatelnými kótami, kdy rozdělení kóty v x_i nezávisí na polohách a kótách $\{(x_j, k_j)\}$, pro něž $x_j < x_i$, je $f^\star(k \mid t) = f(k \mid t)$ nenáhodná funkce.

Jako aplikaci časoprostorového bodového procesu uvádíme modelování výskytu zemětřesení podle [10] založené na ETAS modelu (epidemic-type aftershock sequence), což je zobecněný Hawkesův samobudící se proces.

Výskyty zemětřesení jsou popsány kótovaným bodovým procesem s časovou dynamikou a kótami (x, y, M), kde (x, y) je průmět epicentra na zemský povrch a M síla zemětřesení. Jsou dány předpoklady:

a) $\lambda^\star(t, x, y, M) = j(M)\lambda^\star(t, x, y)$ pro nějakou deterministickou funkci j,

b) intenzita imigrantů je funkcí (x, y)
c) následníci jsou nezávislí, jejich střední počet je $\kappa(M)$,
d) rozdělení času větvení má hustotu pravděpodobnosti $g(t-\tau)$, kde τ je okamžik výskytu předchůdce,
e) rozdělení síly resp. polohy závisí na síle předchůdce M^* a jeho poloze ξ, η s hustotami $j(M \mid M^*)$ resp. $f(x-\xi, y-\eta \mid M^*)$

V zavedeném ETAS modelu je podmíněná intenzita

$$\lambda^\star(t,x,y) = \mu(x,y) + \sum_{i:t_i<t} \kappa(M_i) g(t-t_i) f(x-x_i, y-y_i \mid M_i).$$

Parametrická volba funkcí f a g umožňuje odhad parametrů modelu maximalizací věrohodnosti (2) a následně testování shody modelu s daty založené na myšlence z Věty 1.2.

2 Časoprostorové Coxovy bodové procesy

V další části představíme tři modely časoprostorových bodových procesů s aplikacemi zvláště v epidemiologii a ekologii. Všechny tři modely jsou Coxovy procesy, ovšem s různými typy řídících náhodných polí. Začněme tedy definicí Coxova procesu obecně na $\mathbb{R}^n$.

Definice 2.1. *Buď $\{Z(s) : s \in S\}$, $S \subseteq \mathbb{R}^n$ nezáporné náhodné pole takové, že s pravděpodobností 1 je $s \to Z(s)$ lokálně integrovatelná funkce. X nazveme Coxovým procesem řízeným polem Z (alternativně Coxovým procesem s řídící intenzitou Z), pokud je podmíněné rozdělení X za podmínky $Z = z$ rovno rozdělení Poissonova procesu s funkcí intenzity z.*

Uvažujeme-li Coxův proces na omezené množině $B \subseteq S$, $|B| < \infty$, potom je jeho hustota vzhledem ke standardnímu Poissonovu procesu dána vzorcem

$$f(x) = \mathbb{E}\left[\exp\left(|B| - \int_B Z(s)\mathrm{d}s\right) \prod_{s\in x} Z(s)\right], \quad x \in N(S). \tag{3}$$

Explicitní vyjádření použité střední hodnoty obvykle není k dispozici a numerická aproximace by vyžadovala počítání mnohorozměrných integrálů velké dimenze. Protože je ale díky podmíněné struktuře Coxových procesů a obecným vlastnostem Poissonova procesu možné vyjádřit různé charakteristiky procesu X pomocí charakteristik použitého řídícího pole, máme k dispozici jednoduché momentové metody odhadu parametrů modelu.

Přímo z definice Coxova procesu plyne, že pro míru intenzity procesu X platí

$$\Lambda(B) = \int_B Z(s)\mathrm{d}s, \qquad B \subseteq S, \tag{4}$$

nepodmíněná funkce intenzity je tedy rovna $\rho(s) = \mathbb{E}Z(s)$, a párová korelační funkce je dána vztahem

$$g(s_1, s_2) = \mathbb{E}[Z(s_1)Z(s_2)]/[\rho(s_1)\rho(s_2)]. \tag{5}$$

Obdobně se i další momentové míry a faktoriální momentové míry dají vyjádřit pomocí momentů náhodného pole Z. Další výhodou Coxových procesů je, že při vhodné volbě modelu pro řídící pole Z můžeme získat velmi flexibilní popis časoprostorové kovarianční struktury pozorovaného procesu X, použitelný pro nejrůznější reálné aplikace.

Dobře interpretovatelnou variantou Coxových procesů jsou takzvané log-Gaussovské Coxovy procesy (LGC) pro které $Z(s) = \exp(Y(s))$, kde $Y(s)$ je Gaussovské pole se střední hodnotou $\mu(s) = \mathbb{E}Z(s)$ a kovarianční funkcí $c(s_1, s_2) = \mathbb{C}\mathrm{ov}(Z(s_1), Z(s_2))$. Aby byl odpovídající Coxův proces správně definován, je třeba splnit jistá kritéria na hladkost kovarianční funkce viz [8]. V modelu nepožadujeme stacionaritu procesu Y, ale předpokládáme-li translační invariantnost a izotropii kovarianční funkce c dostáváme velmi jednoduché vztahy mezi μ a c (charakteristikami procesu Y) a funkcí intenzity ρ a párovou korelační funkcí $g(s_1, s_2) = g(\|s_1 - s_2\|)$ procesu X

$$g(\|s_1 - s_2\|) = \exp(c(s_1, s_2)), \tag{6}$$

$$\rho(s) = \exp(\mu(s) + c(s, s)/2). \tag{7}$$

Odhady v LGC modelech se provádí metodou minimálního kontrastu, kdy odhady parametrů parametrizujících μ a c a tedy i celý LGC model jsou hodnoty argumentu minima integrovaných rozdílů mezi teoretickou hodnotou a neparametrickým odhadem $\hat{g}$ funkce g

$$\int_{a_1}^{a_2} \{(\log \widehat{g}(r))^b - (\log g(r))^b\}^2 \mathrm{d}r. \tag{8}$$

Logaritmu se používá kvůli stabilizaci rozptylu, a_1, a_2 a b jsou volené konstanty.

Při použití Coxových procesů pro časoprostorové modelování se neužívá přístup pomocí kótovaných bodových procesů z první části našeho článku. Přímočará volba (viz [3]) je uvažovat v definici 1 prostor $S = [0, \infty) \times \mathbb{R}^n$, kde první rozměr odpovídá času a n je v reálných aplikacích rovno 2 nebo 3. Nejprve však popíšeme jiný model z [5] (oba užívají LGC proces), kde jde o prostorový Coxův proces měnící se v čase jako proces rození. Data, na něž je model aplikován, jsou pozice rostlinek dvou různých druhů plevele na ječmenném poli pozorované v diskrétních časových okamžicích během několika týdnů po jeho přeorání. Tedy začínáme s prázdnou konfigurací a postupně nám body přibývají. Zde $X_i(t), t \geq 0$ značí prostorový proces v $\mathbb{R}^2$ závisející na čase t, dvěma druhům plevele odpovídá $i = 1, 2$. Podmíněně na Gaussovském procesu Y na $\mathbb{R}^2$ jsou $X_i(t)$ nezávislé Poissonovy procesy rození, na $S = [0, \infty) \times \mathbb{R}^2$ mají míru intenzity $\gamma_i \times \nu_i$. Předpokládáme, že γ_i jsou absolutně spojité deterministické míry na $[0, \infty)$ a

$$\nu_i(B) = \int_B \exp(Y_i(s))\mathrm{d}s, \qquad B \in \mathcal{B}^2$$

$$Y_i = \omega V + \sigma_i U_i + m_i, \qquad i = 1, 2,$$

kde V, U_1, U_2 jsou nezávislé centrované Gaussovské procesy s jednotkovým rozptylem a korelačními funkcemi r, r_1, r_2. Ty byly voleny izotropní v exponenciálním tvaru $r_i(a) = \exp(-a/\beta_i)$. Parametry modelu β, β_1, β_2, ω, σ_1 a $\sigma_2 > 0$ ("prostorové" parametry) se odhadují metodou minimálního kontrastu a m_i je deterministická střední hodnota prostorového Gaussovského procesu. Ověření odhadnutého modelu se provádí simulačními testy různých charakteristik jako třeba funkce prázdného prostoru F (viz [13]) či párové korelační funkce g.

Právě předvedený LGC model je sice časoprostorový, ale díky součinovému tvaru intenzity $\gamma_i \times \nu_i$ a nezávislosti náhodných polí Y_i na čase je časoprostorová interakce a závislost dosti omezená. Větší flexibilitou se v tomto ohledu vyznačuje model z článku [3], který používá pro definici podmiňovací míry intenzity opravdu časoprostorový Gaussovský proces $Y(t, s)$.

Článek se zabývá epidemiologickou aplikací. Situace, kterou má daný Coxův proces modelovat, jsou výskyty určité nemoci v různých místech sledovaného regionu oznamované v diskrétních, ale vzhledem k rychlosti změny intenzity výskytu této nemoci velmi častých časových intervalech. Cílem je odhadnout z pozorovaných případů intenzitu rizika vzniku nemoci v daném čase a zvláště její lokální zvýšení.

Coxův proces je v tomto případě vhodným modelem, protože jak bylo ukázáno v [1] existuje dualita mezi časoprostorovou nehomogenitou rizika a časoprostorovým shlukováním pozorovaného procesu jednotlivých případů onemocnění. Každý takový shluk případů tedy odpovídá lokálně zvýšené řídící intenzitě $Z(t, s)$ a ta se v modelu musí měnit sdruženě v prostoru a čase. Řídící pole Z je dáno rovnicí

$$Z(t, s) = \lambda(s) \exp\{Y(t, s)\}, \tag{9}$$

kde $Y(t, s)$ je stacionární Gaussovský proces a $\lambda(s)$ je deterministická funkce. Zde $\lambda(s)$ popisuje změny v prostorové intenzitě ohrožené populace a $Y(t, s)$ odpovídá riziku nakažení chorobou v čase t a místě s.

Protože pozorování jsou prováděna v diskrétních časových okamžicích $t_1, \ldots, t_n$, je řídící intenzita prostorového Coxova procesu případů zaznamenaných mezi časy t_1 a t_2 rovna $\lambda(s) \int_{t_1}^{t_2} \exp\{Y(t, s)\}\mathrm{d}t$, $s \in \mathbb{R}^2$. Tato veličina ovšem nedefinuje prostorový LGC proces a není známo její přesné rozdělení. Proto se pro počet případů $X_{t_i}(B)$ mezi t_{i-1} a t_i v $B \in \mathcal{B}^2$ při pevném Y volí model Poissonova rozdělení

$$X_{t_i}(B) \sim Poisson\left[(t_i - t_{i-1}) \int_B \exp\{Y(t_i, s)\}\lambda(s)ds\right],$$

spoléhající se na dostatečně malé rozdíly mezi časy t_i a t_{i-1}. Z výpočetních důvodů se diskretizuje proces Y také prostorově, rozdělením celého sledovaného území na velké množství buněk. V takto upraveném modelu už jsou k dispozici jednoduchá analytická vyjádření prostorových intenzit ρ_{t_i} a párové korelační funkce g a opět je možno použít metodu minimálního kontrastu porovnáním s jejich neparametrickými odhady z dat.

Poslední model, který v tomto přehledu ukážeme, je poněkud složitější a používá jinou třídu Coxových procesů než předešlé dva. Jsou to takzvané G–shot noise Coxovy procesy (GSNC) zavedené v [2].

Buď $\{u_j, w_j\} \subset S \times [0,\infty),\ S \in mathcalB^k$ realizace Poissonova bodového procesu Π na $S \times [0,\infty)$ s mírou intenzity součinového tvaru

$$\nu_{\kappa,\alpha,\tau}(A \times B) = (\kappa(A)/\Gamma(1-\alpha)) \times \int_B w^{-\alpha-1} \exp(-\tau w)\mathrm{d}w,$$

$A \subseteq S,\ B \subseteq [0,\infty)$, kde $\alpha < 1$ a $\tau \geq 0$ jsou parametry s $\tau > 0$ pro $\alpha \leq 0$ a κ je nezáporná a nenulová Radonova míra na S. Realizaci $\{u_j, w_j\}$ jednoznačně odpovídá takzvaná G-míra

$$m(A) = \sum_j w_j \delta_{u_j}(A), \qquad A \subseteq S$$

kde δ značí Diracovu míru v bodě u_j. Řídící pole G-shot noise Coxova procesu X je pak definováno vzorcem

$$Z(s) = \sum_j k(s, u_j) w_j,$$

kde $k(\cdot, u)$ je jádro (pro jednoduchost můžeme předpokládat, že $k(\cdot, u)$ je hustota spojité náhodné veličiny). Pro $\alpha < 0$ je $\{u_j\}$ Poissonův proces s mírou intenzity $\frac{\tau}{-\alpha}\kappa(\cdot)$ a $\{u_j\}$ jsou nezávislé na hodnotách veličin w_j, které jsou vzájemně nezávislé a mají všechny stejné Gamma rozdělení $\Gamma(-\alpha, \tau)$. Situace je složitější pro $\alpha \geq 0$, protože pak máme nekonečně mnoho bodů $\{u_j\}$ i pro omezenou množinu S.

V práci [4] byly GSNC procesy rozšířeny na časoprostorové GSNC procesy. Zde analyzovaná data byla stejná jako v [5], tedy vývoj růstu plevele, ale nyní byl každý druh analyzován zvlášť.

Základní myšlenka časoprostorového rozšíření spočívá v definování rodiny G-měr $m_t, t \geq 0$ na S, odpovídajících intenzitám $\nu_{\kappa_t,\alpha,\tau}$ s

$$\kappa_t(\mathrm{d}s) = B(t)\,\mathrm{d}s,$$

kde $B(t), t \geq 0$ je kumulativní distribuční funkce, tak, že i rozdíly $(m_{t+\triangle t} - m_t)$ jsou G-míry a jsou nezávislé na m_t. Rodina odpovídajících intenzit $Z_t(\cdot)$ (s použitím jádra $k(\cdot, u)$ nezávislého na čase t) pak určuje časoprostorový GSNC proces X na $S \times [0,\infty)$. Takto definovaný proces má nezávislé přírůstky v čase, takže je možné interpretovat výsledný bodový vzorek jako součet podle času nezávislých prostorových GSNC procesů (připomeňme si, že toto neplatí pro LGC model, protože ten má nezávislé přírůstky v čase pouze podmíněně na řídící Gaussově intenzitě a součet dvou LGC procesů také není LGC proces.)

Stejně jako v případě LGC procesů se i zde parametry odhadují metodou minimálního kontrastu z ρ a g a ověření odhadnutého modelu se provádí

simulačními testy vybraných charakteristik. Co se týče simulování GSNC procesu na omezeném okně S, je třeba řešit problém okrajových efektů, a to způsobených jednak jádrovými funkcemi s neomezeným nosičem (k intenzitě GSNC procesu X pozorovaného v S přispívají i body u_j nacházející se velmi daleko od S), jednak pro $\alpha \geq 0$ faktem, že $\mathrm{card}(\{u_j\}) = \infty$. To se řeší jednak simulací Poissonova procesu $\{u_j\}$ na větším okně než je S, jednak oříznutím počtu $\{u_j\}$ dostatečně velkou konstantou. Jsou k dispozici i odhady takto způsobené chyby v simulované řídící intenzitě, viz [2], [4].

Reference

[1] Bartlett, M. (1964) *Spectral analysis of two-dimensional point processes.* Biometrika **51**, 299 – 311.

[2] Brix, A. (1999) *Generalized gamma measures and shot-noise Cox processes.* Advances in Appl. Probab. **31**, 929 – 953.

[3] Brix, A. and Diggle, P. J. (2001) *Spatiotemporal prediction for log-Gaussian Cox processes.* J. R. Stat. Soc. Ser. B **63**, 823 – 841.

[4] Brix, A. and Chadoeuf, J. (2002) *Spatio-temporal modeling of weeds by shot-noise G Cox processes.* Biometrical Journal **44**, 83 – 99.

[5] Brix, A. and Moller, J. (2001) *Space-time multi type log Gaussian Cox processes with a view to modelling weeds.* Scand. J. Statist. **28**, 471 – 488.

[6] Daley D.J., Vere-Jones D. (2003) *An Introduction to the Theory of Point Processes, Vol. I: Elementary Theory and Methods. Second Ed.* Springer.

[7] Hawkes A.G. (1971) *Spectra of some self-exciting and mutually exciting point processes.* Biometrika **58**, 83 – 90.

[8] Moller J.; Syversveen, A. R.; Waagepetersen, R. P. (1998) *Log Gaussian Cox processes.* Scand. J. Statist. **25**, 451 – 482.

[9] Ogata Y. (1988) *Statistical models for earthquake occurrences and residual analysis for point processes.* J. Amer. Statist. Assoc. **83**, 9 – 27.

[10] Ogata Y. (1998) *Space-time point process models for earthquake occurrences.* Ann. Inst. Statist. Math. **50**, 379 – 402.

[11] Schoenberg F.P., Brillinger D.R., Guttorp P. (2002) *Point processes, spatial-temporal.* In: Encyclopedia of Environmetrics, Ed. by El-Shaarawi A.H., Piegorsch W.W., Wiley, **3**, 1573 – 1577.

[12] Snyder D.L., Miller M.I. (1991) *Random Point Processes in Time and Space.* Wiley, New York.

[13] Stoyan, D., Kendall, W. S., Mecke, J. *Stochastic geometry and its applications.* Chichester: Wiley.

[14] Watanabe S. (1964) *On discontinuous additive functionals and Levy measures of a Markov process.* Japanese J. Math. **34**, 54 – 70.

Poděkování: Tato práce vznikla za podpory grantů GAČR 201/03/0946 a MSM 113200008.

Adresa: V.Beneš, M.Prokešová, KPMS MFF UK, Sokolovská 83,
186 75 Praha 8
E-mail: benesv@karlin.mff.cuni.cz, prokesov@karlin.mff.cuni.cz

POZNÁMKY KE SHLUKOVÉ ANALÝZE PRVOKŮ

Martin Betinec

Klíčová slova: Shluková analýza, hlavní komponenty, fylogenetické stromy, Trichomonadinæ.

Abstrakt: Příspěvek se zabývá vlivem parametrů shlukové analýzy (tj. volbou kódování nukleotidů, metriky a metodou shlukování) na fylogenetickou klasifikaci prvoků čeledi *Trichomonadinæ* a konfrontuje její výsledky s metodou hlavních komponent.

1 Úvod

Shluková analýza pronikla do nejrůznějších odvětví biologie např. formě fylogenetických stromů (v angl. *phylogenetic trees*, resp. *evolutionary trees*). Ty představují způsob popisu vývoje druhů organismů, tzn. jejich vzájemné příbuznosti jak ve smyslu identifikace daného vztahu tak i jeho kvantifikace.

Prostředkem umožňujícím žádaný popis je dendrogram vzniklý často pomocí některé z metod (hierarchické) shlukové analýzy. V této souvislosti se nabízejí dvě otázky. Jak závisí výsledný dendrogram na parametrech shlukové analýzy? Jaká je spolehlivost výsledného popisu, který byl získán na základě zkoumaného vzorku, vzhledem k vlastnostem celé populace? Odpovědí na druhou otázku se zabýval můj příspěvek na ROBUSTU'02, viz [2]. Nyní se budeme věnovat první z nich.

V části 2 se zaměříme na shlukovou analýzu a její výsledky porovnáme v části 3 se závěry analýzy hlavních komponent.

Výsledky budou ilustrovány na datech získaných skupinou parazitologů vedenou Doc.RNDr. Jaroslavem Flégrem,CSc. z PřF UK.[1] Jedná se o 22 vzorků téhož genu prvoků čeledi *Trichomonadinæ*.

2 Shluková analýza

Pod názvem shluková analýza se skrývá celá řada postupů (viz např. [5]), my se v dalším soustředíme na *hierarchické slučovací metody,* které se užívají ve fylogenetice. Výsledky této skupiny metod je možno zobrazit pomocí *dendrogramu*, který lze interpretovat jako hledaný popis evoluce.

2.1 Hierarchické shlukování

Označme množinu m studovaných objektů $\{O_1, \cdots, O_m\}$ symbolem $\mathcal{O}$.

[1] viz. `http://prfdec.natur.cuni.cz/~flegr/`

2.1.1 Hierarchické shlukování $\mathcal{H}$ budiž systém $\{\mathcal{A}_i\}_i$, kde $\emptyset \neq \mathcal{A}_i \subset \mathcal{O}$, tak že $\forall \mathcal{A}_i, \mathcal{A}_j \subset \mathcal{O}$ platí, že buď $\mathcal{A}_i \cap \mathcal{A}_j = \emptyset$, nebo $\mathcal{A}_i \cap \mathcal{A}_j = \mathcal{A}_i$, nebo $\mathcal{A}_i \cap \mathcal{A}_j = \mathcal{A}_j$, přičemž $\exists (i,j)$, pro níž $\mathcal{A}_i \cap \mathcal{A}_j = \mathcal{A}_j$.

$\mathcal{H}$ lze vyjádřit jako posloupnost rozkladů $\{\mathcal{K}_i\}_{i=0}^{m-1}$ množiny $\mathcal{O}$, kde $\mathcal{K}_i = \{\mathcal{A}_{i,1}, \cdots, \mathcal{A}_{i,m-i}\}$, tak, že pro pevné i jsou shluky $\mathcal{A}_{i,j}$ disjunktní ($j = 1, \cdots, m-1$), $\bigcup_{j=1}^{m-i} \mathcal{A}_{i,j} = \mathcal{O}$ a $\mathcal{K}_i$ je zjemněním $\mathcal{K}_{i+1}$, viz část 2.1.4.

2.1.2 Stratifikované hierarchické shlukování nechť je takové hierarchické shlukování $\mathcal{H}$, které je jednak *úplné* (tzn. platí $\mathcal{O} \in \mathcal{H}$ a $\{O_i\} \in \mathcal{H}$ pro $\forall i = 1, \cdots, m$), a zároveň je na $\mathcal{H}$ definován funkcionál h, který přiřazuje každému shluku $\mathcal{A} \in \mathcal{H}$ jeho shlukovací hladinu $h(\mathcal{A})$.

2.1.3 Koeficient nepodobnosti shluků D je funkcionál přiřazující každé dvojici shluků $(\mathcal{A}_i, \mathcal{A}_j)$ číslo $D(\mathcal{A}_i, \mathcal{A}_j)$, které vyjadřuje ohodnocení podobnostních vztahů mezi shluky $(\mathcal{A}_i, \mathcal{A}_j)$. Požadujeme, aby pro $\forall i, j$ platilo
$D(\mathcal{A}_i, \mathcal{A}_j) \geq 0$ a $D(\mathcal{A}_i, \mathcal{A}_j) = D(\mathcal{A}_j, \mathcal{A}_i)$, přičemž $D(\mathcal{A}_i, \mathcal{A}_i) = 0$.

2.1.4 Hierarchická aglomerativní shlukovací procedura s koeficientem nepodobnosti D nalezne pro množinu objektů $\mathcal{O}$ posloupnost jejích rozkladů $\{\mathcal{K}_i\}_{i=0}^{m-1}$. Zároveň každému shluku každého z rozkladů přiřadí jeho slučovací hladinu, a to v následující rekurzi:

1. Počáteční rozklad $\mathcal{K}_0$ množiny $\mathcal{O}$ tvoří shluky identické s původními objekty, tj. $\mathcal{A}_{0,i} := O_i$, pro $i = 1, \cdots, m$, tedy $\mathcal{K}_0 = \{\mathcal{A}_{0,1}, \cdots, \mathcal{A}_{0,m}\}$. Zároveň položme $h(\mathcal{A}_{0,i}) = 0$ pro $i = 1, \cdots, m$.

2. Do k. kroku ($1 \leq k \leq m-1$) vstupuje rozklad $\mathcal{K}_{k-1} = \{\mathcal{A}_{k-1,1}, \cdots, \mathcal{A}_{k-1,m-k+1}\}$. Buď $\mathcal{A}_{k-1,u}, \mathcal{A}_{k-1,v} \in \mathcal{K}_{k-1}$, taková dvojice shluků, pro níž $D(\mathcal{A}_{k-1,u}, \mathcal{A}_{k-1,v}) = \min_{\mathcal{A},\mathcal{B} \in \mathcal{K}_{k-1}} D(\mathcal{A}, \mathcal{B}) =: \mu_k$. Do rozkladu $\mathcal{K}_k$ beze změny zařadíme všechny shluky $\mathcal{A}_{k-1,j} \in \mathcal{K}_{k-1}$ (nezměněny zůstávají i hladiny $h(\mathcal{A}_{k-1,j})$) s výjimkou $\mathcal{A}_{k-1,u}$ a $\mathcal{A}_{k-1,v}$, místo nichž do $\mathcal{K}_k$ přidáme shluk $(\mathcal{A}_{k-1,u} \cup \mathcal{A}_{k-1,v})$, přičemž klademe $h(\mathcal{A}_{k-1,u} \cup \mathcal{A}_{k-1,v}) := \mu_k$.

3. Na závěr platí $\mathcal{K}_{m-1} = \{\mathcal{A}_{m-1,1}\} = \mathcal{O}$, navíc $h(\mathcal{A}_{m-1,1}) = \mu_{m-1}$

2.1.5 Podobnostní strom označuje stratifikované hierarchické shlukování $\mathcal{H}$, pokud pro $\forall \mathcal{A}, \mathcal{B} \in \mathcal{H}$ platí $\mathcal{A} \subset \mathcal{B} \Longrightarrow h(\mathcal{A}) \leq h(\mathcal{B})$.

Aby $\mathcal{H}$ byl podobnostním stromem, musí koeficient nepodobnosti shluků D splňovat následující podmínku (ta je nutná a postačující, viz [5]):

Nechť pro shluky $\mathcal{P}, \mathcal{Q}$ nabývá D minima. Označme $D(\mathcal{P}, \mathcal{Q}) = \mu$, pak pro libovolný jiný shluk $\mathcal{U}$ musí platit $D(\mathcal{P} \cup \mathcal{Q}, \mathcal{U}) \geq \mu$.

Podobnostní strom lze graficky jednoznačně reprezentovat *dendrogramem*. Na jeho svislé ose se vynáší hladina slučování příslušných shluků, které jsou na vodorovné ose seřazeny tak, aby bylo možno zobrazit jejich postupné slučování, viz obr. 2.

Podobnostní strom dané množiny objektů $\mathcal{O}$ nevyjde za všech okolností stejně. Závisí na volbě shlukovací metody (tj. D), na způsobu, jimž měříme

vzdálenost objektů (tj. na metrice), a v případě nominálních znaků i na způsobu jejich kódování. Vlivu těchto charakteristik se budeme věnovat podrobněji, předtím ještě zmiňme jednu zvláštnost spojenou s použitím shlukové analýzy pro fylogenetické stromy.

2.2 Klasifikace

Užitečnost shlukové analýzy spočívá zejména *v generování hypotéz o klasifikaci* zkoumaných objektů.

Navrhovaná klasifikace objektů $\mathcal{K}$ je totožná s rozkladem $\mathcal{K}_i = \{\mathcal{A}_{i,1}, \cdots, \mathcal{A}_{i,m-i}\}$ pro nějž platí, že $\mu_i \leq h_{kr} \leq \mu_{i+1}$, kde h_{kr} je zvolená mezní hodnota a μ_i, resp. μ_{i+1} jsou maximální slučovací hladiny rozkladů $\mathcal{K}_i$, resp. $\mathcal{K}_{i+1}$, blíže viz 2.1.4. Mez h_{kr} se obvykle klade mezi dvojici sousedních slučovacích hladin s největším (interpretovatelným) rozdílem.

Průběh slučování objektů uvnitř každého ze shluků $\mathcal{A}_{i,j}$ finálního rozkladu $\mathcal{K}_i$ výslednou klasifikaci neovlivní. Interpretujeme-li však dendrogram jako evoluční strom, pak se oproti výše zmíněnému zajímáme o celý průběh slučování, tj. o celý dendrogram od kořene až po listy.

2.3 Vliv kódování

Datový soubor se skládá z 22 vzorků téže části genomu prvoků čeledi *Trichomonadinæ*, každý o délce 1870 nukleotidů.

Nukleotidy jsou čtyři: *adenin, guanin, cytosin* a *thymin*. První dva patří mezi *puriny* zatímco druhá dvojce mezi *pyrimidiny*. Ve dvojité šroubovici DNA se proti sobě vyskytují vždy dvojice $A \leftrightarrow T$ a $C \leftrightarrow G$.

2.3.1 Způsoby kódování Pokud bychom s daty zacházeli jako s nominálními veličinami, byli bychom nuceni se omezit na sledování shod daných vzorků v jednotlivých nukleotidech.

Vhodné kódování nukleotidů umožní zacházet se znaky jako s kardinálními veličinami, a tedy i odrážet výše zmíněné vlastnosti nukleotidů.

V článku ([4]) autor navrhuje použít (A,G,C,T) = $(1, 2, 5, 6)$, tento způsob nazývám *originální*. Vzhledem k výše zmíněným vlastnostem nukleotidů je toto kódování nesymetrické, proto stojí za úvahu jeho varianty, jež označuji jako *opačné* – (A,G,C,T) = $(2, 1, 6, 5)$, tj. opačná nesymetrie komplementárních bází, a *posunuté* – (A,G,C,T) = $(1, 2, 6, 5)$, srov. [2]. Pro zjištění, zda je vhodně zvolena vzdálenost oddělující puriny od pyrimidinů, je možno ji zvětšit, např. (A,G,C,T) = $(1, 2, 15, 16)$ – toto kódování označme jako *vzdálené* (v obr. 2 jako *long*).

Ve všech případech byly stromy pěstovány metodu *nejbližšího souseda* za použití *eukleidovské* metriky, blíže viz část 2.4.1.

2.3.2 Výsledky Dendrogramy vzniklé z *originálního*, *posunutého* i *opačného* kódování vyšly (až na měřítko na svislé ose) totožné.

Výsledek vzniklý ze *vzdáleného* kódování se od ostatních liší jen v nejnižších patrech – tím, zda se objekt `GPO` sloučí nejprve se shlukem {`CYG, TUMS`} (jako v případě první trojice způsobů), anebo s dvojicí {`AA9, OAM`} (*vzdálené* kódování), viz horní řádek obrázků 2.

Vzhledem k robustnosti výsledků vůči vlivu kódování budeme v dalším používat *originální* kódování.

2.4 Vliv metriky

Nechť je i-tý prvok reprezentován vektorem $\boldsymbol{x}_i = (x_{i,1}, \cdots, x_{i,p})$, kde $i = 1, \cdots, 22$ a $p = 1870$.

2.4.1 Metrika je zobrazení d: $\boldsymbol{R}^p \times \boldsymbol{R}^p \longrightarrow \boldsymbol{R}_0^+$, které pro $\forall \boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{x}_k \in \boldsymbol{R}^p$, splňuje vlastnosti reflexivity $d(\boldsymbol{x}_i, \boldsymbol{x}_j) = 0 \iff \boldsymbol{x}_i = \boldsymbol{x}_j$, symetrie $d(\boldsymbol{x}_i, \boldsymbol{x}_j) = d(\boldsymbol{x}_j, \boldsymbol{x}_i)$ a trojúhelníkové nerovnosti $d(\boldsymbol{x}_i, \boldsymbol{x}_j) \leq d(\boldsymbol{x}_i, \boldsymbol{x}_k) + d(\boldsymbol{x}_k, \boldsymbol{x}_j)$.

Většina používaných metrik je odvozena od *Minkowského* metriky

$$d_t(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left(\sum_{k=1}^{p} |x_{i,k} - x_{j,k}|^t \right)^{\frac{1}{t}} \qquad t = 1, 2, 3, \cdots, \tag{1}$$

např. *manhattanská* (též *city-block*) metrika vznikne z d_t pro $t = 1$. Již zmiňovaná *eukleidovská* metrika volbou $t = 2$. *Supremální* metrika vznikne z d_t jako: $d_\infty(\boldsymbol{x}_i, \boldsymbol{x}_j) = \lim_{t\to\infty} d_t(\boldsymbol{x}_i, \boldsymbol{x}_j) = \max_{k=1,\cdots,p} |x_{i,k} - x_{j,k}|$. Pro uvedené metriky platí: $d_1(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq d_2(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq \cdots \geq d_\infty(\boldsymbol{x}_i, \boldsymbol{x}_j)$.

Pro srovnání sledujme ještě chování *canberrské* metriky, která vztahuje vzdálenost složek k dvojnásobku jejich aritmetického průměru (tj. relativně snižuje váhu téhož rozdílu se vzrůstající vzdáleností od nuly):

$$d_{Cb}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{k=1}^{p} \frac{|x_{i,k} - x_{j,k}|}{|x_{i,k} + x_{j,k}|} \tag{2}$$

2.4.2 Výsledky Použití *eukleidovské* metriky vydá stejný strom jako užití metriky *manhattanské* a téměř stejný jako *canberrská* metrika. Ta se liší jen tím, že odděluje `pentatrichomonas` těsně před oddělením shluku {`tenax, vaginalis`}, zatímco u obou předešlých je tomu naopak, viz obr. 2. Vzhledem k blízkosti slučovacích hladin těchto shluků jde o změnu zanedbatelnou.

Supremální metrika je z podstaty věci nevhodná. Vzhledem k použitému kódování může rozdíl složek $(x_{i,k} - x_{j,k})$ vektorů $\boldsymbol{x}_i$ a $\boldsymbol{x}_j$ nabýt pouze hodnot $1, 3, 4$ a 5. Pro téměř libovolnou dvojici $\boldsymbol{x}_i$ a $\boldsymbol{x}_j$ se tak mezi 1870 složkami najde alespoň jedna dvojice $1 \leftrightarrow 6$, tedy $d_\infty(\boldsymbol{x}_i, \boldsymbol{x}_j) = 5$.

Ze zjištěného plyne, že smysluplně použité metriky výsledek podstatně neovlivní. V dalším budeme používat metriku *eukleidovskou.*

2.5 Vliv shlukovací metody

Jednotlivé shlukovací metody se od sebe liší volbou koeficientu nepodobnosti D mezi shluky (ty označme $\mathcal{A}$ a $\mathcal{B}$).

2.5.1 Zkoumané metody – všechny splňují podmínku podobnostního stromu (viz část 2.1.5), na rozdíl od některých neuváděných, leč častých metod (např. *mediánová*, *centroidní*), jejichž dendrogramy nelze interpretovat jako evoluční stromy.

1. *Metoda nejvzdálenějšího souseda (complete linkage, furthest neighbour)* – zkráceně značena *FN* – klade $D_{FN}(\mathcal{A},\mathcal{A}) = 0$ a pro $\mathcal{A} \neq \mathcal{B}$ $D_{FN}(\mathcal{A},\mathcal{B}) = \max_{O_i\in\mathcal{A},O_j\in\mathcal{B}} d(O_i,O_j)$.
2. *Metoda nejbližšího souseda (single linkage, nearest neghbour, friend of friends)* – zkráceně *NN*: $D_{NN}(\mathcal{A},\mathcal{B}) = \min_{O_i\in\mathcal{A},O_j\in\mathcal{B}} d(O_i,O_j)$.
3. *Metoda průměrné nepodobnosti shluků (group average)*: klade $D_{GA}(\mathcal{A},\mathcal{A}) = 0$ a označuje-li $|\mathcal{A}|$ počet prvků $\mathcal{A}$, pak pro $\mathcal{A} \neq \mathcal{B}$:
$$D_{GA}(\mathcal{A},\mathcal{B}) = \frac{1}{|\mathcal{A}|\cdot|\mathcal{B}|}\sum_{(i,j)} d(O_i,O_j)\,.$$
4. *Wardova a Wishartova metoda* Pro sloučeni shluků $\mathcal{A},\mathcal{B}$ do shluku $\mathcal{C}$ minimalizuje přírůstek vnitro-shlukové variability $I_{AB} = E_C - (E_A + E_B)$, kde $E_A = \sum_{i:O_i\in\mathcal{A}}\sum_j (o_{ij} - \overline{o}_j)^2$.

2.5.2 Výsledky Všechny metody (viz obr. 2) tvoří shluky {`cirka, 209, CYG2, M3, skupina1, CYG, TUMS, Q9, AA9, OAM`} a {`nonconforma, augusta, mobilensis, foetus`}.

Všechny až na *Wardovu* metodu k prvnímu jmenovanému shluku přiřadí i `GPO`. Navíc oddělují `trichomitus` samostatně od ostatních a vyčleňují samostatný shluk {`KAJ, LMA, SL`}.

Metody *průměrných nepodobností, nejbližšího a nejvzdálenějšího souseda* produkují téměř totožné stromy. Liší se pořadím shlukování `Q9` a `pentatrichomonas`. `Q9` se slučuje na velmi nízké hladině – buď se spojí nejprve se shlukem {`CYG, TUMS`} (jako v případě metod *prům. nepodobností* a *NN*), anebo s dvojicí {`AA9, OAM`} (metoda *FN*). `Pentatrichomonas` se buď nejprve sloučí s hlavním shlukem a až následně se připojí shluk {`tenax, vaginalis`} (metoda *NN*), anebo nejdříve splyne `pentatrichomonas` s {`tenax, vaginalis`}, a celá trojice pak sroste s hlavním shlukem (zbylé dvě metody), vše se děje na velmi blízkých hladinách.

2.6 Závěr

Zjistili jsme, že dendrogramy jsou takřka nedotčeny jak volbou kódování, tak i metriky. Největší dopad na výsledek vykazují shlukovací metody. Nabízejí

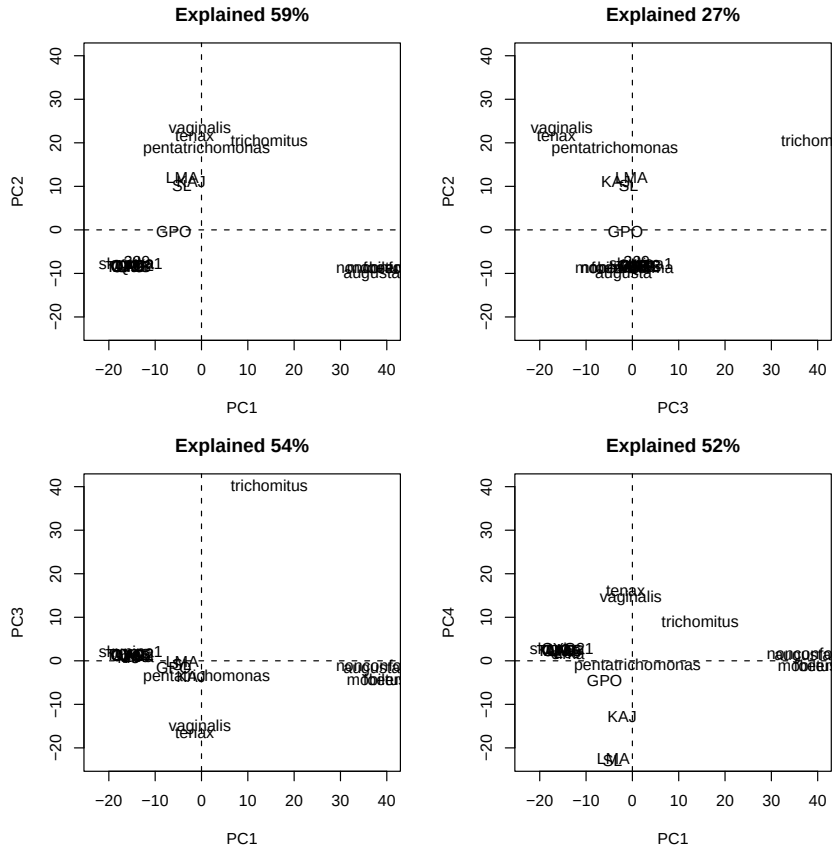
se dvě skupiny řešení: podle metod *prům. nepodobností, NN a FN* a podle *Wardovy* metody.

Zatímco první možnost by navrhovala rozčlenit objekty do 3 shluků (pokud na více, pak vzhledem k blízkosti hladin až do 8), se shlukem {`noncon-forma,...,foetus`}, samostatným `trichomitem` a zbytkem. Druhé řešení by sice též klasifikovalo do 3 shluků, ty by však kromě prvního shluku měly odlišné složení.

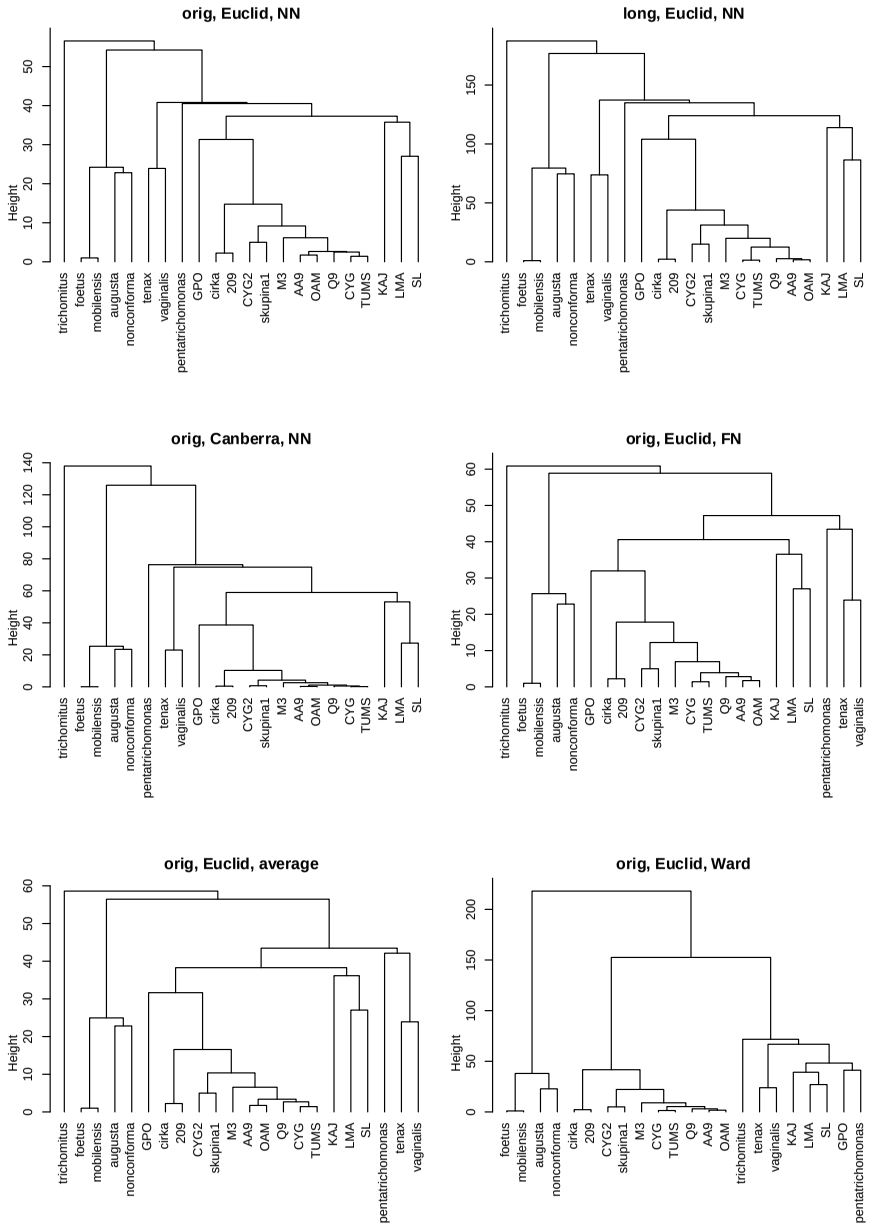
Při rozhodování o tom, na které řešení lze více spoléhat, nám může pomoci projekce dat do hlavních komponent.

3 Analýza hlavních komponent

Téměř 80% variability v datech lze vysvětlit projekcí do prvních čtyř faktorů (označme je $PC_1, \dots PC_4$). Situaci lze nalézt na obr. 1. Na základě diagnos-



Obrázek 1: Projekce objektů.



Obrázek 2: Srovnání výsledků různých kódování, metrik a metod.

tických koeficientů (viz např. [1]) lze stanovit vlivnost jednotlivých objektů vůči danému faktoru i jejich zkreslení v projekci na daný faktor. Výsledné hodnoty není možné z prostorových důvodů uvádět zde, zájemce odkazuji na [3]. Z obr. 1 je patrno, že důležitou roli hrají zejména:

První faktor (PC_1), který zřetelně odděluje shluk {`augusta, foetus, mobilensis, nonconforma`} od ostatních (odtržení tohoto shluku je v souladu s výsledky shlukové analýzy, viz obr. 2). Tento shluk, jehož zástupci jsou projekcí na PC_1 zároveň nejméně zkresleni, přispívá takřka ze 70% k variabilitě vysvětlené podél PC_1 (43% celkové). Nízkou míru zkreslení projekcí na PC_1 vykazují i `AA9, CYG, OAM, Q9, TUMS` a `M3, skupina1, CYG2`.

Třetí faktor(PC_3), který výrazně odděluje `trichomitus` od zbylých vzorků. To odpovídá první skupině dendrogramů. `Trichomitus` významně přispívá k variabilitě vysvětlené podél PC_3 (73.5% ze celkových 11%), zároveň je tento objekt projekcí do PC_3 zkreslen zdaleka nejméně ze všech.

Podél zbylých faktorů jsou objekty rozmístěny relativně rovnoměrně.

3.1 Závěr

Na základě pohledu do první faktorové roviny lze pochopit výsledek *Wardovy a Wishartovy metody*. Další dimenze PC_3 však ukazuje, že oddělení `trichomitu` je natolik výrazné, že pro fylogenetickou analýzu lze před metodou *Wardovou* upřednostnit metody *prům. nepodobností a nejbližšího a nejvzdálenějšího souseda,* z nichž ovšem na základě PCA nelze žádnou jednoznačně prohlásit za nejlepší.

Reference

[1] Aluja T., Morineau A. (1999). *Aprender de los datos: el análisis de los componentes principales*. EUB, Barcelona, 42 – 52.

[2] Betinec M. (2002). *O spolehlivosti vývojových stromů*. ROBUST'02, Sborník prací Dvanácté zimní školy JČMF (eds. J. Antoch, G. Dohnal, J. Klaschka), JČMF, Praha, 1 – 15.

[3] `http://betinec.matfyz.cz/doc/robust04/robust04.html`

[4] Efron B., Halloran E., Holmes S. (1996). *Bootstrap confidence levels for phylogenetic trees*. Proc. Natl. Acad. Sci. USA **93**, 13429 – 13434.

[5] Lukasová A. Šarmanová J. (1985). *Metody shlukové analýzy*. SNTL, Praha, 63 – 72.

Poděkování: Rád bych vyjádřil svůj dík Doc. RNDr. Jaroslavu Flégrovi, CSc. a jeho spolupracovníkům za poskytnutí dat a cenných konzultací. Tato práce je podporována výzkumným záměrem MSM 113200008.

Adresa: M. Betinec, Katedra sociologie, FF UK Praha, Celetná 20,
116 42 Praha 1

E-mail: `betinec@matfyz.cz`

REGRESSION WITH TRUNCATED DATA

Marek Brabec

Keywords: Regression, truncated data, spectral estimation.

Abstract: In this paper, we will present an example of the analysis of historical height data. First, we will discuss some conceptual problems related to the fact that no representable surveys are available for historical periods of interests (18th and 19th centuries). Then we will state a statistical model that can be used to correct available data (Swedish soldiers' measurements) for their selectivity with respect to the general population height distribution. The model is based on truncated normal regression. There, we will concentrate on periodicity in the height data, whose time series is spanning more than a century. In addition to presenting some explorative spectral estimates, we will discuss some problems and features related to maximum likelihood estimation in the model.

1 Introduction

In this paper, we will present an interesting practical problem encountered by anthropologists and historians. The goal is to estimate dynamics of human population mean heights in the past, from historical data (18th and 19th centuries). As one can expect, the task is complicated by the fact that informative data of high quality are difficult to obtain. Namely, no reasonably representative height surveys were organized then. Small samples are available from historical records here and there (e.g. family records of aristocrats etc.), their relevance for population height distribution is dubious, to say the very least, however. On the other hand, large amounts of systematically and rather precisely measured data are available from military records. Since the soldiers came from various social strata, geographical areas etc., historians think that military drafts spatially covered the concurrent (healthy adult male) population to a reasonable extent, [8], [9]. Even if we take this for granted, one substantial problem remains, however. To illustrate it, we consider the fact that while the recent adult population heights are known to follow normal distribution rather closely and there is no substantial reason why their historical counterparts should not behave similarly, the military sample shows consistently substantial positive skewness. Boxplots in Figure 1 demonstrate the situation for one particular sub-group of soldiers (born in Swedish rural areas) from the dataset that was collected by Richard Steckel and Lars Sandberg, [7], and which we will analyze subsequently.

Why is this happening? The answer is simple: military sample, as it stands, is not representative for the healthy male population as a whole. It is highly selective in the sense that the army disliked short men. Its preference for taller soldiers was embodied in a simple policy: men shorter than

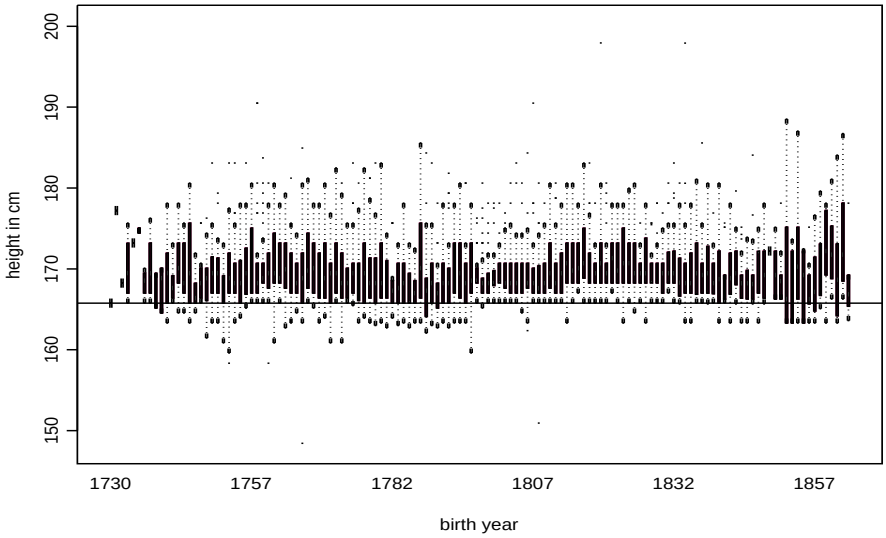


Figure 1: Height of rural-born soldiers,distributional summaries by birthyear.

a prescribed value (minimum height requirement, MHR) should not be drafted. Therefore, only those at or above MHR should appear in the military sample, theoretically. Comparing boxplots in Figure 1 with a presupposed MHR of 165.76 cm (horizontal line), we can see that while the policy really led to the apparent right skew, there are some data below MHR as well. We checked with prof. Komlos (economic historian from University of Munich, who introduced us to the substantive problem of historical height distribution estimation) and made sure that these are not coding errors. Although it was educational to learn that (even in kingdom od Sweden ...) a few individuals maintained to get into the draft while being (even seriously) below the MHR, one quickly realizes that while data obtained from these individuals can tell something about how effectively the policy was implemented, they do not tell much about the general male population distribution. Therefore, it became customary among historians to discard them and analyze only those above MHR, [6].

While "saturated" (i.e. on year-by-year basis) analyses for this data type have been attempted, [6], (including analyses of this particular dataset [7]), current interest of economic and anthropometric historians focuses on analysis of some generalizable long-term properties of the yearly (indexed by birthyear) height time series (like trends, periodicity etc.). Here, we will focus on the periodicity properties (after some simple trend and inter-regional

corrections). This was the task which was brought to us by our customer (prof. Komlos), who was interested in getting picture of the periodicity properties in an explorative style. This interest is connected to the general attention that economic historians pay to height (and other anthropological variables that thay see as possible indicators of "biological standard of living"), see e.g. [11]. The ultimate idea is that biological changes (like height fluctuations) might (to some extent) reflect changes in economic conditions (e.g. through food price and hence food availability) and hence can be potentially used as economic indicators more relevant to human well being than traditional econometric characteristics like GDP. Although one can be a bit skeptical about this upshot, many economic historians take this route, see e.g. [4]. Present investigation was motivated by somewhat more modest goal of comparison between periodic properties of height series and some economic (e.g. food price) series as a rough check of sensibility of the previous attempts to use heights as indicators (then, for instance the periodicity properties should be roughly "similar"). Obviously, such comparisons straightforwardly lead to the need to estimate the spectra.

Nevertheless, it is immediately clear, that the fact that all the heights below MHR are discarded precludes standard statistical/spectral analysis and calls for a model that corrects for this complication. We will formulate one such model in the section 2.

2 Model

The available data consist of (about 17 thousand) measurements of adult Swedish army soldiers, 22 years or older at the time of measurement, born between 1711 and 1864). To assess periodicity properties of the Swedish healthy male population height time series (indexed by birthyear), we outlined the following simple (linear) model (1) with normally distributed errors. Its form has been proposed after certain amount of data explorations and discussions with anthropometric history experts.

$$Y_{tij} = \mu + \alpha_i + \beta t + \Sigma_{k=1}^{F} \left(\delta_{1k} \cos(2\pi t f_k) + \delta_{2k} \sin(2\pi t f_k)\right) + \epsilon_{tij} \qquad (1)$$

where:

- Y_{tij} is the height of j-th man from general Swedish population, born in year t at i-th birth location.
- Time is indexed by birthyear, $t = 1$ corresponds to 1711.
- Birth location is indexed as $i = 0$ for "unknown", $i = 1$ for "rural", $i = 2$ for "urban".
- $\epsilon_{tij} \sim N\left(0; \sigma^2\right)$, independently across t, i, j's

Due to its linear (additive) structure, interpretation of the model is rather simple. It tries to assess amount of variability associated with periodic movements of various frequencies $f_1, \ldots f_F$ after correcting for possible birth-location differences and for possible (common) linear trend (as a simple form of non-stationarity). It is precisely the correction, together with the fact that the data are not balanced (having different sample sizes for different birth-years) which calls for a formulation in regression style and which precludes straightforward periodogram estimation based on standard estimators, [2]. Note that in the non-trigonometric part, the model resembles analysis of covariance. It fits a common linear trend shifted up or down differently at different birth locations, allowing for different average heights at rural/urban locations, a phenomenon which has been well documented for both historical and recent heights [5]. We use a particular parametrization with $\alpha_0 = 0$, which means that the mean height of a men born at unknown location (either rural or urban) is given by μ plus linear and trigonometric terms in time. α_1 corresponds to the difference between mean height of men born at the same year at rural and unknown locations. Similarly α_2 corresponds to difference between urban and unknown locations.

In order to roughly mimic periodogram estimation of variability at Fourier frequencies, we choose frequencies $f_i = \frac{1}{75}, \frac{1.5}{75}, \frac{2}{75}, \frac{3}{75} \ldots, \frac{37}{75}$, which are close to what the Fourier frequencies would be, if we had a single series of length 150 (with frequency $\frac{1.5}{75}$ added, based on preliminary data explorations).

Model (1) is nicely interpretable and would be easily fitted to data from historical Swedish population-representative male height surveys. The only flaw is that unfortunately no such surveys were performed. Available military data are non-representative of the general male population due to the MHR enforcement (and discarding the below-MHR measurements, see 1). Nevertheless, the mis-representation can be corrected rather easily, if we think of the military sample as of a sample from general population, which is left-truncated at the MHR. Then, for the available military data Y'_{tij}, we get the truncated regression model (2) from the original OLS model (1).

$$\begin{aligned}
Y'_{tij} &\quad \text{remains unobserved if } Y_{tij} < \tau_t \\
Y'_{tij} &= Y_{tij} \qquad \text{if} \qquad Y_{tij} \geq \tau_t \\
Y_{tij} &= \mu + \alpha_i + \beta t + \Sigma_{k=1}^{F} \left(\delta_{1k} \cos(2\pi t f_k) + \delta_{2k} \sin(2\pi t f_k)\right) + \epsilon_{tij} \\
\epsilon_{tij} &\sim N\left(0; \sigma^2\right) \\
\alpha_0 &= 0
\end{aligned} \tag{2}$$

Note that the MHR (τ_t) can generally vary in time. Ideally, it should be known from military regulations. Practically, it is not completely known and have been "expertly estimated" by historians (prof. Komlos). We have used both constant and time-varying MHR estimates and found that in terms of the main goal of the analysis. i.e. of rough spectral shape assessment, there are no substantial differences. More careful version with time-varying MHR was used to get results in section 3, however.

We make use maximum likelihood approach to estimate unknown parameters (μ, α_1, α_2, β, σ^2, $\delta_{11}, \ldots \delta_{1F}, \delta_{21}, \ldots \delta_{2F}$) of the model (2). Because of truncation, the model becomes nonlinear and no explicit formulas for the MLE's are available. Therefore, we maximize the loglikelihood (which is still rather easy to wright down) numerically, using a Newton-Raphson-like routine. We use the S-plus, especially the censorReg environment, [10] for the necessary computations.

3 Results and discussion

Before discussing the periodicity properties, we did some tests of the model (2). Namely, we used asymptotic likelihood ratio test (LRT) to check whether: i) birth-location specific intercepts are necessary ($p < 0.0001$ for $H_0 : \alpha_1 = \alpha_2 = 0$), ii) linear time trend is necessary ($p < 0.0001$ for $H_0 : \beta = 0$). No opening for a substantial simplification of the model structure was detected here.

To assess periodicity, one can look at MLE estimates $\hat{\delta}_{1k}$, $\hat{\delta}_{2k}$, $k = 1, 2, \ldots, 38$. Or, more conveniently at $\hat{\gamma}_k = \delta_{1k}^2 + \delta_{2k}^2, k = 1, 2, \ldots 38$ (respectively their simple transformation to decibels: $10.log_{10}(\hat{\gamma})$). They are analogous to periodogram estimate of "raw spectrum". Figure 2 compares raw $\hat{\gamma}_k$'s (dots) with their smoothed versions (solid line for the smooth, dotted line for pointwise computed 95% confidence interval limits). Smoothing is

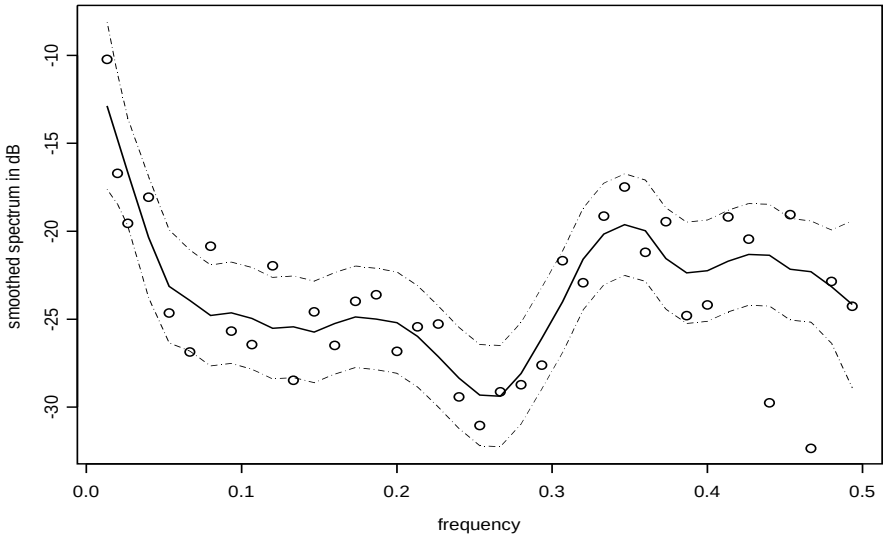


Figure 2: Spectrum estimate.

based on loess (locally linear) regression, namely its robust version, [3], with unequal weighting of $\hat{\gamma}_k$'s, according to their asymptotic variances obtained as a byproduct of the MLE fitting procedure.

From there, we can see that the raw estimates fall reasonably within the confidence limits (although these are necessarily too narrow to be thought of as simultaneous confidence bands due to their construction which guarantees only pointwise and not simultaneous coverage), except for estimates at two large frequencies. The overall shape of the spectrum is of non-trivial shape. Especially low frequencies are prominent (most likely remnants of some long-term trend that is not estimated separately in model 2 and hence is confounded with long-periodic trigonometric part). Frequencies around 0.25 (periods of about four years) seem to contribute less to the height series changes. On the other hand, high frequency part of the spectrum is presented substantially (especially periods shorter than, say 3 years). This is interesting, since this picture resembles results of [11], who did spectral analysis on another height data (not truncated and collected in a completely different way, at different times and locations). To get additional checks of the results concerning overall spectral shape, we re-estimated it under several alternative model modifications in the sensitivity analysis style. We have tried: i) both time-varying and constant expert estimates of MHR, ii) free and σ^2-restricted models ($\sigma^2 = \sigma_0^2$ with externally expert-supplied σ_0^2 – an approach that has been advocated in the past, [1]) as a way to circumvent correlation between mean-related and scale parameter estimates introduced by truncation, iii) combination of left truncation and additional interval censoring that can be suspected in connection with rounding, iv) simultaneous left and right truncation (to assess the possibility of over-representation of extremely tall men in the military sample in addition to the MHR complication), v) addition of quadratic trend, vi) omission of any trend whatsoever. Variants i) through vi) influenced the non-trigonometric part of the model and absolute values of the spectrum to various extent. Nevertheless, the spectrum shape remained remarkably similar and insensitive to the model perturbations considered.

In general, we note that while intercept-like coefficients (μ, α_1, α_2) are rather difficult to estimate precisely, the slope-type coefficients (β, δ's) are estimated much more precisely. This is because the likelihood surface has a near-ridge e.g. in the μ-σ^2 plane (see Figure 3 for a situation with one particular birhtyear-year of 1751). Consequently, periodicity properties can be appreciated much better than intercepts determining average height per se. Vertical shift is more uncertain and hence it is much harder to answer questions about "absolute height", compared to questions about shape of their changes in time.

Apart from overall (smoothed) spectrum estimate (which was required by the customer for explorative purposes), one can think of testing individual harmonic terms contributions (e.g. $H_0 : \delta_{1k} = \delta_{2k} = 0$). For simplicity, the screening tests can be performed as Wald tests (using asymptotic variance-

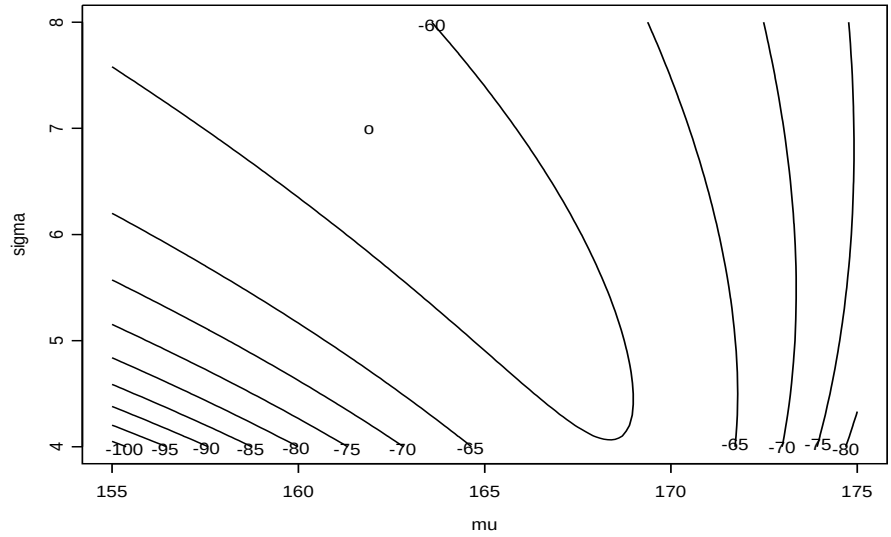


Figure 3: Likelihood surface example.

covariance matrix of parameter estimates). Only few components were significant, namely those corresponding to periods of 75, 50, 25 and 2.885 years. Resulting reduced model (with full non-trigonometric part and four harmonic components only) can be tested against the original model via LRT, and it yields p=0.47. From this, one would get the impression, that the original model can be dramatically simplified. We do not recommend such simplification, however. Coefficient estimates for different sine and cosine terms are correlated here (unlike in the classical Fourier analysis, [2]). This lack of orthogonality is due to the fact that the data are unbalanced, that we do not work exactly with Fourier frequencies and because of truncation.

It is of practical interest that the smooth spectral density estimate $\hat{s}(f)$ can be integrated as $I_{(p_i,p_{i+1})} = \int_{\frac{1}{p_i}}^{\frac{1}{p_{i+1}}} \hat{s}(f)df$ to get some idea about amount of periodic variability in various period intervals (p_i, p_{i+1}). These can be standardized to get proportions. We have estimated proportions of variability in four sub-intervals of $(2, 15)$ years interval (which has been intensively investigated in connection with buisness cycle in the past, [11]). They are: $\frac{I_{(2,3)}}{I_{(2,15)}} = 0.58$, $\frac{I_{[3,5)}}{I_{(2,15)}} = 0.20$, $\frac{I_{[5,7)}}{I_{(2,15)}} = 0.09$, $\frac{I_{[7,10)}}{I_{(2,15)}} = 0.07$, $\frac{I_{[10,15)}}{I_{(2,15)}} = 0.06$, corresponding rather nicely to proportions estimated in [11] by different methods under different circumstances.

References

[1] A'Hearn B. (2004). *A restricted maximum likelihood estimator for truncated height samples.* Economics and Human Biology **2**, 1, 5 – 19.

[2] Brockwell P.J., Davis R.A. (1991). *Time series: Theory and methods.* Springer, New York.

[3] Cleveland W.S., Devlin S.J. (1988). *Locally-weighted regression: An approach to regression analysis by local fitting.* J. Am. Statist. Assoc. **83**, 596 – 610.

[4] Easterlin R.A. (2000). *The worldwide standard of living since 1800.* J. Econ. Perspect. **14**, 7 – 26.

[5] Floud R., Wachtel K., Gregory A. (1990). *Height, health and history. Nutritional status in the United Kingdom, 1750-1980.* Cambridge University Press. Cambridge.

[6] Heintel M. (1996). *Historical height samples with shortfall, a computational approach.* History and Computing **8**, 1, 24 – 37.

[7] Heintel M., Sandberg L., Steckel R. (1998). *Swedish historical heights revisited: New estimation techniques and results.* Proceedings of the conference The biological standard of livinging in comparative perspective. Komlos J., Baten J. (eds.) Franz Steiner Verlag. Stuttgart.

[8] Komlos J. (1989). *Nutrition and economic development in the eighteen century Habsburg monarchy: An anthropometric history.* Princeton University Press. Princeton, NJ.

[9] Komlos J. (1993). *The secular trend in the biological standard of living in the United Kingdom, 1730-1860.* Economic History review **46**, 115 – 144.

[10] Meeker W.Q., Duke S.D. (1981). *CENSOR - A user-oriented computer program for life data analysis.* The American Statistician **35**, 2, 112.

[11] Woitek U. (2003). *Height cycles in the 18th and 19th centuries.* Economic and Human Biology **2**, 145 – 288.

Address: M. Brabec, Department of Biostatistics and Computing Services, National Institute of Public Health, Šrobárova 48, 100 42 Praha 10, Czech Republic

E-mail: `mbrabec@szu.cz`

ROZUMÍ SI STATISTIKA S MEDICÍNOU?

Václav Čapek

Klíčová slova: Aplikovaná statistika, výuka.

Abstrakt: Příspěvek si klade za úkol seznámit čtenáře se zkušenostmi autora z běžné statistické praxe na půdě Lékařských fakult Univerzity Karlovy v Praze. Měl by vést k zamyšlení o tom, jak je statistika chápána, přijímána a používána těmi, kdo jsou její spotřebitelé.

1 Úvodní zamyšlení

Čtenář bude jistě souhlasit, že statistiku potřebuje spousta odborníků z různých oborů. Potřebují ji, aby mohli objektivním způsobem zpracovat a vyhodnotit měření, která v rámci své práce realizují. Před námi, statistiky, teď stojí několik otázek:

- Mají tito odborníci dostatečné znalosti statistiky?
- Mají k dispozici vhodnou statistickou literaturu?
- Mají možnost se ve statistice vzdělávat způsobem, který je pro ně optimální?
- Co jim můžeme nabídnout my, statistici?
- Chceme jim to nabídnout?
- Umíme to nabídnout?

Poslední dvě otázky tak trochu útočí do vlastních řad. Máme-li však být objektivní, musíme si položit i takové otázky.

Pokud jsme dostatečně upřímní, pak se zřejmě shodneme na tom, že bohužel na ani jednu z výše uvedených otázek nemůžeme jednoznačně odpovědět: ano, samozřejmě. A to je chyba. Tento příspěvek si klade za cíl vzbudit ve čtenáři právě tyto pochybnosti a motivovat ho k tomu, aby se i on pokusil současný stav zlepšit.

Autor tohoto textu ve svých závěrech vychází ze své praxe, kdy konzultoval, z pohledu statistiky, diplomové a dizertační práce přátel a pomáhal se zpracováním dat pro 1. LF UK a některé farmaceutické společnosti.

2 Trocha praxe

Podívejme se teď, s jakými prohlášeními je možné se ve statistické praxi setkat:

- Spočítal jsem si něco v Excelu, ale nevím, co to znamená.
- Můžeme předpokládat, že data jsou normální? Já myslím, že ano. Co to znamená?
- Aha – to jsou ty složitý vzorečky. Z toho mě málem vyhodili. To mi sem nedávejte. Já nechci nic složitého, mně stačí t-test.
- Na statistiku jsem nechodil. Nikdy mě nenapadlo, že to budu potřebovat.

Výše uvedené věty demonstrují neduhy současného stavu. Odborník nestatistik se snaží získat výsledek za každou cenu, nerozumí použité metodě, snaží se minimalizovat složitost, chce to mít co nejdříve za sebou. Navíc v době, kdy náš odborník studuje, nemá představu, k čemu mu bude statistika později dobrá. Proto si z absolvovaných přednášek skoro nic neodnáší.

Co s tímto stavem udělat? Máme dvě možnosti. Buď naučíme nestatistiky úplně celou statistiku tak, aby byli soběstační. A nebo jim ukážeme taje statistiky, necháme je lehce nahlédnout do celé její šíře a ukážeme jim možné cesty. Řekneme jim, na koho se obrátit. Věřím, že i čtenář cítí, že druhá cesta je lepší.

3 Současný stav výuky

Vyberme opět pro ilustraci několik vět, se kterými se statistik běžně setkává:

- Statistika je šílenost.
- To byly pořád jenom matice. Co to je to pé?
- Nikdy jsem nepochopil, proč to musí být vždycky 5%.
- Proč nám o těch neparametrických metodách neřekli?
- Teď, když to vidím, tak bych chodil i na víc semestrů statistiky.

Z těchto vět je patrné, že zřejmě podáváme statistiku příliš matematickou, s malým důrazem na pochopení vlastní filozofie a na aplikace. Nejspíš si plně při výkladu neuvědomujeme, jaký typ vzdělání mají naši posluchači. Asi statistiku vykládáme tak, jak bychom ji vykládali matematikům. Zamysleme se nyní pořádně nad výše uvedenými výroky a zkusme najít tu správnou formu výuky statistiky pro nestatistiky. Zkusme se vžít do jejich potřeb a zeptejme se sami sebe, co bychom se jako nestatistici chtěli dozvědět na přednášce s názvem Statistika? Bylo by to odvození maticového tvaru odhadu metodou nejmenších čtverců, nebo vyprávění o tom, kdy tento odhad, ať už se spočítá jakkoliv, má smysl použít a kdy ne?

4 Návrh osnovy výuky

Jak by tedy měla vypadat ta správná osnova výuky statistiky pro nestatistiky? Podívejme se na následující body:

- základní představa o rozdělení a jeho charakteristikách
- výjimečnost normálního rozdělení, CLV
- srovnání klasických a robustních metod
- odhady parametrů a konfidenční intervaly
- testování hypotéz
- regrese
- analýza rozptylu
- návrhy experimentů
- nezávislost
- mnohonásobná porovnávání

Tento návrh plyne ze zkušeností autora z jeho osobní praxe. Jednotlivá témata jsou seřazena dle důležitosti a návaznosti. Všimněme si relativně mohutného zapojení robustních a neparametrických metod. Důraz je kladen zejména na filozofii statistiky a na interpretaci výsledků.

Nematematikům nepomůže k pochopení problému znalost přesných matematických základů přednášené látky. Potřebují téma přiblížit jinak. Nevadí, v jejich případě, pokud se při výkladu dopouštíme drobných nepřesností. Jde o pochopení smyslu statistiky a získání přehledu o dostupných metodách a zejména o komplexnosti celého problému.

Pokud má být taková výuka úspěšná, musí zároveň probíhat i vhodná kampaň, jejímž cílem bude zvýšení motivace potenciálních posluchačů výuku úspěšně absolvovat. Kampaň musí nestatistikům vysvětlit, jaké jsou její přínosy právě pro ně a pro jejich obor. Statistika nesmí být chápána jako povinná matematika, ale jako užitečný a nenahraditelný nástroj. Statistik pak je v roli konzultanta, který má, v jakémkoliv projektu, svoje pevné místo.

5 Související znalosti

Závěrem zmiňme fakt, že od statistika, který se věnuje statistické praxi, se neočekává pouze perfektní znalost statistiky samotné. V okamžiku, kdy statistik připravuje závěrečnou zprávu, píše článek, oponenturu, připravuje klinické hodnocení, zpracovává grant nebo výběrové šetření, musí se vypořádat s uměním jazyka. Musí umět věc vysvětlit a diskutovat o ní, musí umět výsledky práce celého týmu ve své zprávě vhodně prodat. Čato se setká s tím, že ke své práci potřebuje znalost zákona. I tohle všechno patří do náplně práce statistika.

6 Závěr

Cílem tohoto článku bylo vzbudit o popisované problematice povědomí, nastínit klíčové problémy a navrhnout možná řešení. Autor si je vědom toho, že řešení neexistuje jediné, v nějakém směru optimální. Současný stav je však nevyhovující, pojďme proto hledat alespoň nějaké řešení!

Poděkování: Tato práce je podporována výzkumným záměrem MSM 113200008.

Adresa: V. Čapek, Karlova Univerzita v Praze, Katedra pravděpodobnosti a matematické statistiky, Sokolovská 83, 186 75 Praha 8 – Karlín

E-mail: `capek@karlin.mff.cuni.cz`

ZAJIŠTĚNÍ V POJIŠŤOVNICTVÍ A JEHO MATEMATICKÉ ASPEKTY

Tomáš Cipra

Abstrakt: O existenci zajištění v pojišťovnictví a o jeho fungování veřejnost příliš neví, přestože je to jeden z pilířů současného pojišťovnictví. Málokterý pojištěný u nás tuší, že značná část pojistného, které zaplatil české pojišťovně, putuje po převodu na eura nebo švýcarské franky do zahraničních zajišťoven (např. do největších světových zajišťoven Munic Re v Mnichově a Swiss Re v Curychu) a že naopak tyto zajišťovny hradí podstatnou část škody, kterou pojištěný utrpěl při pojistné události. Žádná pojišťovna u nás si nedovolí (zvlášť po povodňových zkušenostech) pracovat bez zajištění, neboť se vlastně jedná o „pojištění pojišťovny". Také výše sazeb, které pojišťovny předepisují svým klientům, se z velké míry odvíjí od situace na zajistných trzích, a to zvlášť v současném světě klimatických změn a narůstajících přírodních a společenských katastrof.

Tento příspěvek nejprve prezentuje základní principy zajištění, které se poněkud liší od principů přímého pojištění (po právní, metodologické i výpočetní stránce). Z hlediska statistiky je zde nutné zdůraznit fakt, že velké zajišťovny disponují velmi rozsáhlými a kvalitními statistickými archivy diverzifikovanými přes rozsáhlá geografická území, které často dávají po příslušném statistickém zpracování (nebo i ve zdrojové podobě) k dispozici zajišťovaným pojišťovnám. Dále se příspěvek soustřeďuje na některé matematické postupy využívané v zajištění, které mají kořeny především v teorii pravděpodobnosti a matematické statistice. Protože součástí dnešního zajištění je alternativní přenos rizik ART (Alternative Risk Transfer), který se snaží převést pojistná rizika nezvládnutelná pojišťovnami na finanční trhy, referuje příspěvek i o těchto postupech využívajících především finanční matematiku.

1 Základní pojmy a principy zajištění

Zajištění je vlastně "pojištění pojišťovny". Zajišťovná pojišťovna se v tomto kontextu obvykle označuje jako prvopojistitel a zajišťující zajišťovna jako zajistitel. Následující tabulky 1-3 uvádí některé aktuální údaje, které mají souvislost se současným stavem zajištění ve světě:

1.1 Význam zajištění

Význam zajištění spočívá mimo jiné v následujících pozitivních skutečnostech:

- zvýšení kapacity pojistitele
- homogenizace pojistného kmene
- rozprostření a diverzifikace pojistných rizik (viz obrázky 1-3)

- dosažení finančních výhod
- získání profesionálních služeb zajistitele

Událost	Datum	Oblast	Počet obětí	Pojištěná škoda (mil. USD)
Hurikán Andrew	23.08.1992	USA (Bahamy)	38	20 185
Teroristický útok v USA	11.09.2001	USA (New York aj.)	3 122	19 000
Zemětřesení v Northridge	17.01.1994	USA (Kalifornie)	60	16 720
Tajfun Mireille	27.09.1991	Japonsko	51	7 338
Větrná smršť Daria	25.01.1990	Francie, UK aj.	95	6 221
Větrná smršť Lothar	25.12.1999	Francie, Švýcarsko aj.	80	6 164
Hurikán Hugo	15.09.1989	Portoriko, USA aj.	61	5 990
Záplavy a bouře (záp. Evropa)	15.10.1987	Francie, UK aj.	22	4 674
Větrná smršť Vivian	25.02.1990	západní a střední Evropa	64	4 323
Tajfun Bart	22.09.1999	Japonsko	26	4 293

Pramen: Sigma 2002, No. 1

Tabulka 1: Deset celosvětově největších pojištěných škod katastrofického charakteru za období 1970-2001 v mil. USD indexovaných k roku 2001 (s vyloučením odpovědnostních škod).

Typ katastrofy	Počet katastrof	Počet obětí	Pojištěná škoda (mil. USD)
Přírodní katastrofy	111	22 803	10 010
Velké požáry a výbuchy	40	921	3 748
Letecké katastrofy	17	785	1 094
Lodní katastrofy	22	1 609	
Silniční a železniční katastrofy	75	2 061	
Důlní neštěstí	18	959	68
Zřícení budov a mostů	5	156	
Terorismus, sociální nepokoje	4	3 165	19 398
Jiné velké katastrofy	23	591	74
Celkem	315	33 050	34 392

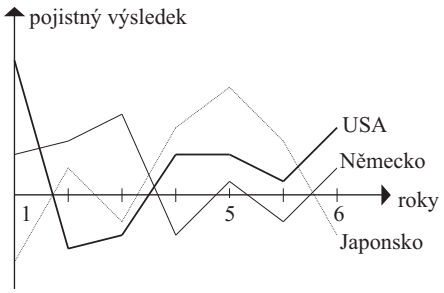
Pramen: Sigma 2002, No. 1

Tabulka 2: Pojištěné škody katastrofického charakteru podle škodních kategorií globálně za rok 2001.

Zajistitel		Předepsané zajistné (mil. USD)	Retro-cese (%)	Škodní průběh (%)	Náklad. koeficient (%)
Munich Re	NP	13 072,3	17,0	104,5	30,6
	ŽP	3 882,5	10,0	NA	23,3
Swiss Re	NP	10 265,7	10,2	95,0	29,0
	ŽP	5 428,0	7,5	NA	4,7
General Cologne Re	NP	5 830,0	8,9	133,9	26,4
	ŽP	2 005,0	7,3	81,8	22,3
Lloyd's	NP	5 743,6	9,0	NA	NA
	ŽP				
GE Global	NP	5 551,0	30,2	101,6	38,9
	ŽP	1 841,0	23,8	84,2	32,9
Hannover Re	NP	4 837,9	41,4	99,4	16,3
	ŽP	1 579,6	26,3	NA	NA
Gerling Global Re	NP	3 462,3	13,7	109,2	25,9
	ŽP	1 037,4	18,8	64,2	22,3
AXA Corp.Solutions	NP	3 294,8	35,9	97,5	29,6
	ŽP				
Berkshire Hathaway	NP	2 953,0	1,0	117,0	5,0
	ŽP				
SCOR	NP	2 809,2	18,0	100,0	29,0
	ŽP	493,7	16,0	83,0	27,0

Pramen: Reinsurance 33, 2002, No. 3

Tabulka 3: Deset celosvětově největších zajistitelů za rok 2001 seřazených podle předepsaného zajistného v mil. USD pro neživotní pojištění po odečtení retrocese.

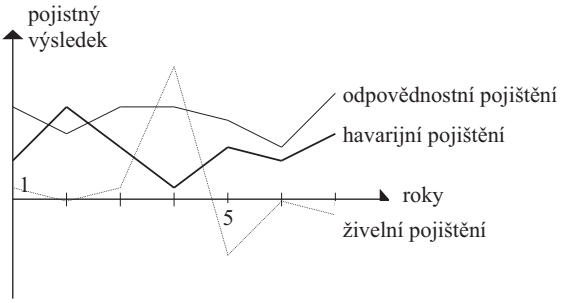


Obrázek 1: Teritoriální diverzifikace pojistných výsledků pomocí zajištění.

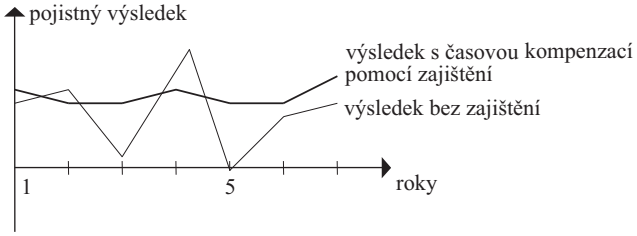
1.2 Právní aspekty zajištění

Jako základní právní principy zajištění se uvádí:

- *princip odškodnění* (indemnity): odškodňuje se jednoznačná finanční ztráta, kterou utrpěl prvopojistitel (tj. zajišťovaná pojišťovna);



Obrázek 2: Produktová diverzifikace pojistných výsledků pomocí zajištění.



Obrázek 3: Časová diverzifikace pojistných výsledků pomocí zajištění.

- *princip dobré víry* (utmost good faith): zajistná smlouva má do jisté míry charakter "džentlmenské dohody" spoléhající na serióznost smluvních partnerů, tj. prvopojistitele (neboli zajišťované pojišťovny) a zajistitele (neboli zajišťující zajišťovny);
- *princip smluvního společenství zájmů* (privity of contract): prvopojistitel zůstává ve vztahu k původnímu pojištěnému za dané riziko plně odpovědný (zajistná smlouva je právně zcela oddělena od původní pojistné smlouvy mezi klientem pojišťovny a pojišťovnou).

Velmi důležitou právní úlohu hrají také tzv. klauzule zajistných smluv, které jsou specialitou zajistných vztahů:

(1) *n-hodinová klauzule* (hour clause; n-Stundenklausel):

  – používá se při potenciálních přírodních nebo společenských katastrofách s delší škodní expozicí a znamená, že u dané škodní události jsou kryty pouze ty škody, které se kumulují nejdéle během intervalu délky n hodin;

  – obvykle je:

    - 48 hodin: pro vichřice a hurikány;

  - 72 hodin: pro zemětřesení (včetně mořských), rozsáhlé požáry (např. celých území), vulkanické erupce, politická rizika (např. občanské nepokoje);
  - 168 hodin: pro povodně a záplavy;
  - 504 hodin: pro záplavy;
- důležitá jsou v této souvislosti také upřesnění, zda
  - pojištěná rizika různého typu budou spadat pod jednu katastrofu (např. záplavy způsobené hurikánem);
  - daná katastrofa je omezena také územně (např. povodím řeky při povodních, územím města při nepokojích apod.).

(2) *Klauzule: věcný rozsah zajistného krytí:*

- *All Risks* zajištění kryje všechna rizika kromě těch, která jsou explicitně uvedena jako výluky (tj. "co není vyloučeno, je automaticky zajištěno");
- *Named Perils* zajištění kryje jen explicitně uvedená rizika (tj. "co není uvedeno, je automaticky ze zajištění vyloučeno").

(3) *Klauzule: indexace:*

- v případech, kdy zajistný vztah trvá delší dobu (několik let), se může během takové doby v důsledku inflace značně zvýšit cenová úroveň škod a při neměnné výši priority by se zvýšilo škodní zatížení zajistitele; proto se provádí indexovaní;
- *výluky ze zajištění:* jejich důvodem může být:
  - stejná výluka v pojistných podmínkách prvopojistitele;
  - problém pojistitelnosti (např. jaderná nebo válečná rizika včetně teroristických činů)
  - potenciální překročení kapacity zajistitele (např. ekologická rizika včetně kontaminace radonem či azbestem).

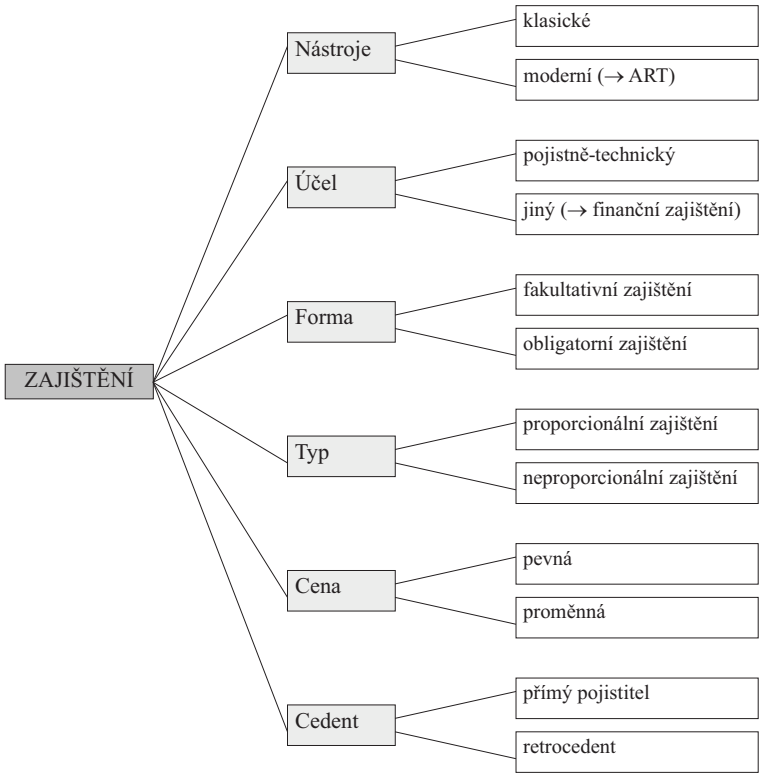
(4) *Klauzule: sdílení osudu a jednání:*

- Zajistitel je podřízen:
  - stejným vnějším podmínkám ovlivňujícím průběh pojištění jako prvopojistitel (např. klimatickým podmínkám);
  - všem rozhodnutím a jednáním, které prvopojistitel v rámci daného pojištění provádí (např. uzavírání, odmítání, změny a výpovědi pojistných smluv, stanovení a změny pojistných podmínek, kalkulace pojistného, správa pojištění, likvidace škod aj.).

(5) *Klauzule: arbitráž:*

- případné spory mezi smluvními stranami v zajištění řeší rozhodčí komise.

2 Klasifikace zajištění



Obrázek 4: Klasifikace zajištění.

2.1 Formy zajištění: fakultativní a obligatorní

Fakultativní zajištění: prvopojistitel a zajistitel zvažují situaci případ od případu, přičemž prvopojistitel není smluvně povinen příslušnou pojistnou smlouvu k zajištění nabídnout a zajistitel není smluvně povinen ji k zajištění přijmout.

Obligatorní zajištění: při splnění podmínek z rámcové zajistné smlouvy (reinsurance treaty) má zajistitel právo a zároveň povinnost převzít příslušné části rizika z jednotlivých pojistných smluv daného portfolia; přitom ve smyslu principu dobré víry: zajistitel věří, že prvopojistitel bude postupovat při uzavírání zajišťovaného pojistného obchodu a jeho správě kvalifikovaně a vůči zajistiteli spravedlivě, zatímco prvopojistitel se spoléhá na promptní plnění zajistitele v případě potřeby.

2.2 Typy zajištění

Proporcionální zajištění: pojistná částka, pojistné plnění a pojistné se zde dělí mezi prvopojistitele a zajistitele ve sjednaném poměru. V praxi se nejčastěji využívají dva typy proporcionálního zajištění:

- *kvótové zajištění:* poměr pro dělení rizika mezi prvopojistitele a zajistitele je pro každou pojistnou smlouvu stejný;
- *surplus* (nebo *excedentní zajištění*): prvopojistitel ceduje v každé pojistné smlouvě jen tu část rizika, která přesahuje pevně sjednanou hodnotu stejnou pro všechny pojistné smlouvy (odtud ovšem plyne, že na rozdíl od kvótového zajištění poměr pro dělení rizika mezi prvopojistitele a zajistitele může být pro každou pojistnou smlouvu jiný).

Ze statistického hlediska je u proporcionálního zajištění pravděpodobnostní rozdělení pojistného plnění ponechaného prvopojistitelem na jeho vlastní vrub stejné jako před cesí až na jiné měřítko:

Jestliže X_P označuje pojistné plnění ponechané prvopojistitelem z původní hodnoty X před cesí na základě proporcionálního vztahu

$$X_P = \alpha \cdot X$$

pak pro odpovídající pravděpodobnostní hustoty f_P a f náhodných veličin X_P a X platí

$$f_P(x) = f(x/\alpha)/\alpha$$

Dojde přitom sice k proporcionálnímu zmenšení střední hodnoty a směrodatné odchylky původního pojistného plnění

$$E(X_P) = \alpha \cdot E(X) \quad \sigma(X_P) = \alpha \cdot \sigma(X)$$

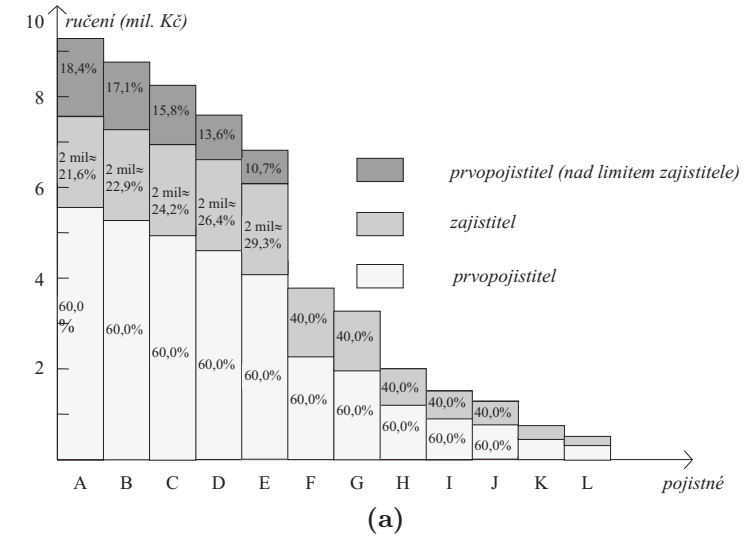
ale nezmění se příslušný variační koeficient

$$\frac{\sigma(X_P)}{E(X_P)} = \frac{\sigma(X)}{E(X)}$$

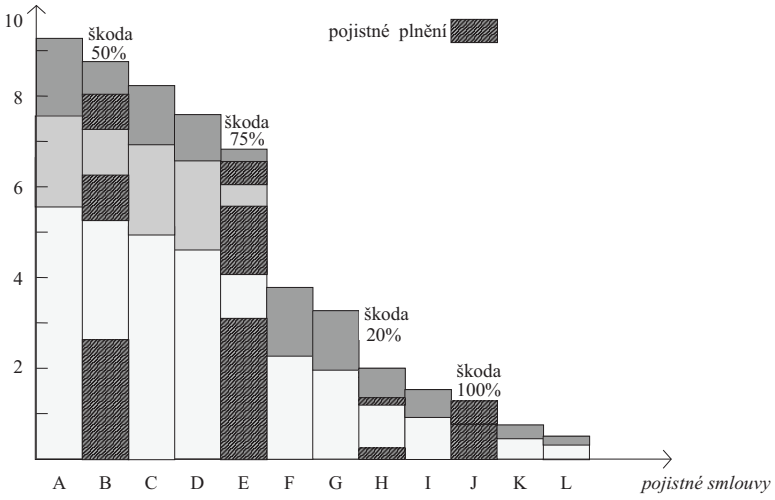
(tj. nezmění se relativní fluktuace pojistného plnění ponechaného na vlastní vrub). Z praktického hlediska se ovšem pro prvopojistitele vylepší jeho solventnostní pozice, neboť v absolutním vyjádření pojistné plnění na jeho vlastní vrub klesne, ale nezmění se jeho volný kapitál důležitý právě pro výkaz solventnosti.

U proporcionálního zajištění není shora omezena výše plnění prvopojistitele, a proto nemusí být dostatečnou ochranou proti vysokým škodám. Z toho důvodu se v takových případech používá neproporcionální zajištění:

Neproporcionální zajištění (nebo *škodové zajištění*): zajistitel za speciálně stanovené zajistné zde přebírá po vzniku škody tu část pojistného plnění prvopojistitele, která přesáhne sjednaný vlastní vrub prvopojistitele nazývaný *priorita:*



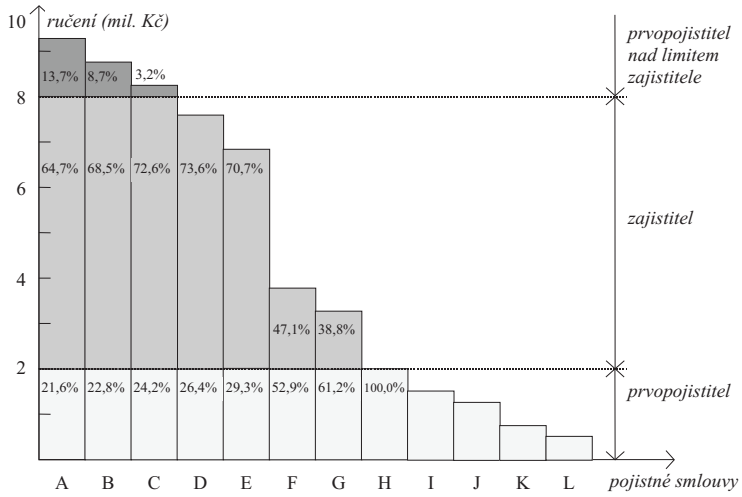
(a)



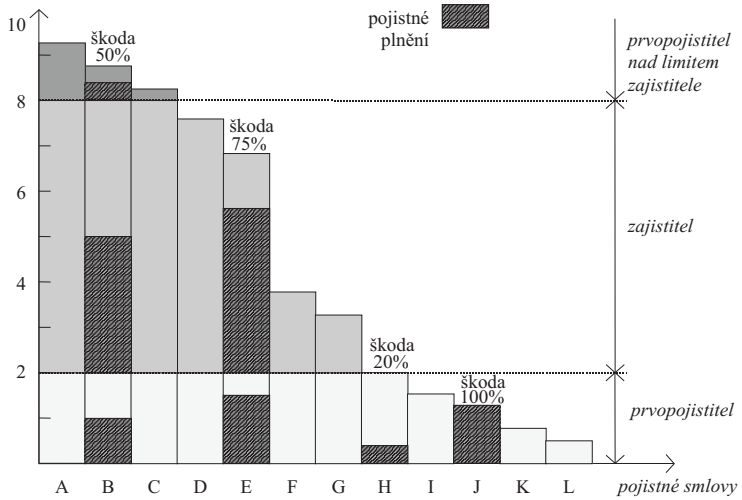
(b)

Obrázek 5: Příklad kvótového zajištění (kvóta zajistitele $q = 40\%$ a limit zajistitele $L = 2$ mil. Kč): (a) ručení; (b) pojistné plnění.

- dochází zde tedy k omezení výše plnění prvopojistitele shora a nikoli ad hoc k proporcionálnímu dělení odpovědnosti mezi prvopojistitele a zajistitele;
- plnění zajistitele je z toho důvodu určováno výhradně výší skutečně vzniklých škod přesahujících prioritu. V praxi se opět nejčastěji využívají dva typy neproporcionálního zajištění;



(a)



(b)

Obrázek 6: Příklad surplusu (vlastní vrub prvopojistitele $s = 2$ mil. Kč a limit zajistitele ve výši tří maxim, tj. $L = 6$ mil. Kč): (a) ručení; (b) pojistné plnění.

- *XL zajištění* (nebo *zajištění škodního nadměrku*): pevně sjednaná priorita se v souvislosti s dalším členěním XL zajištění uplatňuje buď zvlášť pro jednotlivé pojistné smlouvy, nebo souhrnně pro více pojistných smluv zasažených současně nějakou katastrofickou událostí s kumulací škod;

- *SL zajištění* (nebo *zajištění ročního nadměrku*): spoluúčast prvopojistitele se zde uplatňuje v rámci celoročního objemu škod a má často tvar mezní hranice pro škodní průběh, nad níž zajistitel plní.

Ze statistického hlediska dochází u neproporcionálního zajištění k podstatné redukci fluktuací (měřených směrodatnou odchylkou) všech hodnot ponechaných prvopojistitelem na jeho vlastní vrub: jestliže označuje a prioritu prvopojistitele a L limit zajistitele, pak např. střední hodnota pojistného plnění prvopojistitelem má tvar

$$E(X_P) = \int_0^a x \cdot f(x)dx + a \cdot \int_a^{a+L} f(x)dx + \int_{a+L}^{\infty} (x - L) \cdot f(x)dx$$

V praxi se provádí podrobnější členění neproporcionálního zajištění:

(1) *WXL/R zajištění* (nebo *zajištění škodního nadměrku jednotlivých rizik*) zajišťuje prvopojistitele proti jednotlivým (velkým) škodám:
   - je-li nějaká (jednotlivá) pojistná smlouva ze zajišťovaného portfolia postižena pojistnou událostí s pojistnými nároky převyšujícími prioritu prvopojistitele, pak vzniklý nadměrek hradí zajistitel (ale jen do výše jeho vrstvy)

$$X_Z = \begin{cases} 0 & \text{pro} \quad X \le a\,, \\ X - a & \text{pro} \quad X > a \end{cases}$$

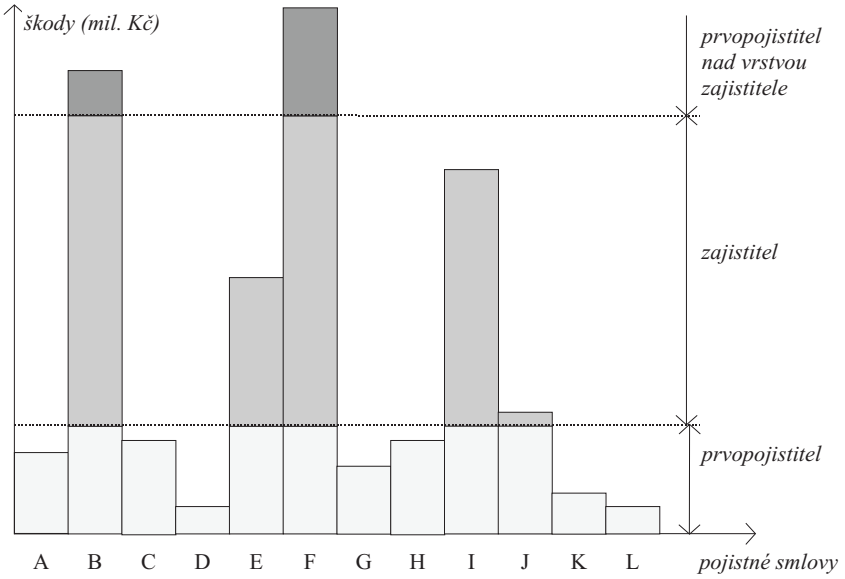
     kde $a (a > 0)$ je priorita prvopojistitele a X_Z označuje pojistné plnění zajistitele (tj. zajistné plnění) z původního pojistného plnění X;
   - v praxi se pak někdy používá zápis typu

$$3 \text{ mil. Kč } xs \text{ } 0,5 \text{ mil. Kč}$$

     kde 3 mil. Kč je vrstva zajistitele a 0,5 mil. Kč je priorita prvopojistitele.

(2) *WXL/E zajištění* (nebo *zajištění škodního nadměrku jednotlivých událostí*) zajišťuje prvopojistitele proti kumulaci škod vzniklých vždy v důsledku jedné škodní události, která zde ale ještě nemá charakter přírodní katastrofy (např. úrazové nebo cestovní pojištění účastníků autobusového zájezdu, požární pojištění bytového družstva):
   - je-li několik pojistných smluv ze zajišťovaného portfolia postiženo jednou škodní událostí s celkovými pojistnými nároky převyšujícími prioritu prvopojistitele, pak vzniklý nadměrek hradí zajistitel

$$X_Z = \begin{cases} 0 & \text{pro} \quad \displaystyle\sum_{i=1}^{n} X_i \le a\,, \\ \displaystyle\sum_{i=1}^{n} X_i - a & \text{pro} \quad \displaystyle\sum_{i=1}^{n} X_i > a \end{cases}$$



Obrázek 7: Příklad WXL/R zajištění (uvedeny jsou jen ty pojistné smlouvy ze zajišťovaného portfolia, ve kterých došlo k pojistným událostem).

kde opět $a(a > 0)$ je priorita prvopojistitele, X_Z je zajistné plnění, $X_1, \ldots, X_n$ je soubor pojistných plnění z dané škodní události v n postižených pojistných smlouvách;

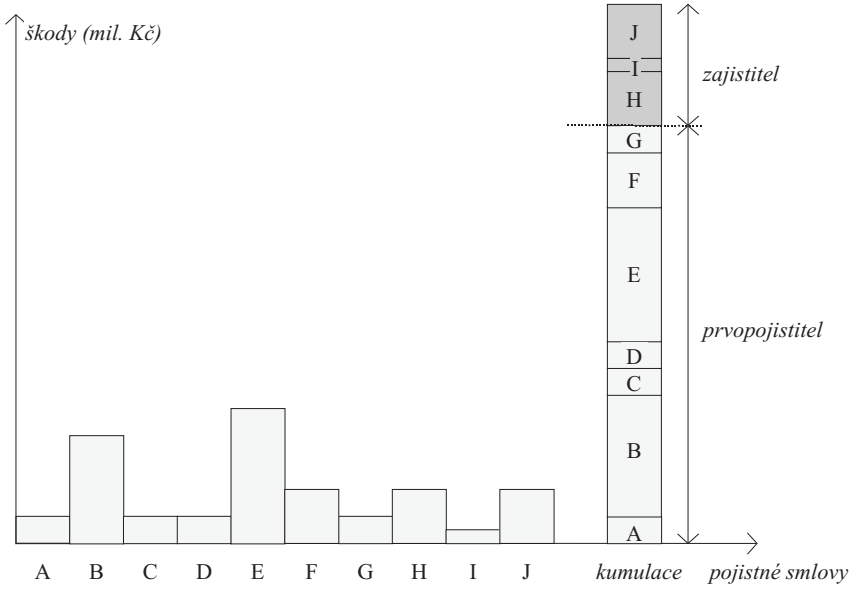
– pro aktivaci WXL/E je nutná expozice jedné škodní události ve více pojistných smlouvách (např. v úrazovém nebo životním pojištění se předepíše minimální počet osob, jejichž postižení danou škodní událostí je pro aktivaci zajištění nutné).

(3) *CatXL zajištění* (nebo *zajištění škodního nadměrku katastrofické události*) se shoduje se zajištěním WXL/E až na katastrofický charakter škodní události, v jejímž důsledku obvykle dochází k podstatné kumulaci škod.

(4) *SL zajištění* (nebo *zajištění ročního nadměrku* nebo *časového nadměrku*): priorita prvopojistitele se zde uplatňuje v rámci celoročního objemu škod a má často tvar mezní hranice pro škodní průběh (tj. pro poměr pojistného plnění vůči pojistnému), nad níž zajistitel plní do sjednaného limitu

$$X_Z = \begin{cases} 0 & \text{pro} \quad X/P \leq p\,, \\ X - p \cdot P & \text{pro} \quad p < X/P \leq l\,, \\ (l-p) \cdot P & \text{pro} \quad l < X/P\,, \end{cases}$$

kde $p(p > 0)$ je priorita prvopojistitele, $l(l > 0)$ je limit zajistitele a X_Z označuje zajistné plnění.



Obrázek 8: Příklad CatXL zajištění (uvedeny jsou jen ty pojistné smlouvy ze zajišťovaného portfolia, které byly postiženy příslušnou katastrofickou událostí).

(5) *Zajištění nejvyšších škod* (*Largest Claims Reinsurance* LCR(p)) znamená, že zajistitel hradí p největších škod (p je dané přirozené číslo, $p < n$), které nastaly během platnosti zajistné smlouvy:

$$X_Z = X_{(1)} + X_{(2)} + \cdots + X_{(p)}$$

kde $X_{(1)} \geq X_{(2)} \geq \cdots \geq X_{(p)} \geq \cdots \geq X_{(n)}$ jsou škody $X_1, X_2, \ldots, X_n$ z daného roku uspořádané podle velikosti a X_Z označuje zajistné plnění.

(6) *ECOMOR zajištění (excédent du cout moyen relatif)* znamená, že zajistitel v daném roce hradí jen ty části škod, které přesáhly p-tou největší škodu (p je opět dané přirozené číslo, $p < n$):

$$\begin{aligned} X_Z &= (X_{(1)} - X_{(p)}) + \cdots + (X_{(p-1)} - X_{(p)}) = \\ &= X_{(1)} + \cdots + X_{(p-1)}) - (p-1) \cdot X_{(p)} \end{aligned}$$

3 Pojistná matematika v zajištění

Budeme používat následující značení:

N_P, S_P, X_P, Z_P	počet pojistných událostí v zajišťovaném portfoliu, jednotlivá pojistná částka, jednotlivé pojistné plnění, celkové pojistné plnění v zajišťovaném portfoliu, a to vždy na vrub prvopojistitele

N_Z, S_Z, X_Z, Z_Z mají stejný význam, ale na vrub zajistitele

${}_XF(x) = P(X \le x)$: distribuční funkce pojistného plnění X

${}_ZF(z) = P(Z \le z)$: distribuční funkce celkového poj. plnění Z

${}_SF(s) = P(S \le s)$: distribuční funkce pojistné částky S

${}_{\check{S}S}F(\chi) = P(\check{S}S \le \chi)$: distribuční funkce škodního stupně $\check{S}S = X/S$

a podobně pro pravděpodobnostní hustoty: např.

${}_Xf_Z(x)$ pravděpodobnostní hustotu náhodné veličiny X_Z apod.

3.1 Kvótové zajištění (s kvótou q)

$$\begin{aligned} N = N_P &= N_Z \\ X_P = (1-q)\cdot X\,, &\quad X_Z = q\cdot X \\ Z_P = (1-q)\cdot Z\,, &\quad Z_Z = q\cdot Z \end{aligned}$$

$$\begin{aligned} E(X_P) = (1-q)\cdot E(X)\,, &\quad E(X_Z) = q\cdot E(X) \\ \mathrm{var}\,(X_P) = (1-q)^2\cdot \mathrm{var}\,(X)\,, &\quad \mathrm{var}\,(X_Z) = q^2\cdot \mathrm{var}\,(X) \\ {}_XF_P(x) = {}_XF(x/(1-q))\,, &\quad {}_XF_Z(x) = {}_XF(x/q) \\ {}_Xf_P(x) = \frac{{}_Xf(x/(1-q))}{1-q}\,, &\quad {}_Xf_Z(x) = \frac{{}_Xf(x/q)}{q} \end{aligned}$$

3.2 Surplus

$$\alpha = \begin{cases} 1 & \text{pro} \quad S \le s \\ \dfrac{s}{S} & \text{pro} \quad S > s \end{cases}$$

$$\begin{aligned} X_P = \alpha\cdot X\,, &\quad X_Z = (1-\alpha)\cdot X \\ E(X_P) = \alpha\cdot E(X)\,, &\quad E(X_Z) = (1-\alpha)\cdot E(X) \\ \mathrm{var}\,(X_P) = \alpha^2\cdot \mathrm{var}\,(X)\,, &\quad \mathrm{var}\,(X_Z) = (1-\alpha)^2\cdot \mathrm{var}\,(X) \\ \frac{\sigma(X_P)}{E(X_P)} &= \frac{\sigma(X_z)}{E(X_z)} = \frac{\sigma(X)}{E(X)} \end{aligned}$$

$$\begin{aligned} {}_XF_P(x) = {}_XF(x/\alpha)\,, &\quad {}_XF_Z(x) = {}_XF(x/(1-\alpha)) \\ {}_Xf_P(x) = \frac{{}_Xf(x/\alpha)}{\alpha}\,, &\quad {}_Xf_Z(x) = \frac{{}_Xf(x/(1-\alpha))}{(1-\alpha)} \end{aligned}$$

(je-li $\alpha = 1$, pak zřejmě $X_Z = 0$ a některé předchozí vztahy je nutné modifikovat).

Za předpokladu nezávislosti náhodných veličin S a $\check{S}S$:

$$\begin{aligned} {}_XF_P(x) &= P(X_P \le x) = \int_0^s {}_{\check{S}S}F\left(\frac{x}{u}\right) d\,{}_SF(u) + {}_{\check{S}S}F\left(\frac{x}{s}\right)\,(1-{}_SF(s))\,, \\ E(X_P^j) &= E(\check{S}S^j)\cdot\left\{\int_0^s u^j d\,{}_SF(u) + s^j\cdot(1-{}_SF(s))\right\} = \\ &= E(\check{S}S^j)\cdot E(S^j)\cdot {}_SF^{(j)}(s) \end{aligned}$$

kde

$${}_SF^{(j)}(s) = \frac{\int_0^s u^j d\,{}_SF(u) + s^j\cdot(1-{}_SF(s))}{E(S^j)}$$

Pokud má navíc náhodná veličina Z složené Poissonovo rozdělení $CP(\lambda,\,{}_XF)$ (tj. speciálně $N\sim P(\lambda)$):

$$\begin{aligned} E(Z_P) &= \lambda\cdot E(X_P) = \lambda\cdot E(\check{S}S)\cdot E(S)\cdot {}_SF^{(1)}(s) = E(Z)\cdot {}_SF^{(1)}(s) \\ \mathrm{var}\,(Z_P) &= \lambda\cdot E(X_P^2) = \lambda\cdot E(\check{S}S^2)\cdot E(S^2)\cdot {}_SF^{(2)}(s) = \\ &= \mathrm{var}\,(Z)\cdot {}_SF^{(2)}(s) \end{aligned}$$

3.3 XL zajištění

Pro jednoduchost uvažujme pouze WXL/R zajištění s prioritou a:

$$N_P = N$$

$$\begin{aligned} P(N_Z=n) &= \sum_{i=n}^{\infty} P(N=i)\cdot\binom{i}{n}\cdot p_a^n\cdot(1-p_a)^{i-n} \\ & E(N_Z) = p_a\cdot E(N) \\ \mathrm{var}\,(N_Z) &= p_a\cdot(1-p_a)\cdot E(N) + p_a^2\cdot\mathrm{var}\,(N) \end{aligned}$$

kde

$$p_a = P(X>a) = 1-{}_XF(a)$$

Speciálně při $N\sim P(\lambda)$:

$$\begin{aligned} P(N_Z=n) &= e^{-\lambda\cdot p_a}\cdot\frac{(\lambda\cdot p_a)^n}{n!} \\ E(N_Z) &= p_a\cdot\lambda \\ \mathrm{var}\,(N_Z) &= p_a\cdot\lambda \end{aligned}$$

a při $N\sim NB(\alpha,p)$:

$$
\begin{aligned}
P(N_Z = n) &= \binom{\alpha + n - 1}{n} \left(\frac{p}{p + p_a \cdot (1-p)}\right)^{\alpha} \left(1 - \frac{p}{p + p_a \cdot (1-p)}\right)^{n} \\
& X_P = \min(a, X), \quad X_Z = \max(0, X - a) \\
{}_X F_P(x) &= \begin{cases} {}_X F(x) & \text{pro} \quad x < a\,, \\ 1 & \text{pro} \quad x \geq a\,, \end{cases} \qquad {}_X F_Z(x) = {}_X F(a + x) \\
E(X_P) &= \int_0^a x d\,{}_X F(x) + a \cdot (1 - {}_X F(a)) = \int_0^a (1 - {}_X F(x))\, dx \\
E(X_Z) &= \int_a^\infty x d\,{}_X F(x) - a \cdot (1 - {}_X F(a)) = \\
&= \int_a^\infty (1 - {}_X F(x))\, dx = E(X) - E(X_P) \\
\mathrm{var}\,(X_P) &= 2 \int_0^a x \cdot (1 - {}_X F(x))\, dx - [E(X_P)]^2 \\
\mathrm{var}\,(X_Z) &= 2 \left\{ \int_a^\infty x\,(1 - {}_X F(x))\, dx - a \int_a^\infty (1 - {}_X F(x))\, dx \right\} - \\
& - [E(X_Z)]^2
\end{aligned}
$$

Obecně pro j-té momenty náhodných veličin X_P a X_Z platí:

$$
\begin{aligned}
E(X_P^j) &= \int_0^a x^j d\,{}_X F(x) + a^j \cdot (1 - {}_X F(a)) \\
E(X_Z^j) &= \sum_{i=1}^{j} \binom{j}{i} \cdot (-a)^{j-i} \cdot [E(X^i) - E(X_P^i)] \\
& Z_P = \sum_{i=1}^{N} X_P, \quad Z_Z = \sum_{i=1}^{N} X_Z
\end{aligned}
$$

Při obvyklých předpokladech v kolektivním modelu rizika:

$$
\begin{aligned}
E(Z_P) &= E(N) \cdot E(X_P) \\
\mathrm{var}\,(Z_P) &= E(N) \cdot \mathrm{var}\,(X_P) + \mathrm{var}\,(N) \cdot [E(X_P)]^2
\end{aligned}
$$

a analogicky pro náhodnou veličinu Z_Z.

Pro variační koeficient je obvykle (v nepatologických případech)

$$\frac{\sigma(Z_P)}{E(Z_P)} < \frac{\sigma(Z)}{E(Z)} < \frac{\sigma(Z_Z)}{E(Z_Z)}$$

tj. u prvopojistitele dochází k redukci variačního koeficientu (vlastní vrub ponechaný prvopojistitelem je stabilnější než původní nezajištěné portfolio), u zajistitele je tomu naopak

$$
\begin{aligned}
E(Z_P) &= E(Z) \cdot {}_X F^{(1)}(a) \\
\mathrm{var}\,(Z_P) &= \mathrm{var}\,(Z) \cdot {}_X F^{(2)}(a)
\end{aligned}
$$

kde

$$ {}_XF^{(j)}(a) = \frac{\int_0^a x^j d\,{}_XF(x) + a^j \cdot (1 - {}_XF(a))}{E(X^j)} $$

3.4 SL zajištění

$$ Z_P = \min(p \cdot P, X), \quad Z_Z = \max(0, X - p \cdot P) $$

kde priorita p představuje omezení pro škodní průběh X/P. Dále

$$ \begin{aligned} E(Z_P) &= E(Z) \cdot {}_ZF^{(1)}(p \cdot P) \\ \mathrm{var}\,(Z_P) &= \mathrm{var}\,(Z) \cdot {}_ZF^{(2)}(p \cdot P) \end{aligned} $$

kde

$$ {}_ZF^{(j)}(p \cdot P) = \frac{\int_0^{p \cdot P} z^j d\,{}_ZF(z) + (p \cdot P)^j \cdot (1 - {}_ZF(p \cdot P))}{E(Z^j)} $$

4 Výpočet zajistného

4.1 Model založený na Paretově rozdělení

Paretovo rozdělení je vhodné pro modelování výše škod. Přitom v zajišťovací praxi je častá následující situace:

Úkol: stanovit nettozajistné pro neproporcionální zajištění s prioritou prvopojistitele a a vrstvou (limitem) zajistitele L.

Data k dispozici: škody z minulých let překračující vhodně nastavenou hodnotu *OP (observation point)*, kde *OP* je mnohem menší než budoucí priorita a (vzhledem k nízké hodnotě *OP* je statistický vzorek takových škod mnohem bohatší).

Výše škody X_a nad hodnotou a se modeluje Paretovým rozdělením s pravděpodobnostními charakteristikami:

$$ \begin{aligned} f_a(x) &= \frac{b \cdot a^b}{x^{b+1}}, \quad x \geq a \\ F_a(x) &= 1 - \left(\frac{a}{x}\right)^b, \quad x \geq a \\ E(X_a) &= \frac{a \cdot b}{b-1} \quad \text{pro } b > 1 \\ \mathrm{var}\,(X_a) &= \frac{a^2 \cdot b}{(b-1)^2(b-2)} \quad \text{pro } b > 2 \end{aligned} $$

($b > 0$ je parametr, který musí být odhadnut z dat). Zajistitel bude v roce zajistné smlouvy plnit nad prioritou a jen do výše vrstvy L, takže střední

výše jeho plnění *EXL (expected XL)* bude

$$\begin{aligned} EXL &= \int_a^{a+L} (x-a)\cdot f_a(x)dx + \int_{a+L}^{\infty} L\cdot f_a(x)dx = \\ &= \begin{cases} \dfrac{a}{1-b}\cdot(RL^{1-b}-1) & \text{pro} \quad b\neq 1 \\ a\cdot \ln RL & \text{pro} \quad b=1 \end{cases} \end{aligned}$$

kde RL je relativní délka vrstvy (*relative layer*)

$$RL = \frac{a+L}{a}$$

Pro výpočet celkové výše plnění zajistitele nutný dále průměrný počet škod $LF(a)$ (*loss frequency*), které v zajišťovaném portfoliu během roku překročí prioritu a (z minulých dat jsme totiž schopni spolehlivě odhadnout jen průměrný počet (aktualizovaných) škod $LF(OP)$ nad hodnotou OP; pro mnohem větší prioritu a je obvykle pozorovaná hodnota $LF(a)$ vzhledem k malému počtu škod překračujících a značně nespolehlivá):

$$LF(a) = LF(OP)\cdot P(X_{OP} > a) = LF(OP)\cdot(1-F_{OP}(a)) = LF(OP)\cdot\left(\frac{OP}{a}\right)$$

Hledané nettozajistné NP_Z by mělo odpovídat zajistnému plnění:

$$NP_Z = LF(a)\cdot EXL = \begin{cases} LF(OP)\cdot OP^b\cdot\dfrac{a^{1-b}}{1-b}\cdot(RL^{1-b}-1) & \text{pro } b\neq 1 \\ LF(OP)\cdot OP\cdot\ln RL & \text{pro } b=1 \end{cases}$$

kde a, L a OP jsou dané hodnoty, $LF(OP)$ a b jsou odhadnuté hodnoty z minulých (aktualizovaných) dat.

Odhad parametru b se realizuje pomocí dvou přístupů:

(1) Aplikuje se ad hoc hodnota ověřená praktickými zkušenostmi s podobnými zajišťovanými portfolii. Např. WXL/R zajištění požárních včetně živelních rizik mívá b v rozmezí od 1,0 do 2,5 (speciálně pro pojištění průmyslových rizik kolem 1,2 a pro pojištění majetku obyvatelstva od 1,8 do 2,5), CatXL zajištění mívá b kolem 1,0 (speciálně pro riziko zemětřesení kolem 0,8 a pro riziko vichřic v Evropě kolem 1,3).

(2) Použije se hodnota parametru odhadnutá z minulých dat prvopojistitele. Jestliže např. $X_{OP,1}, X_{OP,2}, \ldots, X_{OP,n}$ byly všechny pozorované (aktualizované) škody překračující v minulém roce hodnotu OP, pak maximálně věrohodný odhad parametru b lze najít jako

$$b = \frac{n}{\sum_{i=1}^{n} \ln \frac{X_{OP,i}}{OP}}$$

Příklad. Úkolem je najít pomocí Paretova modelu nettozajistné pro příští rok zajistné smlouvy WXL/R s parametry 5 mil. Kč xs 1 mil. Kč v pojištění staveb občanů, jestliže na základě minulých (aktualizovaných) dat byl průměrný počet škod, které v zajišťovaném portfoliu během příštího roku překročí částku 250 000 Kč, odhadnut ve výši 9,36 škod.

Řešení: Pro výpočet použijeme Paretovo rozdělení s ad hoc hodnotou parametru $b = 2$:

$$RL = \frac{1\ 000\ 000 + 5\ 000\ 000}{1\ 000\ 000} = 6$$

$$\begin{aligned} NP_Z &= LF(a) \cdot EXL = LF(OP) \cdot OP^b \cdot \frac{a^{1-b}}{1-b} \cdot (RL^{1-b} - 1) = \\ &= 9,36 \cdot 250\ 000^2 \cdot \frac{1\ 000\ 000^{1-2}}{1-2} \cdot (6^{1-2} - 1) = 0,59 \cdot 833\ 333 = \\ &= 487\ 500 \text{ Kč} \end{aligned}$$

4.2 Metoda scénářů

- používá se pro stanovení nettozajistného při CatXL zajištění přírodních katastrof;
- je založena na odhadu jejich *škodní periody* (*škodní frekvence*): během této periody by hledané nettozajistné mělo právě zaplatit zajistné plnění;
- vytváří se přitom určité škodní scénáře, pomocí nichž se odhadují příslušné škodní periody;
- nevýhodou metody je značná nepřesnost, a proto se spíše používá jako podpůrný prostředek při obchodních jednáních mezi prvopojistitelem a zajistitelem.

Příklad. Metoda scénářů pro výpočet nettozajistného v rámci zajištění CatXL (5 mil. Kč xs 1 mil. Kč) v pojištění proti povodním a záplavám s celkovou pojistnou částkou 50 mil. Kč:

- pomocí odhadnutého škodního stupně (tj. podílu pojistného plnění vůči pojistné částce) v prvním sloupci tabulky 4 byla v druhém sloupci odhadnuta průměrná škoda při jednotlivých typech povodňové vody a v třetím sloupci odpovídající zajistné plnění;
- přepočtem na jeden rok bylo v posledním sloupci odhadnuto (roční) nettozajistné. Nettozajistná sazba vyjádřená ale tentokrát v procentech celkové pojistné částky je 0,16%.

Typ povodňové vody	*Škodní stupeň (%)*	*Průměrná škoda (Kč)*	*Zajistné plnění (Kč)*	*Škodní perioda (roky)*	*Netto-zajistné (Kč)*
desetiletá	1	500 000	0	10	0
padesátiletá	5	2 500 000	1 500 000	50	30 000
stoletá	20	10 000 000	5 000 000	100	50 000
Celkem					80 000

Tabulka 4: Příklad metody scénářů.

5 Alternativní přenos rizik (ART)

Jako příklad ART uvedeme tzv. *pojistné dluhopisy*, které zatím patří mezi nejpoužívanější nástroje sekuritizace pojistných rizik (obecně se pro tyto nástroje začíná používat zkratka *ILS insurance-linked securities*). K rozvoji sekuritizace pojistných rizik přispívá mimo jiné:

- Objektivní *ILS spouštěče (ILS triggers)*: Jedná se o objektivně vymezené události, které podmiňují výši finančních toků realizovaných v rámci ILS. Lze je rozlišit do tří kategorií podle jejich indexace:

  (1) Indexace na škodách způsobených přímo dané pojišťovně či zajišťovně.
  (2) Indexace pomocí všeobecně uznávaných škodních indexů měřených renomovanými agenturami. *Škodním indexem* se zde většinou rozumí vhodně standardizovaný průměrný škodní vývoj v daném regionu pro příslušně vymezené pojistné riziko. V současné době jsou rozšířeny *PCS indexy* (Property Claim Services).
  (3) Indexace pomocí parametrických indexů: *Parametrický index* má většinou fyzikální charakter. Typickým příkladem je Richterova stupnice pro zemětřesení nebo počet teplých a chladných dní u derivátů na počasí.

- Burzy obchodující se sekuritizovaným pojistným rizikem: Jedná se přitom o vznik nových specializovaných burz (např. americká elektronická burza *CATEX Catastrophe Risk Exchange*).
- Rozvoj zprostředkovatelů SPV (*Special Purpose Vehicles*): Tyto subjekty hrají mimo jiné důležitou roli právě při emisích cenných papírů vázaných na pojistné riziko.

Nejčastějším typem pojistných dluhopisů jsou *katastrofické dluhopisy (catastrophe bonds, CatBonds)*. Jedná se o vysoce ziskové dluhopisy s rizikem neplnění závazků v případě živelní katastrofy. Tyto dluhopisy mají kuponovou sazbu mnohem vyšší než průměr trhu. V případě živelní katastrofy daného typu však u nich hrozí ztráta celého (resp. části) kuponu a ztráta celé (resp. části) nominální hodnoty, např.:

Výše škody podle indexu PCS	*Nesplacená procentní část nominální hodnoty*			*Odhadnutá pravděpodobnost*
	*Tranše A*1, *A*2	*Tranše B*	*Tranše C*	
$\geq$ 12,0 mld. USD	0%	0%	100%	0,0240
$\geq$ 18,5 mld. USD	20%	33%	100%	0,0100
$\geq$ 21,0 mld. USD	40%	66%	100%	0,0076
$\geq$ 24,0 mld. USD	60%	100%	100%	0,0052
Očekávaná ztráta (v % nom. hodnoty)	0,46%	0,76%	2,40%	

Tabulka 5: Nesplacená procentní část nominální hodnoty v případě Swiss Re California Earthquake Bonds v závislosti na výši pojištěné škody podle příslušného indexu PCS.

(1) *USAA Hurricane Bonds:* jednoleté dluhopisy v celkové nominální hodnotě 0,5 mld. USD byly vázány na riziko hurikánů v pojištění majetku na východním pobřeží a kolem Golfského zálivu;
(2) *Winterthur Windstorm Bonds:* tříleté dluhopisy v nominální hodnotě 4 700 CHF byly vázány na riziko vichřic a krupobití v havarijním pojištění osobních automobilů;
(3) *Swiss Re California Earthquake Bonds:* tříleté dluhopisy v celkové nominální hodnotě 137 mil. USD byly vázány na riziko zemětřesení v Kalifornii;
(4) *Japonské katastrofické dluhopisy:* Japonské pojišťovny vzhledem k absenci škodních indexů typu PCS pro zemětřesení zvolily parametrický přístup ve formě Richterovy stupnice na základě hlášení japonské meteorologické služby.

Princip katastrofických dluhopisů popišme na nejjednodušším příkladu jednoletých dluhopisů s jedním ročním kuponem. Uvažujme jednoletou zajistnou smlouvu, podle níž zajistitel zaplatí po uplynutí jednoho roku částku L, pokud nastala předem specifikovaná živelní katastrofa (např. povodeň určitého rozsahu), a v opačném případě zajistitel neplní (částka L je samozřejmě sjednána před podpisem zajistné smlouvy). Cena takové zajistné smlouvy pro prvojistitele (tj. výše zajistného) se zřejmě určí jako

$$P_Z = \frac{1}{1+i} \cdot [q_{cat} \cdot L + (1 - q_{cat}) \cdot 0] = \frac{1}{1+i} \cdot q_{cat} \cdot L$$

kde i je roční úroková míra a q_{cat} je pravděpodobnost živelní katastrofy, tak jak ji oceňuje zajistný trh. Zajistitel musí před podpisem zajistné smlouvy (řekněme v čase $t = 0$) věrohodně doložit, že za rok (tj. v čase $t = 1$) bude disponovat kapitálovou částkou L, neboť by jinak nesehnal zákazníky. Potřebuje proto získat v čase $t = 0$ vedle zajistného P_Z ještě částku F takovou,

	Čas $t = 0$	*Čas* $t = 1$	
		došlo ke katastrofě (s prstí q_{cat}*)*	*nedošlo ke katastrofě (s prstí* $1 - q_{cat}$*)*
Prvopojistitel	$-Z = -\frac{1}{1+i} \cdot q_{cat} \cdot L$	L	0
Zajistitel = Emitent	$Z + F$	$-L$	$-L$
Investor	$-F = -\frac{1}{1+i} \cdot (1 - q_{cat}) \cdot L$	0	L

Tabulka 6: Finanční toky spojené s jednoletými katastrofickými dluhopisy (dluhopisy se prodávají za nominální hodnotu).

aby investičním zhodnocením kapitálu $P_Z + F$ opravdu obdržel za jeden rok potřebnou částku L

$$(P_Z + F) \cdot (1 + i) = L$$

Emituje proto vysoce ziskové jednoleté dluhopisy v nominální hodnotě F, v jejichž rámci zaplatí držiteli dluhopisů (investorovi) po uplynutí jednoho roku částku L (tj. na konci ročního období splatí nominální hodnotu F a vyplatí opravdu značný objem kuponů ve výši $K = L - F = P_Z \cdot (1+i) + F \cdot i$), pokud nenastala předem specifikovaná živelní katastrofa, a v opačném případě zajistitel nezaplatí držiteli dluhopisů nic a celou částku L použije na zajistné plnění. Nominální hodnota F splňuje vztah

$$F = \frac{1}{1+i} \cdot [q_{cat} \cdot 0 + (1 - q_{cat}) \cdot L] = \frac{1}{1+i} \cdot (1 - q_{cat}) \cdot L$$

Pokud se uvedené dluhopisy prodávají za nominální hodnotu F (tj. za pari), potom příslušné finanční toky jsou v souladu s potřebami všech zúčastněných stran (viz tab. 6):

Uvedené dluhopisy se ovšem prodávají za tržní cenu F^* (obecně $F^* \neq F$), takže dluhopisový trh ocenil pravděpodobnost katastrofy jako q_b (na rozdíl od odhadu q_{cat} provedeného zajistným trhem, přičemž index b je z anglického *bond* pro dluhopis) splňující

$$F^* = \frac{1}{1+i} \cdot (1 - q_{cat}) \cdot L$$

tj.

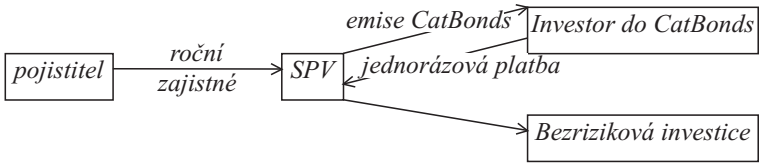
$$q_b = \frac{L - F^* \cdot (1+i)}{L}$$

To pak následně implikuje zajistné ve výši

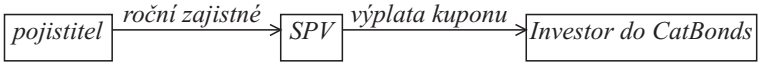
$$Z_b = \frac{1}{1+i} \cdot q_b \cdot L = \frac{L}{1+i} - F^*$$

Názorně lze zajištění pomocí katastrofických dluhopisů zachytit následujícím způsobem:

- *začátek zajistné smlouvy:*

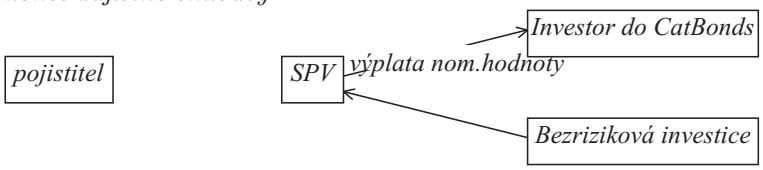


- *bezškodní rok zajistné smlouvy:*



- *škodní rok zajistné smlouvy:*

- *konec zajistné smlouvy:*



Reference

[1] Booth P., Chadburn R., Cooper D., Haberman S., James D. (1999). *Modern actuarial theory and practice.* Chapman and Hall/CRC, London.

[2] Cipra T. (2004). *Zajištění a přenos rizik v pojišťovnictví.* Grada, Praha.

[3] Gerathewohl K. et al. (1976). *Rückversicherung – Grundlagen und Praxis, Band I.* Verlag Versicherungswirtschaft, Karlsruhe.

[4] Gerathewohl K. et al. (1979). *Rückversicherung – Grundlagen und Praxis, Band II.* Verlag Versicherungswirtschaft, Karlsruhe.

[5] Kiln R., Kiln S. (2001). *Reinsurance in practice.* Witherby, London.

[6] Liebwein P. (2000). *Klassische und moderne Formen der Rückversicherung.* Verlag Versicherungswirtschaft, Karlsruhe.

[7] Pfeiffer C. (1994). *Einführung in die Rückversicherung: Das Standardwerk für Theorie und Praxis.* Gabler, Wiesbaden.

[8] Straub E. (1988). *Non-life insurance mathematics.* Springer and Association of Swiss Actuaries, Heidelberg.

[9] Tiller J.E., Fagerberg D. (1990). *Life, health, and annuity reinsurance.* ACTEX Publications, Winsted and Avon, Connecticut.

Poděkování: Tato práce je podporována výzkumným záměrem MSM 113200008.

Adresa: T. Cipra, KPMS, MFF UK, Sokolovská 83, 186 00 Praha 8 - Karlín

E-mail: cipra@karlin.mff.cuni.cz

ASYMPTOTICKÁ ANALÝZA STRATEGIÍ OBCHODOVÁNÍ S AKCIÍ PŘI EXISTENCI TRANSAKČNÍCH NÁKLADŮ

Petr Dostál

Klíčová slova: Obchodní strategie, transakční náklady, asymptotický užitek.

Abstrakt: Uvažujeme investora, který obchoduje s jednou akcií, ale na rozdíl od [1], [2] nic nespotřebovává. Jeho snaha je maximalizovat asymptotické chování očekávaného užitku měřeného užitkovou funkcí s hyperbolickou absolutní averzí vůči riziku (HARA) ve tvaru $U_\gamma(x) = x^\gamma/\gamma$ pro $\gamma < 0$ a $U_0(x) = \ln x$. Předpokládáme, že tržní cena akcie je geometrický Brownův pohyb. Tato omezení nám umožňují odvodit optimální intervalové strategie v téměř explicitní podobě. Tyto strategie jsou optimální i mezi všemi rozumnými strategiemi. V případě logaritmické užitkové funkce jsou odvozené strategie optimální i v modelu, který dostaneme rozumnou časovou transformací původního modelu geometrického Brownova pohybu. V ostatních případech jsou odvozené strategie optimální pouze při deterministické změně času.

1 Úvod

Předpokládejme, že tržní cena akcie X_t je geometrický Brownův pohyb

$$dX_t = \mu X_t\,dt + \sigma X_t\,dW_t, \qquad X_0 = x_0 > 0. \tag{1}$$

Nejprve budeme předpokládat, že depozitní část portfolia není úročena. Označme Y_t tržní cenu portfolia a G_t pozici investora na trhu v čase $t \geq 0$. Dále budeme označovat H_t počet akcií v portfoliu. Nyní můžeme vyjádřit tržní cenu akciové části portfolia v následujících dvou tvarech $G_tY_t = H_tX_t$. Dále budeme předpokládat, že platíme $(1+b)$-násobek tržní ceny akcie, abychom tuto akcii obdrželi. Na druhou stranu obdržíme $(1-c)$-násobek tržní ceny akcie, kterou prodáme. Rozdíly v cenách interpretujeme jako transakční náklady. Snadno zjistíme, že následující hodnota

$$Y_t(1+bG_t) = Y_t + bH_tX_t \quad \text{resp.} \quad Y_t(1-cG_t) = Y_t - cH_tX_t \tag{2}$$

zůstává stejná před a po provedení nákupu resp. prodeje. Tyto vztahy můžeme zapsat v difereneciální podobě

$$d\ln Y_t = -\vartheta_+(G_t)\,d^+G_t - \vartheta_-(G_t)\,d^-G_t, \tag{3}$$

kde $\vartheta_+(x) = \frac{b}{1+bx}$ a $\vartheta_-(x) = \frac{c}{1-cx}$ a kde d^+G_t a d^-G_t jsou diferenciály dvou neklesajících adaptovaných procesů reprezentující nárůst resp. pokles pozice způsobený nákupem či prodejem akcie. Pokud s akcií neobchodujeme, tak se

pozice investora G_t chová jako difúzní proces s driftem $B(x)$ a difúzí $S^2(x)$, kde

$$B(x) = x(1-x)[\mu - \sigma^2 x], \qquad S(x) = \sigma x(1-x). \tag{4}$$

Pokud obchodujeme, je G_t semimartingal se stochastickým diferenciálem

$$dG_t = B(G_t)\,dt + S(G_t)\,dW_t + d^+G_t - d^-G_t. \tag{5}$$

Výkyvy v tržní ceně portfolia Y_t jsou jednak způsobeny změnami tržní hodnoty akcie X_t a jednak tržní hodnota portfolia klesá o zaplacené transakční náklady, tj.

$$dY_t = H_t\,dX_t - Y_t\vartheta_+(G_t)\,d^+G_t - Y_t\vartheta_-(G_t)\,d^-G_t \tag{6}$$

$$= Y_t[G_t(\mu\,dt + \sigma\,dW_t) - \vartheta_+(G_t)\,d^+G_t - \vartheta_-(G_t)\,d^-G_t]. \tag{7}$$

Tato rovnost má řešení ve tvaru $Y_t = Y_0 \exp\{L_t\}$, kde

$$L_t = \int_0^t G_s\mu - \frac{1}{2}\,\sigma^2 G_s^2\,ds + \sigma\int_0^t G_s\,dW_s - \int_0^t \vartheta_+(G_s)\,d^+G_s - \int_0^t \vartheta_-(G_s)\,d^-G_s.$$

My se dále více zaměříme na strategie, které neobchodují, pokud se pozice G_t nachází v intervalu (α,β) a které akcii nakupují nebo prodávají tak, aby tato pozice neopustila interval $[\alpha,\beta]$. V takovýchto případech je diferenciál d^+G_t resp. d^-G_t soustředěn na množině $[G_t = \alpha]$ resp. $[G_t = \beta]$. Můžeme tedy psát $\vartheta_+(G_t)\,d^+G_t = \vartheta_\alpha\,d^+G_t$, resp. $\vartheta_-(G_t)\,d^-G_t = \vartheta_\beta\,d^-G_t$, kde $\vartheta_\alpha = \vartheta_+(\alpha) = \frac{b}{1+b\alpha}$ a $\vartheta_\beta = \vartheta_-(\beta) = \frac{c}{1-c\beta}$. Jako kritérium optimality budeme uvažovat maximalizaci asymptotického vývoje očekávaného užitku při volbě užitkových funkcí $U_0(x) = \ln x$ a $U_\gamma(x) = \frac{x^\gamma}{\gamma}$, kde $\gamma < 0$, tj.

$$\max \lim_{t\to\infty} \frac{1}{t} E \ln Y_t \qquad \text{resp.} \qquad \min \lim_{t\to\infty} \frac{1}{t} \ln E Y_t^\gamma. \tag{8}$$

2 Logaritmická užitková funkce

Na chvíli budeme uvažovat nulové transakční náklady, tj. $b = c = 0$. V tomto případě bychom měli maximalizovat $\lim_{t\to\infty} \frac{1}{t}\,E\int_0^t G_s\mu - \frac{1}{2}\sigma^2 G_s^2\,ds$. Funkce $x \mapsto x\mu - \frac{1}{2}\sigma^2 x^2$ nabývá maxima v bodě $\theta := \mu/\sigma^2$. Neexistuje tedy lepší strategie než $[\theta,\theta]$. Takováto strategie je neobchodující v případě, že $\theta \in \{0,1\}$. V těchto dvou případech je strategie $[\theta,\theta]$ optimální i v případě nenulových transakčních nákladů. Těmito případy $\theta = 0, 1$ se dále už tedy zabývat nebudeme. Nyní existují dvě cesty, kterými se dá pokračovat. Mohli bychom použít ergodickou teorii. Místo toho využijeme teorii martingalů. Tato cesta je založena na tom, že jsme schopni nalézt hladkou funkci f a konstantu ν takovou, že následující proces je martingal

$$\ln Y_t - f(G_t) - \nu t. \tag{9}$$

Z martingalové konvergence pak dostaneme, že

$$\lim_{t\to\infty}\frac{1}{t}E\ln Y_t=\nu+\lim_{t\to\infty}\frac{1}{t}Ef(G_t)=\nu \tag{10}$$

je tou hodnotou, kterou bychom měli maximalizovat. Hladká funkce f splňuje martingalovou podmínku (9), pokud splňuje následující ODE

$$f'(x)B(x)+\frac{1}{2}f''(x)S^2(x)=\mu x-\frac{1}{2}\sigma^2x^2-\nu \tag{11}$$

s okrajovými podmínkami

$$f'(\alpha)=-\vartheta_\alpha=-\frac{b}{1+b\alpha}\qquad\text{a}\qquad f'(\beta)=\vartheta_\beta=\frac{c}{1-c\beta}. \tag{12}$$

Označíme-li $h:=f'$, dostaneme obyčejnou diferenciální rovnici prvního řádu

$$h(x)B(x)+\frac{1}{2}h'(x)S^2(x)=\mu x-\frac{1}{2}\sigma^2x^2-\nu \tag{13}$$

s okrajovými podmínkami $h(\alpha)=-\frac{b}{1+b\alpha}$ a $h(\beta)=\frac{c}{1-c\beta}$. Protože (13) je ODE prvního řádu, jsme schopni vyjádřit obecné řešení této rovnice pomocí metody variace konstant a zodpovědět otázku, kdy tato rovnice má řešení vyhovyjícím okrajovým podmínkám (12). Možná volba funkce f (splňující (11) a (12)) je jakákoli primitivní funkce k h. Jedna taková možná volba f je

$$f(x)=2\rho\,a_0\left|\frac{1}{x}-1\right|^{2\rho}+a_1\ln\left|\frac{x}{1-x}\right|+\ln\left|\frac{1}{1-x}\right| \tag{14}$$

v případě, že $\rho:=\theta-\frac{1}{2}\neq 0, \nu=-\rho\sigma^2a_1$, kde $\xi_\alpha=\alpha\frac{1+b}{1+b\alpha}, \xi_\beta=\beta\frac{1-c}{1-c\beta}$ a kde

$$a_0=\frac{\xi_\beta-\xi_\alpha}{\left|\frac{1}{\beta}-1\right|^{2\rho}-\left|\frac{1}{\alpha}-1\right|^{2\rho}},\qquad a_1=-\frac{\left|\frac{\beta}{1-\beta}\right|^{2\rho}\xi_\beta-\left|\frac{\alpha}{1-\alpha}\right|^{2\rho}\xi_\alpha}{\left|\frac{\beta}{1-\beta}\right|^{2\rho}-\left|\frac{\alpha}{1-\alpha}\right|^{2\rho}}. \tag{15}$$

V případě, že $\rho=0$ a $\nu=-\frac{\sigma^2}{2}a_1$, můžeme volit

$$f(x)=a_0\ln\frac{x}{1-x}+\frac{a_1}{2}\ln^2\frac{x}{1-x}+\ln\frac{1}{1-x},\qquad\text{kde} \tag{16}$$

$$a_0=\frac{\ln\frac{\alpha}{1-\alpha}\xi_\beta-\ln\frac{\beta}{1-\beta}\xi_\alpha}{\ln\frac{\beta}{1-\beta}-\ln\frac{\alpha}{1-\alpha}},\quad a_1=-\frac{\xi_\beta-\xi_\alpha}{\ln\frac{\beta}{1-\beta}-\ln\frac{\alpha}{1-\alpha}}. \tag{17}$$

Protože bychom měli maximalizovat hodnotu $\nu=\lim_{t\to\infty}\frac{1}{t}E\ln Y_t$, je naším úkolem najít maximum funkce

$$u(\alpha,\beta)=\frac{\left|\frac{\beta}{1-\beta}\right|^{2\rho}\xi_\beta-\left|\frac{\alpha}{1-\alpha}\right|^{2\rho}\xi_\alpha}{\frac{1}{2\rho}\left[\left|\frac{\beta}{1-\beta}\right|^{2\rho}-\left|\frac{\alpha}{1-\alpha}\right|^{2\rho}\right]}\quad\text{resp.}\quad u(\alpha,\beta)=\frac{\xi_\beta-\xi_\alpha}{\ln\frac{\beta}{1-\beta}-\ln\frac{\alpha}{1-\alpha}}$$

podle toho, zda $\rho \neq 0$ či $\rho = 0$, a to na jedné z množin $T = \{(\alpha, \beta), 0 < \alpha < \beta < 1\}$ resp. $T = \{(\alpha, \beta), 1 < \alpha < \beta < 1/c\}$ resp. $T = \{(\alpha, \beta), -1/d < \alpha < \beta < 0\}$ podle toho, zda $\theta \in (0, 1)$ resp. $\theta \in (1, \infty)$ resp. $\theta \in (-\infty, 0)$.

Věta 1. *Funkce $u(\alpha, \beta)$ má právě jeden stacionární bod na množině T, který lze charakterizovat následujícími rovnostmi*

$$\xi_\alpha = \theta - \omega, \qquad \xi_\beta = \theta + \omega, \tag{18}$$

kde ω je (pokud $\theta \neq \frac{1}{2}$) jediné řešení rovnice

$$\ln \frac{1+b}{1-c} + \frac{1}{\rho}\left[\theta \ln \left|\frac{\theta+\omega}{\theta-\omega}\right| + (\theta-1)\ln\left|\frac{1-\theta+\omega}{1-\theta-\omega}\right|\right] = 0 \tag{19}$$

na $[0, |\theta| \wedge |1-\theta|)$. Pokud $\theta = \frac{1}{2}$, je ω jediné řešení následující rovnice

$$\ln \frac{1+b}{1-c} + 2\ln \frac{\frac{1}{2}+\omega}{\frac{1}{2}-\omega} = \frac{2\omega}{\frac{1}{4}-\omega^2} \tag{20}$$

na intervalu $[0, \frac{1}{2})$. Funkce u nabývá svého maxima na T v tomto stacionárním bodě. Navíc, rozdíl mezi levou a pravou stranou (19) resp. (20) je na odpovídajícím intervalu ryze monotónní funkce v ω.

Poznamenejme, že hodnoty α, β lze následně obdržet ze vzorců $\alpha = \xi_\alpha/(1+b-b\xi_\alpha), \beta = \xi_\beta/(1-c+c\xi_\beta)$. Nyní předpokládejme, že funkce u nabývá svého maxima na T v bodě (α, β). Dále definujme $\mathbf{F}(x) := f(x)$ pro $x \in [\alpha, \beta]$, $\mathbf{F}(x) := C_\alpha - \ln(1+bx)$ pro $x \in (-1/b, \alpha), \mathbf{F}(x) := C_\beta - \ln(1-cx)$ pro $x \in (\beta, 1/c)$, kde C_α, C_β jsou konstanty zvolené tak, aby funkce $\mathbf{F}$ byla spojitá na intervalu $(-1/b, 1/c)$.

Věta 2. *Nechť Y_t označuje tržní cenu portfolia a G_t pozici investora na trhu, pak*

$$\ln Y_t - \mathbf{F}(G_t) - \nu t \tag{21}$$

je součet supermartingalu a neroustoucího procesu za přepokladu, že zvolená strategie udržuje pozici G_t odraženou od extrémních hodnot $-1/b$ a $1/c$, a předpokladu, že zvolená strategie nedovolí, aby tržní cena portfolia klesla na nulu v konečném čase. Navíc, pokud je (21) martingal, pak lze říci, že byla aplikována strategie $[\alpha, \beta]$. Je to zřejmě martingal, pokud je použita strategie $[\alpha, \beta]$.

Z předchozí věty plyne, že neexistuje rozumná strategie s lepší asymptotikou střední hodnoty logaritmu tržní hodnoty portfolia než má strategie $[\alpha, \beta]$. Tato věta nám umožňuje definovat užitek v čase $t \geq 0$ na základě tržní hodnoty portfolia Y_t a pozice G_t pomocí (21). Takto definovaný systém užitků je v čase konzistentní a jako optimální strategii geneuje právě strategii $[\alpha, \beta]$.

3 Mocninná užitková funkce

Pokud by transakční náklady byly nulové, tj. $b = c = 0$, pak by γ-nejlepší strategie udržovala pozici G_t na hodnotě $\frac{\mu}{\sigma^2(1-\gamma)} = \frac{\theta}{1-\gamma}$, což lze nahléhnout z následujícího vyjádření $Y_t^\gamma = Y_0^\gamma \mathcal{E}_t \cdot \exp\{\gamma N_t\}$, kde

$$N_t = \int_0^t \mu G_s - \frac{1}{2}\sigma^2(1-\gamma)G_s^2\,ds - \int_0^t \vartheta_+(G_s)\,d^+G_s - \int_0^t \vartheta_-(G_s)\,d^-G_s, \quad (22)$$

$$\mathcal{E}_t = \exp\left\{\gamma\sigma\int_0^t G_s\,dW_s - \frac{1}{2}\gamma^2\sigma^2\int_0^t G_s^2\,ds\right\}. \quad (23)$$

V případě nulových transakčních nákladů $\vartheta_+(G_t) = \vartheta_-(G_t) = 0$ strategie $[\frac{\theta}{1-\gamma}, \frac{\theta}{1-\gamma}]$ totiž dává

$$EY_t^\gamma = EY_0^\gamma \exp\left\{\frac{\gamma}{2}\frac{\mu^2}{\sigma^2(1-\gamma)}t\right\} = EY_0^\gamma \exp\left\{\frac{\sigma^2}{2}\frac{\gamma\theta^2 t}{1-\gamma}\right\}, \quad (24)$$

což je menší nebo rovno než střední hodnota (24) v případě, že bychom uvažovali jakoukoli jinou intervalovou strategii, neboť funkce $x \mapsto \mu x - \frac{1}{2}\sigma^2(1-\gamma)x^2$ nabývá maxima $\frac{1}{2}\frac{\mu^2}{\sigma^2(1-\gamma)} = \frac{\sigma^2}{2}\frac{\theta^2}{1-\gamma}$ v bodě $\frac{\theta}{1-\gamma}$. Tato strategie je neobchodující, pokud $\frac{\theta}{1-\gamma} = 0$ nebo $\frac{\theta}{1-\gamma} = 1$, tj. pokud $\theta = 0$ nebo $\theta = 1-\gamma$. Tyto singulární případy budeme dále vynechávat a zaměříme se na strategie typu $[\alpha, \beta]$, kde $0 < \alpha < \beta < 1$, pokud $0 < \theta < 1-\gamma, 1 < \alpha < \beta < 1/c$, pokud $1-\gamma < \theta$ a na strategie typu $-1/d < \alpha < \beta < 0$, pokud $\theta < 0$. Nyní máme opět dvě možnosti, jak pokračovat. První cesta vede přes teorie semigrup lineárních operátorů na spojitých funkcích na $[\alpha, \beta]$ a spočítá ve výpočtu maximální vlastní hodnoty příslušného infinitezimálního generátoru. Jak uvidíme tak i v druhé možnosti se tomuto infinitezimálnímu generátoru nevyhneme a jeho maximální vlastní hodnotu budeme počítat, i když to tak třeba nebude vypadat. Tou druhou možností je nalézt konstantu ν a hladkou funkci f takovou, že

$$Y_t^\gamma g(G_t)e^{-\lambda t} = \exp\{\gamma[\ln Y_t - f(G_t) - \nu t]\} \quad (25)$$

je martingal, kde $g(x) = \exp\{-\gamma f(x)\}$ a kde $\nu = \lambda/\gamma$. Podle Itôovy formule stačí najít ν a f tak, aby platilo

$$\frac{1}{2}g''(x)S^2(x) + g'(x)\tilde{B}(x) + \gamma g(x)\left[\mu x + \frac{1}{2}(\gamma-1)\sigma^2x^2\right] = \lambda g(x), \quad (26)$$

$$g'_+(\alpha) = \gamma\frac{b}{1+b\alpha}\,g(\alpha), \qquad g'_-(\beta) = -\gamma\frac{c}{1-c\beta}\,g(\beta), \quad (27)$$

kde $\tilde{B}(x) = x(1-x)[\mu - (1-\gamma)\sigma^2 x]$. Levá strana (26) uvažovaná jako funkce proměnné x je hodnotou výše uvedeného infinitezimálního generátoru v bodě g semigrupy jejíž maximální vlastní hodnotu hledáme. Podmínka maximality mezi vlastními hodnotami odpovídá požadavku $g(x) = \exp\{-\gamma f(x)\}$, který zajišťuje, že funkce g nemění znaménko na $[\alpha, \beta]$.

Tento problém je možné řešit více-méně explicitně, neboť máme k dispozici fundamentální systém v explicitním tvaru

$$g_{1,2}(x) = \left|\frac{1}{x} - 1\right|^{\rho\mp\Delta} |1-x|^{\gamma}, \text{ pokud } \lambda = \frac{\sigma^2}{2}(\Delta^2 - \rho^2) \neq -\frac{\sigma^2}{2}\rho^2, \text{ resp.}$$

$$g_2(x) = g_1(x)\ln\left|\frac{x}{1-x}\right|, \text{ kde } g_1(x) = \left|\frac{1}{x} - 1\right|^{\rho} |1-x|^{\gamma}, \text{ pokud } \lambda = -\frac{\sigma^2}{2}\rho^2.$$

V tomto případě tak jsme schopni určit asymptotiku $\lambda = \lim_{t\to\infty} \frac{1}{t}\ln EY_t^{\gamma}$ jako implicitní funkci. Pokud $\rho_\alpha = \rho_\beta$, kde $\rho_\alpha := \rho + \gamma\xi_\alpha$ a $\rho_\beta := \rho + \gamma\xi_\beta$, platí

$$\lambda = \frac{\sigma^2}{2}\left(\rho_\alpha^2 - \rho^2\right) = \frac{\sigma^2}{2}\left(\rho_\beta^2 - \rho^2\right). \tag{28}$$

V opačném případě $\lambda = \frac{\sigma^2}{2}(D - \rho^2)$, kde D je jediné řešení rovnice

$$\ln\frac{1/\alpha - 1}{1/\beta - 1} = \int_{\rho_\beta}^{\rho_\alpha} \frac{dx}{x^2 - D} \tag{29}$$

na $\mathbf{R}_\infty \backslash co\{\rho_\alpha^2, \rho_\beta^2\}$, kde $\mathbf{R}_\infty = \mathbf{R}\cup\{\infty\}$ označuje jednobodovou kompaktifikaci reálné přímky a $co\{\rho_\alpha^2, \rho_\beta^2\}$ označuje konvexní obal množiny $\{\rho_\alpha^2, \rho_\beta^2\}$.

Je-li $D = 0$, je pravá strana (29) tvaru $1/\rho_\beta - 1/\rho_\alpha$. Je-li $D > 0$, je pravá strana (29) rovna

$$\frac{1}{2\Delta}\ln\frac{\Delta + \rho_\beta}{\Delta - \rho_\beta}\frac{\Delta - \rho_\alpha}{\Delta + \rho_\alpha}, \qquad \text{kde} \qquad \Delta^2 = D. \tag{30}$$

Pokud $D < 0$, lze pravou stranu (29) zapsat ve tvaru

$$\frac{1}{a}\left[\operatorname{arctg}\left(\frac{\rho_\alpha}{a}\right) - \operatorname{arctg}\left(\frac{\rho_\beta}{a}\right)\right], \qquad \text{kde} \qquad a^2 = -D. \tag{31}$$

Ve všech případech lze psát $\Delta^2 = D$, kde $\Delta \in \mathbf{R}$ resp. $i\Delta \in \mathbf{R}$. Pokud $\rho_\alpha = \rho_\beta =: \Delta$, je $g := g_1$ hledané řešení (26) a (27) kladné na $[\alpha, \beta]$. Pokud $\rho_\alpha \neq \rho_\beta$ a (29) platí pro $D = 0$, pak máme kladné řešení (26) a (27) na $[\alpha, \beta]$ ve tvaru

$$g(x) = g_1(x)\left|\frac{1}{\rho_\alpha} + \ln\frac{1/\alpha - 1}{1/x - 1}\right| = g_1(x)\left|\frac{1}{\rho_\beta} + \ln\frac{1/\beta - 1}{1/x - 1}\right|. \tag{32}$$

Pokud $\rho_\alpha \neq \rho_\beta$ a (29) platí pro nějaké $D = \Delta^2 > 0$, pak jedno z kladných řešení (26) a (27) na $[\alpha, \beta]$ je tvaru $g(x) = g_1(x)|\psi + |\frac{1}{x} - 1|^{2\Delta}|$, kde

$$\psi = \left|\frac{1}{\alpha} - 1\right|^{2\Delta}\frac{\Delta + \rho_\alpha}{\Delta - \rho_\alpha} = \left|\frac{1}{\beta} - 1\right|^{2\Delta}\frac{\Delta + \rho_\beta}{\Delta - \rho_\beta}. \tag{33}$$

Pokud $\rho_\alpha \neq \rho_\beta$ a (29) platí pro nějaké $D = -a^2 < 0$, máme k dispozici kladné řešení (26) a (27) na $[\alpha, \beta]$ ve tvaru

$$g(x) = \left|\frac{1}{x} - 1\right|^{\rho} |1-x|^{\gamma} \cdot 2\sin\left(\varphi - a\ln\left|\frac{1}{x} - 1\right|\right), \quad \text{kde} \tag{34}$$

$$\varphi = a\ln\left|\frac{1}{\alpha} - 1\right| + \operatorname{arccotg}\left(\frac{\rho_\alpha}{a}\right) = a\ln\left|\frac{1}{\beta} - 1\right| + \operatorname{arccotg}\left(\frac{\rho_\beta}{a}\right). \tag{35}$$

Věta 3. *Funkce $\lambda(\alpha, \beta)$ je spojitá na T a na této množině má právě jeden stacionární bod, který lze charakterizovat následujícími rovnostmi*

$$\xi_\alpha = \frac{\theta - \omega}{1 - \gamma}, \qquad \xi_\beta = \frac{\theta + \omega}{1 - \gamma}, \tag{36}$$

kde ω je jediné řešení rovnice $L(\omega) = P(\omega)$ na $[0, |\theta| \wedge |1 - \theta - \gamma|)$, kde

$$L(\omega) := \ln\frac{\theta+\omega}{\theta-\omega}\frac{1-\theta-\gamma+\omega}{1-\theta-\gamma-\omega} + \ln\frac{1+d}{1-c}, \quad P(\omega) := \int_{\frac{1}{2}-\omega}^{\frac{1}{2}+\omega} \frac{dx}{x^2 - D(\omega)},$$

kde $D(\omega) := \rho^2 + \frac{\gamma}{1-\gamma}(\theta^2 - \omega^2)$. Navíc, funkce $\lambda = \lambda(\alpha, \beta)$ nabývá minima v tomto stacionárním bodě. Dále rozdíl $L(\omega) - P(\omega)$ je ryze monotónní funkce na takových intervalech, na kterých je tato funkce spojitá.

Pokud $D(\omega) > 0$, lze funkci $P(\omega)$ počítat podle vzorce

$$P(\omega) := \frac{1}{2\sqrt{D(\omega)}} \ln\frac{\frac{1}{2} - \omega + \sqrt{D(\omega)}}{\frac{1}{2} + \omega + \sqrt{D(\omega)}}\frac{\frac{1}{2} + \omega - \sqrt{D(\omega)}}{\frac{1}{2} - \omega - \sqrt{D(\omega)}}, \qquad \text{resp.}$$

$$P(\omega) := \frac{1}{\sqrt{-D(\omega)}}\left[\operatorname{arccotg}\left(\frac{\frac{1}{2} - \omega}{\sqrt{-D(\omega)}}\right) - \operatorname{arccotg}\left(\frac{\frac{1}{2} + \omega}{\sqrt{-D(\omega)}}\right)\right]$$

v případě, že $D(\omega) < 0$, resp. $P(\omega) = \frac{2\omega}{\frac{1}{4} - \omega^2}$, pokud $D(\omega) = 0$.

Nyní předpokládejme, že funkce λ nabývá svého minima na T v bodě (α, β). Definujme dále $\mathbf{F}(x) := f(x) = -\frac{1}{\gamma}\ln g(x)$ pro $x \in [\alpha, \beta], \mathbf{F}(x) := D_\alpha - \ln(1 + bx)$ pro $x \in (-1/b, \alpha), \mathbf{F}(x) := D_\beta - \ln(1 - cx)$ pro $x \in (\beta, 1/c)$, kde konstanty D_α, D_β jsou zvoleny tak, aby funkce $\mathbf{F}$ byla spojitá na intervalu $(-1/b, 1/c)$. Konečně položme

$$\mathbf{G}(x) := \exp\{-\gamma\mathbf{F}(x)\}. \tag{37}$$

Věta 4. *Nechť Y_t je tržní cena portfolia a G_t je pozice investora na trhu, pak*

$$Y_t^\gamma \mathbf{G}(G_t)e^{-\lambda t} = \exp\left\{\gamma\left[\ln Y_t - \mathbf{F}(G_t) - \nu t\right]\right\}, \tag{38}$$

kde $\nu = \lambda/\gamma$, je součet submartingalu a neklesajícího procesu za předpokladu, že strategie obchodování udržuje pozici G_t odraženou od krajních hodnot $-1/b$

a $1/c$ *a pokud zaručuje, že tržní cena portfolia* Y_t *neklesne na nulu v konečném čase.*

Pokud je proces (38) martingal, lze říci, že byla aplikována strategie $[\alpha, \beta]$. *Proces (38) je zřejmě martingal, pokud je použita strategie* $[\alpha, \beta]$.

Podobně jako v případě logaritmické užitkové funkce z této věty plyne, že neexistuje rozumná strategie s lepším asymptotickým chováním středního užitku než je strategie $[\alpha, \beta]$, měříme-li asymptotický užitek z tržní hodnoty portfolia pomocí funkce $U_\gamma(x)$. Opět můžeme definovat užitek v čase t na základě Y_t a G_t pomocí $\frac{1}{\gamma}(38)$ za předpokladu, že náš cíl je maximalizovat asymptotikcké chování $EU_\gamma(Y_t)$. Mohli bychom také říci, že užitek je roven hodnotě

$$\ln Y_t - \mathbf{F}(G_t) - \nu t, \tag{39}$$

ale ne ve smyslu maximalizace očekávaného užitku, ale ve smyslu minimalizace střední hodnoty exponenciely z γ-násobku takovéhoto druhu užitku.

4 Nenulová úroková míra a změna času

Nechť Z_t označuje tržní cenu akcie v čase t. Označme X_t diskontovanou tržní cenu akcie $X_t = e^{-rt}Z_t$, kde r je konstantní úroková míra. Předpokládejme dále, že $dZ_t = \kappa Z_t\,dt + Z_t\sigma\,dW_t$. Pak $dX_t = \mu X_t\,dt + \sigma X_t\,dW_t$, kde $\mu := \kappa - r$. Definujeme-li Y_t jako diskontovanou tržní cenu portfolia, můžeme použít předchozí výsledky k tomu, abychom odvodili optimální strategie pro tento případ, neboť kritéria optimality jsou invariantní vzhledem k diskontování.

Poznámka ke změně času

Rozšířená optimalita odvozených strategií je založena na větách, které říkají, že nějaké procesy jsou martingaly, pokud použijeme odvozené strategie, zatímco jsou obecně jen super/sub-martingaly $+/-$ nerostoucí proces, pokud se omezíme na strategie udržující pozici investora G_t odraženou od extrémních hodnot $-1/b$ a $1/c$.

Tento druh optimality je stabilní v případě logaritmické užitkové funkce vzhledem k jakékoli rozumné změně času, tj. vzhledem k takovým transformacím času, které neporuší (sub,super)-martingalovou vlastnost našich (sub,super)-martingalů. Zatímco v případě mocninných užitkových funkcí je tato optimalita stabilní vzhledem k deterministickým změnám času v modelu.

Reference

[1] Janeček K., Shreve S.E. (2004). *Asymptotic analysis for optimal investment and consumption with transaction costs.* Fin.&Stochas. **8**, 181-206.

[2] Shreve S., Soner H.M. (1994). *Optimal investment and consumption with transaction costs.* Ann. Applied Probab. **4**, 609 – 692.

Poděkování: Účast na této konferenci byla umožněna na základě podpory z grantu GA ČR 201/03/1027 a výzkumného záměru MSM 113200008.

Adresa: P. Dostál, KPMS, MFF UK, Sokolovská 83, Praha 8 - Karlín

E-mail: dostal@karlin.mff.cuni.cz

SPEED OF CONVERGENCE TO EQUILIBRIUM OF ZERO RANGE PROCESS ON A BINARY TREE

Lucie Fajfrová

Keywords: Interacting particle system, zero range process, spectral gap, Poincare inequality.

Abstract: Let us consider the zero range process with the symmetric random walk on a binary tree as the single particle law. The paper bring out an estimate of a rate of convergence of this process to equilibrium in the following sense. We find a lower bound of a spectral gap of finite volume processes. That is processes on a finite binary tree of height n with a density ρ of particles. We distinguish two classes of speed function. For the constant speed function we show that the spectral goes to zero when $n \to \infty$ no faster than $(n2^n(1+2^n\rho^2))^{-1}$. A better lower bound, uniform in the density, is obtained for a speed function 'near' the linear function: the spectral gap goes to zero no faster than $(n2^n)^{-1}$. These estimates follow up the results of [3] and [5], where the symmetric zero range process on $\mathbb{Z}^d$ is considered.

1 Introduction

The zero range process (ZR) is one of the interacting particle systems which describe the movement of infinitely many indistinguishable particles on some set of nodes (sites) X. In this paper we are interested primarily in the case when X is a binary tree. The particles can move only over edges of this tree, it means that a particle at node x can jump just to a neighbour of x, i.e. a node y such that $(x, y) \in E$, where E is the set of all edges. This movement of particles is considered to be random and in this paper we shall consider only the symmetric and translation invariant case. One can imagine it simply: when considering only one particle then the movement is common symmetric random walk on binary tree. So at each node (except the root) the particle chooses one of the tree possibilities with the same probability. Nevertheless in general there are interactions among particles. The zero range interactions are described by so called *speed function* $g : \mathbb{N} \to [0, \infty)$, $g(0) = 0$. If there are k particles at a node x the rate of jump of one particle from x to its arbitrary neighbour is equal to $g(k)/3$. Note that exactly this function g determine the type of the interactions. So the name "zero range" comes from the fact that g depends only on the number of particles at particular site and hence the interactions appears only among particles at the same site. One of the basic examples is the constant speed function $g(k) = \mathrm{I}_{[k>0]}$. In this case we can interpret the zero range process as a (close) system of infinitely many queues with servers placed at the nodes of a binary tree and these are connected over the edges of this tree. Note that the zero range dynamics doesn't allow any particle to exit or enter and so the whole system is mass-conserving.

Finally the zero range process is a continuous time Markov process with the state space $\mathbb{N}^{\mathrm{X}} = \{\eta : \mathrm{X} \to \mathbb{N}\}$, which is given by infinitesimal generator $\mathcal{L}$ acting on cylinder functions:

$$\mathcal{L}f(\eta) = \sum_{x\in\mathrm{X}} \sum_{y:(x,y)\in E} \frac{1}{3}\, g(\eta(x))\big(f(\eta^{xy}) - f(\eta)\big), \tag{1}$$

for $\eta \in \mathbb{N}^{\mathrm{X}}$, where η is the configuration of particles on X and η^{xy} is the configuration which arises from η after the jump of a particle from x to y. Note that the rate of leaving in the *root* was modified (2/3 instead of uniform 1) in order to have the transition rates uniform 1/3. Nevertheless recall that the speed function g cannot be considered arbitrary. We need to avoid too fast increasing rates of jumps in order to be able to use the techniques of Markov processes. For instance we want to obtain the Markov semigroup of operators corresponding to the generator $\mathcal{L}$. Liggett [6] or Andjel [1] obtained this correspondence under the *Lipschitz condition*:

$$\textit{there exists } a_1 < \infty : \sup_k |g(k+1) - g(k)| \le a_1. \tag{2}$$

Invariant measures for such a Markov process are product measures ν^φ on $\mathbb{N}^{\mathrm{X}}$ with the same marginals $m^\varphi(\cdot) = \nu^\varphi(\eta : \eta(x) = \cdot)$ for every $x \in \mathrm{X}$. The measure m^φ on $\mathbb{N}$ is given by

$$m^\varphi(k) = Z_\varphi \frac{\varphi^k}{g(k)g(k-1)\dots g(1)} \qquad \forall\ k > 0 \tag{3}$$

and $m^\varphi(0) = Z_\varphi$, where Z_φ is a normalizing constant. Specially for the speed function $g(k) = \mathrm{I}_{[k>0]}$ it makes sense to consider any $\varphi \in [0,1)$ and then the measure m^φ is the geometrical distribution with parameter φ for every φ positive and $m^0 = \delta_0$, the Dirac measure siting in zero. More about invariant measures of the zero range processes can be found in [1].

The very usual approach is to study the zero range process using its *finite volume approximations*. It means that we approximate the set X by a sequence of finite sets $\mathrm{X}_1 \subset \mathrm{X}_2 \subset \dots$, $\bigcup \mathrm{X}_n = \mathrm{X}$ and we restrict the state space to $\Omega_{n,K} := \{\eta \in \mathbb{N}^{\mathrm{X}_n} : \sum_{x\in\mathrm{X}_n} \eta(x) = K\}$, the finite configurations of just K particles. In our case we consider X_n as the binary tree of height n with the set of edges E_n. Note that the root is settled at level zero thus the number of nodes of X_n is $2^{n+1}-1$. So a Markov process with the state space $\Omega_{n,K}$ given by generator $\mathcal{L}_n f(\eta) = \sum_{x\in\mathrm{X}_n} \sum_{y:(x,y)\in E_n} 1/3\ g(\eta(x))\big(f(\eta^{xy}) - f(\eta)\big)$ we shall call a *finite volume zero range process*. Recall that for each fixed n and K it is a finite state space Markov process and its description using the generator $\mathcal{L}_n$ can be replaced simply by using the transition rate matrix $Q_{n,K} := \big(\mathrm{I}_{[\zeta=\eta^{xy} \text{ for some } (x,y)\in E_n]} g(\eta(x))/3\big)_{\eta,\zeta\in\Omega_{n,K}}$ or equivalently by using the appropriate transition matrix semigroup $\big(P_t^{n,K}\big)_{t\ge 0}$.

One can obtain the invariant measure for a finite volume ZR by conditioning product measures $\nu_n^\varphi := \big(m^\varphi\big)^{\otimes n}$ on event $\big[\sum_{x\in\mathrm{X}_n} \eta(x) = K\big]$. Formally

$$\nu_{n,K}(\cdot) = \nu_n^\varphi\left(\ \cdot\ \Big|\ \sum_{x\in\mathrm{X}_n} \eta(x) = K\right) \tag{4}$$

is the measure on $\Omega_{n,K}$. It is easy to observe that $\nu_{n,K}$ is moreover reversible. Specially for the speed function $g(k) = \mathrm{I}_{[k>0]}$ the measure $\nu_{n,K}$ is the uniform distribution on $\Omega_{n,K}$.

It is a basic result from the theory of Markov process that for every f on $\mathbb{N}^{\mathrm{X}_n}$ there holds $P_t^{n,K} f \overset{t\to\infty}{\longrightarrow} \mathrm{E}_{\nu_{n,K}} f$. When $\nu_{n,K}$ is the reversible measure then we can even estimate the speed of this convergence using $L_2(\nu_{n,K})$ norm:

$$||P_t^{n,K} f - \mathrm{E}_{\nu_{n,K}} f||_{L_2(\nu_{n,K})} \leq ||f||_{L_2(\nu_{n,K})} \ \exp\{-t|\lambda|\} \tag{5}$$

where λ is the second largest eigenvalue of the transition rate matrix $Q_{n,K}$. Note that $|\lambda|$ is usually called the *spectral gap*. These basic facts about Markov processes and spectral analysis of its generator matrix can be found in [2]. Let us remind here that the spectral gap can be characterized using the *Dirichlet form* of given Markov process

$$\mathcal{D}_{n,K} f := \langle f, -\mathcal{L}_n f\rangle_{\nu_{n,K}} = \frac{1}{2} \sum_{x,y\in \mathrm{X}_n} \frac{1}{3} \mathrm{E}_{\nu_{n,K}} \Big(g(\eta(x))[f(\eta^{xy}) - f(\eta)]^2 \Big), \tag{6}$$

and the variation of measure $\nu_{n,K}$:

$$|\lambda| = \inf_{f\neq 0} \frac{\mathcal{D}_{n,K} f}{Var_{\nu_{n,K}} f}.$$

Once we have for all n, K some constant $\gamma_{n,K}$ satisfying the inequality

$$Var_{\nu_{n,K}} f \leq \gamma_{n,K} \mathcal{D}_{n,K} f \tag{7}$$

for every f on $\mathbb{N}^{\mathrm{X}_n}$, then $1/\gamma_{n,K}$ is a lower bound of the spectral gap for (n, K)–finite volume process. Note that the inequality (7) is called a *Poincare inequality*.

Now we are ready to say what is our aim. It is to find $\gamma_{n,K}$ for every n, K in order to find out the limit behavior of the spectral gap when $n \to \infty$ and $K/|\mathrm{X}_n| \ \to \rho$.

Let us mention here two previous results. At first we mention the paper of Landim at al. [5] where appears the first sharp lower bound on the spectral gap for the zero range process on a cube $\{1, \ldots, n\}^d \subset \mathbb{Z}^d$. They assume that the speed function satisfies the Lipschitz condition (2) and in addition *the linear growth condition*:

$$\textit{there exists } i \in \mathbb{N} \textit{ and } a_2 > 0 \textit{ such that } g(k) - g(l) \geq a_2 \ \forall\, k \geq l + i. \tag{8}$$

For homogeneous symmetric zero range process on $\mathrm{X}_n := \{1, \ldots, n\}^d$ they obtain the following Poincare inequality:

There exists a constant W independent of n and K such that for every f on $\Omega_{n,K}$ the following holds:

$$Var_{\nu_{n,K}}(f) \leq W_0 n^2 \mathcal{D}^*_{n,K}(f), \tag{9}$$

where $\mathcal{D}^*_{n,K} = \frac{1}{2} \sum_{x,y\in \mathrm{X}_n : |x-y|=1} \frac{1}{2d} \mathrm{E}_{\nu_{n,K}} \Big(g(\eta(x))[f(\eta^{xy}) - f(\eta)]^2 \Big)$ is the corresponding Dirichlet form and $\nu_{n,K}$ is the invariant measure, the same as in (4). This inequality implies a spectral gap of at least $1/(Wn^2)$ on a cube of

volume n^d. See Theorem 1.1. in [5] for details. Note that the desired constant $\gamma_{n,K}$ in the Poincare inequality is here uniform in the number K of particles.

At second, in the work of Caputo [3] is a generalization of (9) using quite different approach. They drops geometrical constraint on the dynamics and consider a process on an arbitrary set of the same volume as X_n where jumps are allowed between any two sites rather than only between nearest neighbours. Moreover every these jumps have the same probability and it is equal to $(|\mathrm{X}_n| - 1)^{-1}$. The rest of dynamics given by the speed function g is preserved. Let us call (as in [3]) such a dynamics *the complete graph dynamics* and a corresponding process *the process on the complete graph* X_n. We put the result here for the readers convenience.

Theorem C 1.1. [Theorem 4.1. in [3]] *Let us consider the homogeneous zero range process on a complete graph* X_n *with a speed function* g *satisfying both (2) and (8) conditions.*

Then there exists a constant C *independent of* n *and* K *such that for every* f *on* $\Omega_{n,K}$

$$Var_{\nu_{n,K}} f \leq C\ \mathcal{E}_{n,K} f$$

where

$$\mathcal{E}_{n,K} f = \frac{1}{2} \sum_{x,y \in \mathrm{X}_n} \frac{1}{|\mathrm{X}_n| - 1}\ \mathrm{E}_{\nu_{n,K}} \Big(g(\eta(x))[f(\eta^{xy}) - f(\eta)]^2 \Big) \qquad (10)$$

is the corresponding Dirichlet form and $\nu_{n,K}$ *is the invariant measure; the same as in (4).*

Note that the Poincare inequality is now uniform in both parameters K and n as well. We apply ourselves to the result of this theorem because right an absence of geometrical bounds allows to employ this uniform Poincare inequality for arbitrary graph (X_n, E_n) if we would be able to turn from the complete graph dynamics to the nearest neighbour dynamics. This turning consists in mere comparison between the Dirichlets forms $\mathcal{E}_{n,K}$ and $\mathcal{D}_{n,K}$ and is usually called *the moving particle lemma*. We shall solve this task later in details in section 3, where we prove the moving particle lemma for the zero range process on a connected graph (X_n, E_n).

Nevertheless the conditions (2) and (8) admit only the speed functions which are not so different from the linear one $g(k) = k$. But at the beginning we mentioned interesting case of the speed function: the constant speed function $g(k) = \mathrm{I}_{[k>0]}$, for which the assumption (8) fails. Since the assumption (8) seems to be essential to have a Poincare inequality uniform in parameter K and the technique used in both theorem are based on this uniformity we are forced to use another approach to finding a Poincare inequality for ZR on binary tree with the constant speed function. This is aim of section 4.

2 Poincare inequality of ZR on a binary tree

We are going to state results concerning the Poincare inequality for the Dirichlet form of the zero range process on binary tree as it was described in

the previous section. The first proposition supposes the speed function from a class given by conditions (2) and (8). The second proposition concerns the constant speed function.

Proposition 2.1. *Let us consider the zero range process, where the single particle law is the symmetric random walk on the binary tree of height n and the speed function g satisfies the Lipschitz condition (2) and the linear growth condition (8).*

Then there exists a constant $C < \infty$ *such that for every* $n \in \mathbb{N}$, $K > 0$ *for every* f *on* $\Omega_{n,K}$

$$Var_{\nu_{n,K}} f \leq Cn2^n \mathcal{D}_{n,K} f$$

where $\nu_{n,K}$ *and* $\mathcal{D}_{n,K}$ *are the corresponding invariant measure and Dirichlet form, defined in (4) and (6), respectively.*

Proof. We divide the proof into two steps. The first one is to apply the general result for the zero range process on a complete graph, see Theorem C 1.1. We obtain from it a Poincare inequality $Var_{\nu_{n,K}} f \leq C\ \mathcal{E}_{n,K} f$ which is uniform in the number of particles K and in the volume n. Recall that $\mathcal{E}_{n,K} f$ is the Dirichlet form of the zero range process on the complete graph (10). Our aim is to obtain a Poincare inequality with the Dirichlet form $\mathcal{D}_{n,K}$ appropriate to the original graph structure (6). It is easy to observe that the measure $\nu_{n,K}$ is invariant for each homogeneous symmetric single particle law. Hence $\nu_{n,K}$ is invariant for the zero range process on the complete graph and for the zero range with the original graph structure as well.

The second step lies in comparison between these two Dirichlet forms $\mathcal{E}_{n,K}$ and $\mathcal{D}_{n,K}$, where for both we consider the same invariant measure $\nu_{n,K}$ and the same speed function g. This comparison is carried out for the zero range process on an arbitrary connected graph (X_n, E_n) in the next section 3. Specially for the considered binary tree a result (13) gives the desired comparison which completes this proof. □

Proposition 2.2. *Let us consider the zero range process, where the single particle law is the symmetric random walk on the binary tree of height n and the speed function is* $g(k) = \mathrm{I}_{[k>0]}$.

Then there exists a constant $C < \infty$ *such that for every* $n \in \mathbb{N}$, $K > 0$ *and for every* f *on* $\Omega_{n,K}$

$$Var_{\nu_{n,K}} f \leq C\Big(n2^n + nK^2\Big)\mathcal{D}_{n,K} f,$$

where $\nu_{n,K}$ *and* $\mathcal{D}_{n,K}$ *are the corresponding invariant measure and Dirichlet form, respectively, the special case of (4) and (6) for* $g(k) = \mathrm{I}_{[k>0]}$.

Proof. A proof of this Proposition requires a little more care then the proof of Proposition 2.1 because in this case of speed function the assumptions of Theorem C1.1 are not fulfilled. Moreover there is no constant uniform in number of particles such that a Poincare inequality holds with it. We shall

make use of a known similarity between the zero range and the exclusion dynamics. Let us briefly describe here the exclusion process: The exclusion process is a mass conserving particle system, as well as the zero range process. But there is allowed at most one particle per site (it means that each site is either occupied or vacant) and so interactions arise between particles at some neighbour sites. Anyway it is a Markov process with infinitesimal generator $\mathcal{L}^{EX}\tilde{f}(\tilde{\eta}) =$

$$\sum_{i\in I}\sum_{j\in I:|j-i|=1} \mathrm{I}_{[\ \tilde{\eta}(i)=1,\tilde{\eta}(j)=0]}\frac{1}{2}\big(\tilde{f}(\tilde{\eta}^{ij})-\tilde{f}(\tilde{\eta})\big) = \frac{1}{2}\sum_{i\in I^{-1}}\big(\tilde{f}(\tilde{\eta}^{i,i+1})-\tilde{f}(\tilde{\eta})\big).$$

for $\tilde{\eta}\in\{0,1\}^I$ and $\tilde{f}$ on $\{0,1\}^I$ cylinder function where we assume arbitrary interval $I\subseteq\mathbb{Z}$ as a set of sites with the symmetric random walk on I as a single particle law. I^{-1} stands for an interval I without the maximal element, if it exists, and $\tilde{\eta}^{ij}$ is the configuration which arises from $\tilde{\eta}$ after exchange values on sites i and j.

If I is finite we shall again (like for ZR) consider *the exclusion process on complete graph* I with generator $\frac{1}{|I|-1}\sum_{i\in I^{-1}}\big(\tilde{f}(\tilde{\eta}^{i,i+1})-\tilde{f}(\tilde{\eta})\big)$. The Caputo result concerning the spectral gap for the exclusion dynamics on the complete graph is the same as for ZR on the complete graph with the speed function satisfying (2) and (8), it means that the Poincare inequality is uniform in the volume $|I|$ and the number of particles K as well. See [3, Th.3.1].

The next step of the proof is again a comparison between Dirichlet forms for the complete graph and for the original graph. This time in addition the first mentioned Dirichlet form is considered for the exclusion dynamics. This comparison and also a completion of this prove we postpone to section 4. □

Now we are going to state the main theorem which describes the speed of L_2 convergence to the invariant measures for the zero range processes from the previous. We pay attention to dependence on parameters n, K.

Theorem 2.3. *For the model described in Proposition 2.1 there exists a constants* $0<C<\infty$ *such that for each* $n\in\mathbb{N}$, $K>0$, f *on* $\Omega_{n,K}$ *and* $t>0$ *the following holds:*

$$||P_t^{n,K}f-\mathrm{E}_{\nu_{n,K}}f||_{L_2(\nu_{n,K})}\le\|f\|_{L_2(\nu_{n,K})}\exp\Big\{\frac{-t}{Cn2^n}\Big\}.$$

For the model described in Proposition 2.2 there exists a constants $0<C<\infty$ *such that for each* $n\in\mathbb{N}$, $K>0$, f *on* $\Omega_{n,K}$ *and* $t>0$*:*

$$||P_t^{n,K}f-\mathrm{E}_{\nu_{n,K}}f||_{L_2(\nu_{n,K})}\le\|f\|_{L_2(\nu_{n,K})}\exp\Big\{\frac{-t}{C(n2^n+nK^2)}\Big\},$$

where $P_t^{n,K}$ *and* $\nu_{n,K}$ *are the transition matrix at time* t *and the invariant measure of the corresponding zero range process, respectively.*

Proof. This theorem is a right consequence of the proposition 2.1 and the proposition 2.2, respectively, when also the inequality (5) is applied. □

Note that we can generalize this theorem to arbitrary connected graph (X, E). For the proper statement see [4].

3 Moving particle lemma

We consider a connected graph (X, E) as a set of sites. As an finite approximation we mean a sequence $\{(\mathrm{X}_n, E_n)\}_n$ of finite connected subgraphs such that $\mathrm{X}_n \subset \mathrm{X}_{n+1}$ & $E_n = E \restriction_{\mathrm{X}_n}$ for each n. Of course we want that $\bigcup_n \mathrm{X}_n = \mathrm{X}$. The Dirichlet form of the appropriate zero range process is

$$\mathcal{D}_{n,K} f = \frac{1}{2} \sum_{x \in \mathrm{X}_n} \sum_{y \in \mathrm{X}_n:\ (x,y) \in E_n} \frac{1}{d_m} \mathrm{E}_{\nu_{n,K}} \Big(g(\eta(x)) \big(f(\eta^{xy}) - f(\eta) \big)^2 \Big) \quad (11)$$

where $d_m := \max\{d(x) : x \in \mathrm{X}_n\}$ is the maximal degree of the graph.

Our aim is to estimate the auxiliary Dirichlet form $\mathcal{E}_{n,K}$, see (10), by the Dirichlet form $\mathcal{D}_{n,K}$ of the original process (11) in order to gain a Poincare inequality for $\mathcal{D}_{n,K}$ and $\nu_{n,K}$ as a consequence of Caputo theorem C 1.1.

What is needed to do is to replace every transition $\eta \mapsto \eta^{xy}$ where x, y are arbitrary distinct nodes by a sequence of transitions $\zeta_i \mapsto \zeta_i^{x_i, x_{i+1}}$ just over edges of graph $(x_i, x_{i+1}) \in E_n$ where the sequence $(x_0 = x, x_1, \ldots, x_{r-1}, x_r = y)$ is a path in graph between the nodes x and y. Let us establish a set Γ of paths between each couple of distinct nodes x, y but just one path for each couple. If there exist more then one path we must choose just one of them and we allow just paths without repeated edges. We shall denote $\gamma_{xy} \in \Gamma$, $\gamma_{xy} = \{(x, x_1), (x_1, x_2), \ldots, (x_{r-1}, y)\}$.

Let us fix f on $\Omega_{n,K}$. Then for arbitrary $x \neq y \in \mathrm{X}_n$

$$\Big(f(\eta^{xy}) - f(\eta) \Big)^2 = \Big(\textstyle\sum_{i=0}^{r-1} f(\eta^{x_0, x_{i+1}}) - f(\eta^{x_0, x_i}) \Big)^2.$$

For expected values by using the Schwartz inequality we obtain:

$$\mathrm{E}_{\nu_{n,K}} \Big(g(\eta(x)) \big(f(\eta^{xy}) - f(\eta) \big)^2 \Big) \leq r \sum_{i=0}^{r-1} \mathrm{E}_{\nu_{n,K}} \Big(g(\eta(x)) \big(f(\eta^{x_0, x_{i+1}}) - f(\eta^{x_0, x_i}) \big)^2 \Big)$$

Using reversibility of $\nu_{n,K}$ it is equal to

$$r \sum_{i=0}^{r-1} \sum_{\eta : \eta(x) > 0} \nu_{n,K}(\eta^{x_0, x_i}) g(\eta^{x_0, x_i}(x_i)) \Big(f(\eta^{x_0, x_{i+1}}) - f(\eta^{x_0, x_i}) \Big)^2 =$$

$$= r \sum_{i=0}^{r-1} \sum_{\zeta : \zeta(x_i) > 0} \nu_{n,K}(\zeta) g(\zeta(x_i)) \Big(f(\zeta^{x_i, x_{i+1}}) - f(\zeta) \Big)^2.$$

And summing over all x, y we get: $\mathcal{E}_{n,K} f \leq$

$$\leq \frac{1}{2(|\mathrm{X}_n - 1|)} \sum_{x \in \mathrm{X}_n} \sum_{y \in \mathrm{X}_n, y \neq x} |\gamma_{xy}| \sum_{i=0}^{r-1} \mathrm{E}_{\nu_{n,K}} g(\zeta(x_i)) \Big(f(\zeta^{x_i, x_{i+1}}) - f(\zeta) \Big)^2 =$$

$$= \frac{1}{2(|\mathrm{X}_n| - 1)} \sum_{u \in \mathrm{X}_n} \sum_{v : (u,v) \in E_n} \mathrm{E}_{\nu_{n,K}} g(\zeta(u)) \Big(f(\zeta^{u,v}) - f(\zeta) \Big)^2 \frac{1}{2} \sum_{x \in \mathrm{X}_n} \sum_{y : \gamma_{xy} \ni (u,v)} |\gamma_{xy}|.$$

Now an estimate uniform in $e := (u, v) \in E_n$ of the term $\sum_{xy} \mathrm{I}_{[\gamma_{xy} \ni\ e]} |\gamma_{xy}|$ is required. Let us establish characteristics $\gamma := \max\{|\gamma_{xy}| : x, y \in \mathrm{X}_n\}$, the maximum length of path from Γ, and the measure of bottleneck

$b := \max_{e\in E_n} |\{\gamma_{xy} \ni e : \{x,y\} \subset \mathrm{X}_n\}|$. Then

$$\begin{aligned}\mathcal{E}_{n,K}f &\leq \frac{\gamma\ b}{2(|\mathrm{X}_n|-1)} \sum_{u\in \mathrm{X}_n} \sum_{v:(u,v)\in E_n} \mathrm{E}_{\nu_{n,K}} g(\zeta(u))\Big(f(\zeta^{uv}) - f(\zeta)\Big)^2 \\ &\leq \frac{d_m \gamma b}{|\mathrm{X}_n|-1} \mathcal{D}_{n,K}\ f \end{aligned} \tag{12}$$

for each $n \in \mathbb{N}$, $K > 0$, f on $\Omega_{n,K}$.

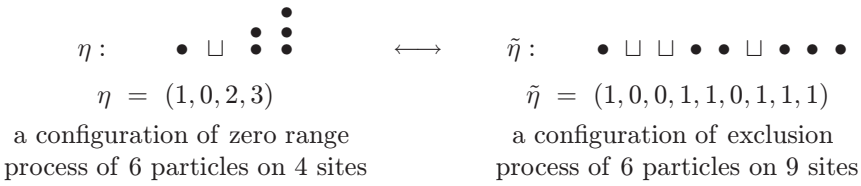
As a consequence we obtained the moving-particle lemma for the discussed binary tree ($d_m = 3,\ \gamma = 2n,\ b = 2^n(2^n - 1)$ and $|\mathrm{X}_n| = 2^{n+1} - 1$):

$$\mathcal{E}_{n,K}f \leq 3n2^n \mathcal{D}_{n,K} f \tag{13}$$

holds for each $n \in \mathbb{N}$, $K > 0$ and f on $\Omega_{n,K}$.

4 Comparison between Dirichlet forms

As it was mentioned behind the proposition 2.2 we shall take advantage of duality between an exclusion process on the line and a zero range process on the line. We only illustrate the easy principle of this duality on the following picture:



Such configurations η and $\tilde{\eta}$ we shall call dual. Notice that the total number of particles K is for dual configurations the same and that the number of sites for the zero range process N is equal to the number of empty sites (holes) in the exclusion process plus one. We shall denote $\kappa : \Omega_{N,K} \to \tilde{\Omega}_{K+N-1,K} := \{\tilde{\eta} \in \{0,1\}^{K+N-1} : \sum \tilde{\eta}(i) = K\}$ the bijection representing this duality. It is easy to see that there is also duality between transitions $\eta \mapsto \eta^{x,x+1}$ and $\tilde{\eta} \mapsto \tilde{\eta}^{i,i+1}$ such that $\kappa(\eta^{x,x+1}) = \tilde{\eta}^{i,i+1}$. Finally we define also a bijection between function spaces: if f is a function $f : \Omega_{N,K} \to \mathbb{R}$ then we put $\tilde{f} : \tilde{\Omega}_{K+N-1,K} \to \mathbb{R}$, $\tilde{\eta} \mapsto f(\kappa^{-1}\tilde{\eta})$.

Our aim is to employ the second mentioned Caputo theorem [3, Th.3.1] which gives a Poincare inequality for the exclusion process on the complete graph with K particles on $K + N - 1$ nodes in this way:

$$Var_{\mu_{N+K-1,K}} \tilde{f} \leq C\mathcal{E}^{EX}_{N+K-1,K}\tilde{f} \quad \forall N\ \forall K\ \forall \tilde{f} \text{ on } \tilde{\Omega}_{K+N-1,K},$$

where $\mu_{N+K-1,K}(\tilde{\eta}) \equiv \binom{K+N-1}{K}^{-1}$ is the invariant measure and $\mathcal{E}^{EX}_{N+K-1,K}\tilde{f}$ is the Dirichlet form, respectively, for the complete graph exclusion dynamics when we consider K particles on $K+N-1$ nodes:

$$\mathcal{E}^{EX}_{N+K-1,K}\tilde{f} = \sum_{i=1}^{N+K-2}\sum_{j=i+1}^{N+K-1}\frac{1}{N+K-2}\mathrm{E}_{\mu_{N+K-1,K}}\mathrm{I}_{\substack{[\tilde{\eta}(i)=1,\\ \tilde{\eta}(j)=0]}}\left(\tilde{f}(\tilde{\eta}^{ij})-\tilde{f}(\tilde{\eta})\right)^2. \quad (14)$$

We know that the invariant measure $\nu_{N,K}$ for the symmetric zero range process with $g(k) = \mathrm{I}_{[k>0]}$ is also uniform measure and sets $\Omega_{N,K}$ and $\tilde{\Omega}_{K+N-1,K}$ have the same volume. Hence $Var_{\nu_{n,K}} f = Var_{\mu_{V_n+K-1,K}}\tilde{f}$ for every f and $\tilde{f}$ dual functions. We obtain

$$Var_{\nu_{n,K}} f \le C\mathcal{E}^{EX}_{N+K-1,K}\tilde{f} \quad \forall N\ \forall K\ \forall f \text{ on } \Omega_{N,K}. \quad (15)$$

The remaining step of the proof of Proposition 2.2 is to estimate the Dirichlet form $\mathcal{E}^{EX}_{N+K-1,K}\tilde{f}$ by the Dirichlet form $\mathcal{D}_{n,K}f$ (11) of the zero range process with original graph structure also with K particles where $N = |\mathrm{X}_n|$.

We omit many details here and bring out only sketch of the remaining part of proof (a detailed proof can be found in [4]). Let us consider again as in section 3 a connected graph (X_n, E_n).

Start with the Dirichlet form $\mathcal{E}^{EX}_{N+K-1,K}\tilde{f}$ for some fixed $\tilde{f}$ on $\tilde{\Omega}_{K+N-1,K}$. At first, let us also fix the sites $i < j$ and $\tilde{\eta} \in \tilde{\Omega}_{K+N-1,K}$ such that $\tilde{\eta}(i) = 1$, $\tilde{\eta}(j) = 0$. Our aim is now to replace the exclusion transition $\tilde{\eta} \longmapsto \tilde{\eta}^{ij}$ by some sequence of zero range transitions $\eta \longmapsto \ldots \longmapsto \kappa^{-1}(\tilde{\eta}^{ij})$. Since we take advantage of duality only if we consider nodes for the zero range dynamics linearly ordered we must order the nodes of graph X_n.

To make the following lines a bit more clear let us set up an operator $T_{i,j}$ on $\tilde{\Omega}_{N,K}$, $T_{i,j} : \tilde{\eta} \longmapsto \tilde{\eta}^{ij}$. We can replace the difference $\tilde{f}(\tilde{\eta})-\tilde{f}(\tilde{\eta}^{ij})$ by a sum

$$\left(\tilde{f}(\tilde{\eta})-\tilde{f}(\tilde{\eta}_\tau)\right)+\sum_{l=i+1}^{j-2}\left(\tilde{f}(T_{l-1,l}(\ldots T_{j-2,j-1}\tilde{\eta}_\tau))-\tilde{f}(T_{l,l+1}(\ldots T_{j-2,j-1}\tilde{\eta}_\tau))\right)$$
$$+\left(\tilde{f}(T_{j-2,j-1}\tilde{\eta})-\tilde{f}(\tilde{\eta}_\tau)\right),$$

where $\tilde{\eta}_\tau = \tilde{\eta}_{\tau(ij)} := T_{j-1,j}(T_{j-2,j-1}(\ldots T_{i+1,i+2}(T_{i,i+1}\tilde{\eta})))$. Filling this term in (14) and applying the Schwartz inequality we obtain:

$$2\sum_{i=1}^{N+K-2}\sum_{j=i+1}^{N+K-1}\frac{1}{N+K-2}\mathrm{E}_{\mu_{N+K-1,K}}\left(\mathrm{I}_{[\tilde{\eta}(i)=1,\tilde{\eta}(j)=0]}\left(\tilde{f}(\tilde{\eta})-\tilde{f}(\tilde{\eta}_\tau)\right)^2\right) \quad (16)$$

$$+2K\sum_{i=1}^{N+K-2}\sum_{j=i+1}^{N+K-1}\frac{1}{N+K-2}\sum_{l=i+1}^{j-1}\sum_{\substack{\tilde{\zeta}:\\ \tilde{\zeta}(l-1)=1,\tilde{\zeta}(l)=0,\tilde{\zeta}(j)=1}}\left(\tilde{f}(T_{l-1,l}\tilde{\zeta})-\tilde{f}(\tilde{\zeta})\right)^2\mu_{N+K-1,K}(\tilde{\zeta}). \quad (17)$$

The term (16) is simple to estimate using just duality and then we get:

$$(16) \le \frac{2(N-1)}{N+K-2}\mathcal{E}_{N,K}f \le \frac{2d_m\gamma b}{N+K-2}\,\mathcal{D}_{N,K}f$$

where the last inequality follows from (12).

The estimate of (17) depends on the linear ordering $c : \mathrm{X}_n \to \{1, \dots, N\}$ of the graph X_n

$$(17) \ \leq 2K^2 \sum_{x \in \mathrm{X}_n} \mathrm{E}_{\nu_{n,K}} \mathrm{I}_{[\zeta(x)>0]} \Big(f(\zeta^{x,c^{-1}(c(x)+1)}) - f(\zeta) \Big)^2 \leq$$

$$\leq 2K^2 \sum_{x \in \mathrm{X}} |\gamma_{x,c(x)+1}| \sum_{u \in \mathrm{X}_n} \sum_{v:(u,v) \in \gamma_{x,c^{-1}(c(x)+1)}} \mathrm{E}_{\nu_{n,K}} \mathrm{I}_{[\eta(u)>0]} \Big(f(\eta) - f(\eta^{uv}) \Big)^2$$

$$= 2K^2 \sum_{u \in \mathrm{X}_n} \sum_{v:(u,v) \in E_n} \mathrm{E}_{\nu_{n,K}} \mathrm{I}_{[\eta(u)>0]} \Big(f(\eta) - f(\eta^{uv}) \Big)^2 \frac{1}{2} \sum_{x:\gamma_{x,c^{-1}(c(x)+1)} \ni (u,v)} |\gamma_{x,c(x)+1}|$$

$$\leq 2K^2 a(c) b(c) \sum_{u \in \mathrm{X}_n} \sum_{v:(u,v) \in E_n} \mathrm{E}_{\nu_{n,K}} \mathrm{I}_{[\eta(u)>0]} \Big(f(\eta) - f(\eta^{uv}) \Big)^2.$$

where we denoted $a(c) = \max_{x \in \mathrm{X}} |\gamma_{x,c^{-1}(c(x)+1)}|$

and $b(c) = \max_{e \in E_n} |\{\gamma_{x,c^{-1}(c(x)+1)} \ni e : x \in \mathrm{X}_n\}$.

The sum in the last term is $2d_m$ multiple of the Dirichlet form (11) so we get the final estimate

$$\mathcal{E}^{EX}_{N+K-1,K} \tilde{f} \ \leq \ \Big(\frac{2d_m \ \gamma \ b}{N+K-2} + 4K^2 d_m a(c) b(c) \Big) \ \mathcal{D}_{n,K} f. \tag{18}$$

For a special case of the binary tree X_n of height n ($N = 2^{n+1} - 1$, the maximal degree $d_m = 3$, the maximal length of path $\gamma = 2n$, the measure of bottleneck $b = 2^n(2^n - 1)$) there exists an ordering of the tree for which $a(c) = n$ and $b(c) = 1$. This completes the proof of Proposition 2.2.

References

[1] Andjel E.D. (1982). *Invariant measures for the zero range process.* Ann.Probab. **10**, 525 – 547.

[2] Brémaud P. (1999). *Markov chains, gibbs fields, Monte Carlo simulation and queues.* Springer-Verlag, New York.

[3] Caputo P. (2002). *Spectral gap inequalities in product spaces with conservation laws.* Paper submitted for the proceedings of the conference "Stochastic analysis on large scale interacting systems", Japan.

[4] Fajfrová L. *Equilibrium behavior of Zero Range Process on binary tree.* PhD. thesis, in preparation.

[5] Landim C., Sethuraman S. and Varadhan S. (1996). *Spectral gap for zero range dynamics.* Ann.Probab. **24**, No. 4, 1871 – 1902.

[6] Liggett T.M. (1972). *Existence theorems for infinite particle systems.* Trans. Amer. Math. Soc. **165**, 471 – 481.

Acknowledgement: Supported by the Grant Agency of the Czech Republic under Grant No. 201/03/0455.

Address: L. Fajfrová, UTIA AV ČR, Pod Vodárenskou věží 4, 182 08 Praha 8

E-mail: `fajfrova@utia.cas.cz`

KLASIFIKAČNÍ PRAVIDLA PRO ELIPTICKY VRSTEVNICOVÁ ROZDĚLENÍ

Marie Forbelská

Klíčová slova: Diskriminační analýza, neparametrické a semiparametrické odhady hustot, součinová jádra, hraniční jádra, jádra s proměnlivou šířkou oken, elipticky vrstevnicové rozdělení, M-odhady.

Abstrakt: Diskriminační analýza používá klasifikační postupy, s jejichž pomocí se objekt popsaný vícerozměrným znakem zařadí do jedné z konečného počtu existujících tříd. Postup klasifikace je založen na určitých předpokladech o vlastnostech objektů, např. na předpokladu o normálním rozdělení náhodného vektoru, charakterizujícího objekt. Pro tento případ byla odvozena klasická lineární a kvadratická diskriminační pravidla. V příspěvku bude ukázáno, že obdobná lineární a kvadratická (t.j. parametrická) klasifikační pravidla lze najít i pro mnohem širší třídu vícerozměrných rozdělení, a to pro tzv. elipticky vrstevnicová rozdělení. Pro tento typ rozdělení bude také popsán neparametrický i semiparametrický přístup vycházející z jádrových odhadů podmíněných hustot figurujících v klasifikačních pravidlech. Popsané klasifikační postupy budou demonstrovány na příkladu simulovaných dat z vícerozměrného Studentova rozdělení.

1 Úvod

Předpokládejme, že množina $\mathcal{G}$ nějakých objektů (jedinců) sestává z k rozlišitelných tříd, skupin či populací $\mathcal{G}_1, \ldots, \mathcal{G}_k$, které jsou po dvou disjunktní a tvoří rozklad množiny objektů $\mathcal{G} = \bigcup_{j=1}^{k} \mathcal{G}_j$ ($\mathcal{G}_i \cap \mathcal{G}_j = \emptyset, i \neq j$). Na každém objektu nás budou zajímat dva statistické znaky X a $\mathbf{Y} = (Y_1, \cdots, Y_m)'$. Znak X má obor hodnot $\{1, \ldots, k\}$ a určuje příslušnost objektu k dané třídě a m-rozměrný znak $\mathbf{Y}$ objekt charakterizuje.

Nechť $\mathbf{W} = (X, \mathbf{Y}')'$ je náhodný vektor definovaný na nějakém pravděpodobnostním prostoru $(\Omega, \mathcal{A}, P)$, který má vzhledem k součinové míře $\mu = \nu_X \times \mu_{\mathbf{Y}}$ hustotu $f_{X\mathbf{Y}}(j, \mathbf{y})$, kde ν_X je čítací míra a $\mu_{\mathbf{Y}}$ je Lebesguova míra, přičemž $f_{X\mathbf{Y}}(j, \mathbf{y}) = p_j f_j(\mathbf{y})$, kde p_j jsou tzv. *apriorní pravděpodobnosti* ($p_j > 0$, $p_1 + \cdots + p_k = 1$) a $f_j(\mathbf{y})$ jsou hustoty vzhledem k Lebesguově míře. Zřejmě $f_j(\mathbf{y})$ je podmíněná hustota $\mathbf{Y}$, když $X = j$ ($j = 1, \ldots, k$, $\mathbf{y} \in \mathbb{R}^m$).

Na základě pozorování $\mathbf{y}$ náhodného vektoru $\mathbf{Y}$ chceme objekt zařadit do jedné z k tříd. Použijeme následující *bayesovské klasifikační pravidlo* (viz např. [3]): pokud při daném $\mathbf{Y} = \mathbf{y}$ pro všechna $j \neq t$ ($t, j = 1, \ldots, k$) platí

$$p_t f_t(\mathbf{y}) \geq p_j f_j(\mathbf{y}), \tag{1}$$

pak objekt zařadíme do t-té třídy.

Při praktickém provádění diskriminační analýzy máme k dispozici k souborů objektů, přičemž víme, který objekt do které třídy patří. Jde o tzv.

trénovací data, na základě kterých provádíme odhady neznámých apriorních pravděpodobností a podmíněných hustot. Pro $j = 1, \ldots, k$ označme počet objektů v j-tém souboru n_j a $N = n_1 + \cdots + n_k$, relativní četnosti $\hat{p}_j = \frac{n_j}{N}$, realizace $\mathbf{y}_{j1}, \ldots, \mathbf{y}_{jn_j}$, vektory výběrových průměrů $\bar{\mathbf{y}}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} \mathbf{y}_{ji}$, matice součtů kvadrátů a součinů $\mathbf{Q}_j = \sum_{i=1}^{n_j} (\mathbf{y}_{ji} - \bar{\mathbf{y}}_j)(\mathbf{y}_{ji} - \bar{\mathbf{y}}_j)'$, výběrové kovarianční matice $\mathbf{S}_j = \frac{\mathbf{Q}_j}{n_j - 1}$ a společnou výběrovou kovarianční matici $\mathbf{S} = \sum_{j=1}^{k} \frac{n_j - 1}{N - k} \mathbf{S}_j$.

2 Parametrická klasifikační pravidla pro normální rozdělení

Pokud pro $j = 1, \ldots, k$ podmíněné rozdělení náhodného vektoru $\mathbf{Y}$ za podmínky, že $X = j$, je m-rozměrné normální $N_m(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ s vektorem středních hodnot $\boldsymbol{\mu}_j$ a kovarianční maticí $\boldsymbol{\Sigma}_j$, resp. $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_k$, pak objekt popsaný vícerozměrným znakem $\mathbf{y}$ zařadíme do t-té třídy, pokud pro všechna $j \neq t$ $(t, j = 1, \ldots, k)$ platí

$$\boxed{D_t(\mathbf{y}) \geq D_j(\mathbf{y})},\ D_j(\mathbf{y}) = -\tfrac{1}{2} \ln |\boldsymbol{\Sigma}_j| - \tfrac{1}{2} (\mathbf{y} - \boldsymbol{\mu}_j)' \boldsymbol{\Sigma}_j^{-1} (\mathbf{y} - \boldsymbol{\mu}_j) + \ln p_j, \quad (2)$$

$$\text{resp. } \boxed{d_t(\mathbf{y}) \geq d_j(\mathbf{y})},\ d_j(\mathbf{y}) = \left(\mathbf{y} - \tfrac{1}{2} \boldsymbol{\mu}_j\right)' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_j + \ln p_j. \quad (3)$$

Diskriminační metoda založená na pravidle (2), resp. (3), v němž neznámé parametry $\boldsymbol{\mu}_j$, $\boldsymbol{\Sigma}_j$, p_j po řadě nahradíme $\bar{\mathbf{y}}_j$, $\mathbf{S}_j$ a $\hat{p}_j$, resp. $\bar{\mathbf{y}}_j$, $\mathbf{S}$ a $\hat{p}_j$, se nazývá *klasická* ***kvadratická***, resp. ***lineární***, *diskriminační analýza.*

3 Parametrická klasifikační pravidla pro elipticky vrstevnicová rozdělení

Rozšíření lineárních a kvadratických diskriminačních pravidel ze třídy normálně rozdělených klasifikačních znaků na třídu elipticky vrstevnicových bylo inspirováno článkem [2] s kvadratickými diskriminačními pravidly pro vícerozměrná Pearsonova rozdělení typu II a VII a faktem, že typické vlastnosti vícerozměrného normálního rozdělení lze zobecnit na mnohem širší třídu elipticky vrstevnicových rozdělení, kam obě jmenovaná rozdělení patří. Při definici elipticky vrstevnicového rozdělení vyjdeme ze značení používaného v [6].

Definice 3.1. *Řekneme, že m-rozměrný náhodný vektor $\mathbf{Y}$ má* elipticky vrstevnicové rozdělení *(ECD rozdělení) s parametry $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ a ϕ, kde $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)' \in \mathbb{R}^m$, $\boldsymbol{\Sigma}$ je pozitivně semidefinitní matice řádu $(m \times m)$, ϕ je nějaká funkce $\phi\colon [0, \infty) \mapsto \mathbb{R}$, jestliže charakteristická funkce $\psi(\mathbf{t})$ náhodného vektoru $\mathbf{Y} = (Y_1, \ldots, Y_m)'$ má tvar $\psi(\mathbf{t}) = \exp(i\mathbf{t}'\boldsymbol{\mu})\ \phi(\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t})$. Pak píšeme $\mathbf{Y} \sim EC_m(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \phi)$. Speciálně, jestliže $\boldsymbol{\mu} = 0$ a $\boldsymbol{\Sigma} = \mathbf{I}_m$, $EC_m(\mathbf{0}, \mathbf{I}_m, \phi)$ se nazývá* sférické rozdělení *a píšeme $\mathbf{Y} \sim S_m(\phi)$.*

Pokud $\mathbf{Y} \sim EC_m(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \phi)$, pak platí (viz např. [6] nebo [5]):

(i) Pokud hodnost matice $rk(\boldsymbol{\Sigma}) = k$, pak $\mathbf{Y} \sim EC_m(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \phi)$ tehdy a jen tehdy, když $\mathbf{Y}$ má *stochastickou reprezentaci* $\mathbf{Y} = \boldsymbol{\mu} + R\mathbf{A}\mathbf{U}_k$, kde $\mathbf{U}_k$

je náhodný vektor s rovnoměrným rozdělením na povrchu jednotkové koule v $\mathbb{R}^k$, R je nezáporná náhodná veličina, R a $\mathbf{U}_k$ jsou nezávislé a $\mathbf{\Sigma}=\mathbf{A}\mathbf{A}'$ je rozklad matice $\mathbf{\Sigma}$ (tj. $\mathbf{A}$ je $(m\times k)$ matice hodnosti k).

(ii) $E\mathbf{Y}<\infty$ (resp. $D\mathbf{Y}<\infty$) tehdy a jen tehdy, když $ER<\infty$ (resp. $ER^2<\infty$), přičemž $E\mathbf{Y}=\boldsymbol{\mu}$ (resp. $D\mathbf{Y}=-2\phi'(0)\mathbf{\Sigma}=\frac{ER^2}{rk(\mathbf{\Sigma})}\mathbf{\Sigma}$).

(iii) Obecně náhodný vektor $\mathbf{Y}$ nemusí mít hustotu vzhledem k Lebesguově míře, avšak je-li absolutně spojitý, musí být matice $\mathbf{\Sigma}=\mathbf{A}\mathbf{A}'$ pozitivně definitní a platí: $\mathbf{Y}=\boldsymbol{\mu}+R\mathbf{A}\mathbf{U}_m\sim EC_m(\boldsymbol{\mu},\mathbf{\Sigma},\phi)$ je absolutně spojitého typu s *hustotou* $f_{\mathbf{Y}}(\mathbf{y})$, právě když pro $\mathbf{y}\in\mathbb{R}^m$ je hustota ve tvaru $f_{\mathbf{Y}}(\mathbf{y})=c_m\,|\mathbf{\Sigma}|^{-\frac{1}{2}}\,g\left((\mathbf{y}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right)$, kde tzv. *generátor hustoty* $g:[0,\infty)\mapsto[0,\infty)$ a $\int_0^\infty r^{\frac{m}{2}-1}g(r)dr<\infty$, $c_m=\frac{\Gamma\left(\frac{m}{2}\right)}{\pi^{\frac{m}{2}}\int_0^\infty r^{\frac{m}{2}-1}g(r)dr}$. Přitom hustotu nezáporné náhodné veličiny lze také vyjádřit pomocí generátoru g, tj.

$$f_{R^2}(s)=\pi^{\frac{m}{2}}\Gamma\left(\tfrac{m}{2}\right)^{-1}s^{\frac{m}{2}-1}c_m g(s). \tag{4}$$

A naopak, označíme-li $\rho^2(\mathbf{y})=(\mathbf{y}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})$, pak

$$f_{\mathbf{Y}}(\mathbf{y})=|\mathbf{\Sigma}|^{-\frac{1}{2}}\,\Gamma\left(\tfrac{m}{2}\right)\pi^{-\frac{m}{2}}\;f_{R^2}\left(\rho^2(\mathbf{y})\right)\rho^{2-m}(\mathbf{y}). \tag{5}$$

(iv) Mějme realizace náhodného výběru $\mathbf{y}_1,\dots,\mathbf{y}_n$ z rozdělení $EC_m(\boldsymbol{\mu},\mathbf{\Sigma},\phi)$ s generátorem hustoty g, který je navíc nerostoucí a spojitý. Označíme-li výběrový vektor středních hodnot $\bar{\mathbf{y}}=\frac{1}{n}\sum_{i=1}^n\mathbf{y}_i$, matici součtů kvadrátů a součinů $\mathbf{Q}=\sum_{i=1}^n(\mathbf{y}_i-\bar{\mathbf{y}})(\mathbf{y}_i-\bar{\mathbf{y}})'$, pak *maximálně věrohodné odhady* parametrů $\boldsymbol{\mu}$ a $\mathbf{\Sigma}$ jsou rovny $\hat{\boldsymbol{\mu}}_{MLE}=\bar{\mathbf{y}}$ a $\hat{\mathbf{\Sigma}}_{MLE}=\frac{m}{z_g}\mathbf{Q}$, kde $z_g>0$ je maximum funkce $l(z)=z^{\frac{mn}{2}}g(z)$. Je-li navíc g diferencovatelná, z_g je řešením $g'(z)+\frac{mn}{2z}g(z)=0$ $(z\geq 0)$, nebo ekvivalentně řešením rovnice $W_g(z)+\frac{mn}{2z}=0$ $(z\geq 0)$, kde $W_g(z)=\frac{d\ln(g(z))}{dz}=\frac{g'(z)}{g(z)}$.

Předpokládejme, že pro $j=1,\dots,k$ jsou $f_j(\mathbf{y})$ m-rozměrnými hustotami

(A) <u>*Pearsonova rozdělení typu II*</u>, značíme $MPII_m(q,\boldsymbol{\mu}_j,\mathbf{\Sigma}_j)$, $q\in\mathbb{R}$, $q>-1$,

(B) <u>*Pearsonova rozdělení typu VII*</u>, značíme $MPVII_m(q,p,\boldsymbol{\mu}_j,\mathbf{\Sigma}_j)$, $p>0$, $q>\frac{m}{2}$ (pro $q=\frac{m+p}{2}$, $p\in\mathbb{N}$ dostaneme *vícerozměrné t-rozdělení*, jestliže navíc $p=1$, pak jde o *vícerozměrné Cauchyovo rozdělení*),

(C) <u>*Kotzova rozdělení*</u>, značíme $MK_m(q,p,s,\boldsymbol{\mu}_j,\mathbf{\Sigma}_j)$, $p,s>0$, $2q+m>2$ (pro $s=1$, $q=1$, $p=\frac{1}{2}$ dostaneme *vícerozměrné normální rozdělení*),

pro které

(A) $g(u)=(1-u)^q$ pro $|u|\leq 1$, $c_m=\frac{\Gamma\left(\frac{m}{2}+q+1\right)}{\Gamma(q+1)\pi^{\frac{m}{2}}}$, $\widehat{\mathbf{\Sigma}}_j^{MLE}=\frac{2q+mn}{n}\mathbf{Q}_j$,

(B) $g(u)=\left(1+\frac{1}{p}u\right)^{-q}$, $c_m=\frac{\Gamma(q)}{\Gamma(q-\frac{m}{2})(\pi p)^{\frac{m}{2}}}$, $\widehat{\mathbf{\Sigma}}_j^{MLE}=\frac{2q-mn}{np}\mathbf{Q}_j$,

(C) $g(u)=u^{q-1}e^{-pu^s}$, $c_m=\frac{sp^{\frac{2q+m-2}{2s}}\Gamma\left(\frac{m}{2}\right)}{\Gamma\left(\frac{2q+m-2}{2s}\right)\pi^{\frac{m}{2}}}$, $\widehat{\mathbf{\Sigma}}_j^{MLE}=m\left(\frac{2ps}{2q+mn-2}\right)^{\frac{1}{s}}\mathbf{Q}_j$.

Klasifikační pravidlo (1) je pro $j \neq t$ $(t, j = 1, \ldots, k)$ ekvivalentní s pravidly

(A) $p_t^{\frac{1}{q}} f_t^{\frac{1}{q}}(\mathbf{y}) = c_m^{\frac{1}{q}} D_t(\mathbf{y}) \geq p_j^{\frac{1}{q}} f_j^{\frac{1}{q}}(\mathbf{y}) = c_m^{\frac{1}{q}} D_j(\mathbf{y})$,
kde $D_j(\mathbf{y}) = p_j^{\frac{1}{q}} |\boldsymbol{\Sigma}_j|^{-\frac{1}{2q}} \left[1 - (\mathbf{y} - \boldsymbol{\mu}_j)' \boldsymbol{\Sigma}_j^{-1} (\mathbf{y} - \boldsymbol{\mu}_j)\right]$,

(B) $-p_t^{-\frac{1}{q}} f_t^{-\frac{1}{q}}(\mathbf{y}) = c_m^{-\frac{1}{q}} D_t(\mathbf{y}) \geq -p_j^{-\frac{1}{q}} f_j^{-\frac{1}{q}}(\mathbf{y}) = c_m^{-\frac{1}{q}} D_j(\mathbf{y})$,
kde $D_j(\mathbf{y}) = -p_j^{-\frac{1}{q}} |\boldsymbol{\Sigma}_j|^{\frac{1}{2q}} \left[1 + \frac{1}{p} (\mathbf{y} - \boldsymbol{\mu}_j)' \boldsymbol{\Sigma}_j^{-1} (\mathbf{y} - \boldsymbol{\mu}_j)\right]$,

(C) $\ln(p_t f_t(\mathbf{y})) = D_t(\mathbf{y}) + \ln c_m \geq \ln(p_j f_j(\mathbf{y})) = D_j(\mathbf{y}) + \ln c_m$,
kde $D_j(\mathbf{y}) = \ln p_j - \frac{1}{2} \ln |\boldsymbol{\Sigma}_j| - p(\mathbf{y} - \boldsymbol{\mu}_j)' \boldsymbol{\Sigma}_j^{-1} (\mathbf{y} - \boldsymbol{\mu}_j)$ a $q = s = 1$.

Jestliže zanedbáme konstanty c_m a neznámé parametry p_j, $\boldsymbol{\mu}_j$ a $\boldsymbol{\Sigma}_j$ po řadě nahradíme odhady $\hat{p}_j$, $\widehat{\boldsymbol{\mu}}_j = \bar{\mathbf{y}}_j$ a $\widehat{\boldsymbol{\Sigma}}_j = \widehat{\boldsymbol{\Sigma}}_j^{MLE}$, dostaneme pro všechna $j \neq t$ $(t, j = 1, \ldots, k)$ ***kvadratické*** *diskriminační pravidlo* (*ECD–QDA*)

$$\boxed{\hat{D}_t(\mathbf{y}) \geq \hat{D}_j(\mathbf{y})}, \text{ kde } \hat{D}_j(\mathbf{y}) = A_j + B_j(\mathbf{y} - \widehat{\boldsymbol{\mu}}_j)' \widehat{\boldsymbol{\Sigma}}_j^{-1} (\mathbf{y} - \widehat{\boldsymbol{\mu}}_j), \qquad (6)$$

přičemž
(A) $A_j = \hat{p}_j^{\frac{1}{q}} |\widehat{\boldsymbol{\Sigma}}_j|^{-\frac{1}{2q}}$, $B_j = \hat{p}_j^{\frac{1}{q}} |\widehat{\boldsymbol{\Sigma}}_j|^{-\frac{1}{2q}}$,
(B) $A_j = \hat{p}_j^{-\frac{1}{q}} |\widehat{\boldsymbol{\Sigma}}_j|^{\frac{1}{2q}}$, $B_j = -\hat{p}_j^{-\frac{1}{q}} \frac{1}{p} |\widehat{\boldsymbol{\Sigma}}_j|^{\frac{1}{2q}}$,
(C) $A_j = \ln \hat{p}_j - \frac{1}{2} \ln |\widehat{\boldsymbol{\Sigma}}_j|$, $B_j = p$ a $q = s = 1$.

Pokud navíc $\boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$ a odhadneme $\widehat{\boldsymbol{\Sigma}} = \frac{1}{N} \sum_{j=1}^{k} n_j \widehat{\boldsymbol{\Sigma}}_j^{MLE}$, dostaneme pro $j \neq t$ $(t, j = 1, \ldots, k)$ ***lineární*** *diskriminační pravidlo* (*ECD–LDA*)

$$\boxed{\hat{d}_t(\mathbf{y}) \geq \hat{d}_j(\mathbf{y})}, \text{ kde } \hat{d}_j(\mathbf{y}) = a_j + b_j \left(\mathbf{y} - \tfrac{1}{2} \widehat{\boldsymbol{\mu}}_j\right)' \widehat{\boldsymbol{\Sigma}}^{-1} \widehat{\boldsymbol{\mu}}_j, \qquad (7)$$

přičemž
(A), (B) $a_j = 0$, $b_j = 1$ a $p_1 = \cdots = p_k$,
(C) $a_j = \ln \hat{p}_j$, $b_j = 2p$ a $q = s = 1$.

4 Neparametrická klasifikační pravidla

Podmíněné hustoty $f_j(\mathbf{y})$ $(j = 1, \ldots, k)$ z klasifikačního pravidla (1) nejprve odhadneme pomocí tzv. *součinových jader*, která jsou součinem m jader jedné proměnné: $\hat{f}_j(\mathbf{y}) = \prod_{t=1}^{m} \hat{f}_{jt}(y_t) = \prod_{t=1}^{m} \frac{1}{n_j} \sum_{i=1}^{n_j} \frac{1}{h_{jt}} K\left(\frac{y_t - y_{ji,t}}{h_{jt}}\right)$, $\mathbf{y} = (y_1, \ldots, y_m)' \in \mathbb{R}^m$, $\mathbf{y}_{ji} = (y_{ji,1}, \ldots, y_{ji,m})' \in \mathbb{R}^m$. *Jádrem* K rozumíme libovolnou symetrickou a ohraničenou funkci, pro niž $\int_{-\infty}^{\infty} K(y) dy = 1$ a $\lim_{y \to \pm\infty} |y| K(y) = 0$. Pro posloupnosti kladných *vyhlazovacích parametrů* (též *šířek oken*) $\{h_{jt} = h_{jt}(n_j)\}_{n_j=1}^{\infty}$ požadujeme, aby $\lim_{n_j \to \infty} h_{jt}(n_j) = 0$ a $\lim_{n_j \to \infty} n_j h_{jt}(n_j) = \infty$.

V práci bude použité buď *Gaussovo jádro* $K(y) = (2\pi)^{-1/2} e^{-y^2/2}$ nebo tzv. *polynomická jádra* $K(y) = \kappa_{rs}(1 - |y|^r)^s I_{[-1,1]}(y)$, kde $\kappa_{rs} = \frac{r}{2B(s+1, 1/r)}$, $r > 0$, $s \geq 0$ a $I_{[a,b]}(y) = \begin{cases} 1 & y \in [a, b] \\ 0 & \text{jinak} \end{cases}$ (pro $r = 2$ a $s = 1, 2, 3$ po řadě dostaneme tzv. *Epanechnikovo*, *biweight* a *triweight* jádro).

Vhodnou metodou volby vyhlazovacího parametru se v diskriminační analýze ukázala např. jednoduchá a rychlá metoda založená na principu horní hranice, tzv. přehlazení (*oversmoothing*) či maximální vyhlazení (*maximal smoothing*), více viz [4] nebo [5].

5 Semiparametrická klasifikační pravidla

Budeme-li předpokládat, že $f_j(\mathbf{y})$ $(j=1,\ldots,k)$ jsou podmíněné hustoty blíže neurčeného elipticky vrstevnicového rozdělení, pak lze s využitím vztahu (5) tyto vícerozměrné hustoty odhadovat např. pomocí jádrových odhadů jednorozměrných hustot $f_{R_j^2}(\rho^2(\mathbf{y}))$. Ve snaze eliminovat vliv odlehlých pozorování, v odhadech $r_{ji}^2=\widehat{\rho}_j^2(\mathbf{y}_{ji})=(\mathbf{y}_{ji}-\widehat{\boldsymbol{\mu}}_j)'\widehat{\boldsymbol{\Sigma}}_j^{-1}(\mathbf{y}_{ji}-\widehat{\boldsymbol{\mu}}_j)$ místo odhadů $\bar{\mathbf{y}}_j$ a $\mathbf{S}_j$ použijeme *Huberovy M-odhady* získané z rekurentních vztahů: $\widehat{\boldsymbol{\mu}}_j^{(0)}=\frac{1}{n_j}\sum_{i=1}^{n_j}\mathbf{y}_{ji}$, $\widehat{\boldsymbol{\Sigma}}_j^{(0)}=\frac{1}{n_j}\sum_{i=1}^{n_j}\left(\mathbf{y}_{ji}-\widehat{\boldsymbol{\mu}}_j^{(0)}\right)\left(\mathbf{y}_{ji}-\widehat{\boldsymbol{\mu}}_j^{(0)}\right)'$, $r_{ji}^{(\nu)}=\sqrt{\left(\mathbf{y}_{ji}-\widehat{\boldsymbol{\mu}}_j^{(\nu-1)}\right)'\left(\widehat{\boldsymbol{\Sigma}}_j^{(\nu-1)}\right)^{-1}\left(\mathbf{y}_{ji}-\widehat{\boldsymbol{\mu}}_j^{(\nu-1)}\right)}$, $w_{ji}^{(\nu)}=\frac{\psi\left(r_{ji}^{(\nu)}\right)}{r_{ji}^{(\nu)}}$, $\widehat{\boldsymbol{\mu}}_j^{(\nu+1)}=\frac{\sum_{i=1}^{n_j}w_{ji}^{(\nu)}\mathbf{y}_{ji}}{\sum_{i=1}^{n_j}w_{ji}^{(\nu)}}$, $\widehat{\boldsymbol{\Sigma}}_j^{(\nu+1)}=\frac{\sum_{i=1}^{n_j}\left(w_{ji}^{(\nu)}\right)^2\left(\mathbf{y}_{ji}-\widehat{\boldsymbol{\mu}}_j^{(\nu)}\right)\left(\mathbf{y}_{ji}-\widehat{\boldsymbol{\mu}}_j^{(\nu)}\right)'}{\sum_{i=1}^{n_j}\left(w_{ji}^{(\nu)}\right)^2}$, přičemž $\psi(s)=\begin{cases}s & |s|\le c,\\ \operatorname{sign}(s)c & |s|>c.\end{cases}$ (více [8], kde se doporučuje $c=2$ a $\nu=1,\ldots,5$).

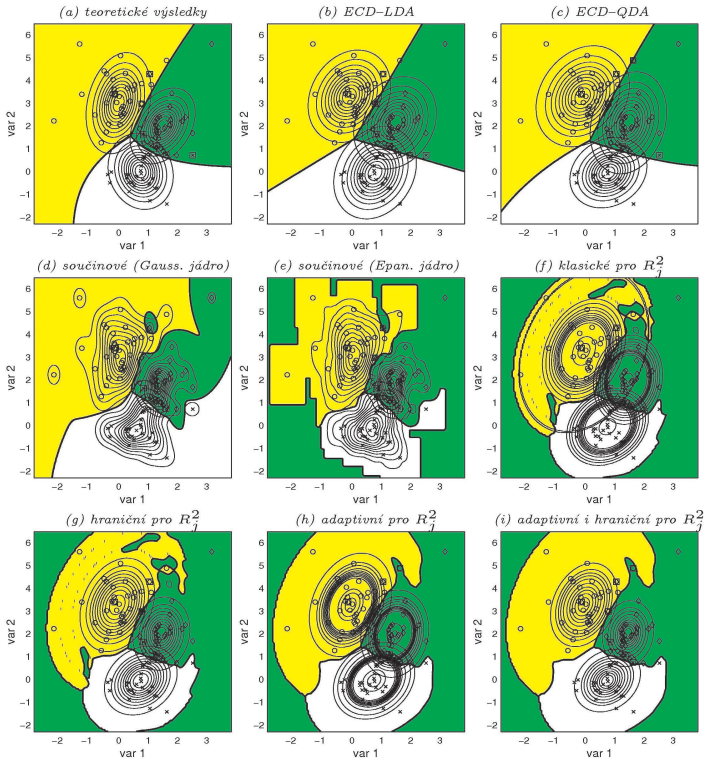
Tím, že stačí uvažovat pouze jednorozměrné jádrové odhady nezáporných náhodných veličin R_j^2, na jedné straně pro velká m ušetříme čas nutný k hledání optimální šířky vyhlazovacích parametrů jednotlivých dimenzí, na druhé straně však nevystačíme s klasickými odhady typu $\tilde{f}_{R_j^2}(s)=\sum_{i=1}^{n_j}\frac{1}{n_j h_j}K\left(\frac{s-r_{ji}^2}{h_j}\right)$, neboť je třeba se vyrovnat s hraničními efekty.
Použijeme proto tzv. *lineární hraniční jádro* typu $(l_x+m_x u)K(u)$ (viz např. [7]) a dostaneme odhad $\tilde{f}_{R_j^2}^b(s)=\frac{1}{n_j}\sum_{i=1}^{n_j}\frac{1}{h_j}\left[l_{\frac{s}{h_j}}+m_{\frac{s}{h_j}}\frac{s-r_{ji}^2}{h_j}\right]K\left(\frac{s-r_{ji}^2}{h_j}\right)$, kde K je jádro, jehož nosič je $[-u_K,u_K]$, $l_u=\frac{a_2(u)}{a_0(u)a_2(u)-a_1^2(u)}$, $m_u=\frac{-a_1(u)}{a_0(u)a_2(u)-a_1^2(u)}$ a $a_l(u)=\int_{-u_K}^{\min\{u,u_K\}}u^l K(u)du$.

Chceme-li kvalitnější jádrové odhady, můžeme použít také jádrové odhady s proměnlivou šířkou oken $\tilde{f}_{R_j^2}^a(s)=\frac{1}{n_j}\sum_{i=1}^{n_j}\frac{1}{h_j\lambda_{ji}}K\left(\frac{s-r_{ji}^2}{h_j\lambda_{ji}}\right)$, kde $\lambda_{ji}=\left(\tilde{f}_{R_j^2}(r_{ji}^2)/g_j\right)^{-\alpha_j}$, $0\le\alpha_j\le 1$, $\ln g_j=\frac{1}{n_j}\sum_{i=1}^{n_j}\ln\tilde{f}_{R_j^2}(r_{ji}^2)$, nazvěme je krátce *adaptivní* (viz [10], kde se doporučuje volit $\alpha_j=\frac{1}{2}$).

6 Příklad s vícerozměrným Studentovým rozdělením

K ilustraci popsaných klasifikačních metod použijeme simulovaná data z vícerozměrného t-rozdělení, tj. z $MPVII_m(q,p,\boldsymbol{\mu},\boldsymbol{\Sigma})$ pro $p\in\mathbb{N}$, $q=\frac{m+2}{2}$. Při generování pseudonáhodných čísel $\mathbf{y}=(y_1,\ldots,y_m)'$ využijeme vztah $\mathbf{y}=\boldsymbol{\mu}+[p(1-z)/z]^{1/2}\mathbf{A}\frac{\mathbf{v}}{\|\mathbf{v}\|}$, kde $\boldsymbol{\Sigma}=\mathbf{A}\mathbf{A}'$, z je pseudonáhodné číslo z rozdělení $\beta\left(q-\frac{m}{2},\frac{m}{2}\right)$ a $\mathbf{v}=(v_1,\ldots,v_m)'$ jsou nezávislá pseudonáhodná čísla ze standardizovaného normálního rozdělení (odvození viz [5]). Pomocí simulací tak vytvoříme trénovací soubor, který je směsí tří dvourozměrných t-rozdělení s 5 stupni volnosti pro $n_1=n_2=n_3=30$, $\boldsymbol{\mu}_1=(0,3)'$, $\boldsymbol{\mu}_2=(0.75,0)'$, $\boldsymbol{\mu}_3=(1.5,2)'$, $\boldsymbol{\Sigma}_1=\begin{pmatrix}0.25 & 0.125\\ 0.125 & 1\end{pmatrix}$, $\boldsymbol{\Sigma}_2=\begin{pmatrix}0.25 & -0.046875\\ -0.046875 & 0.5625\end{pmatrix}$ a $\boldsymbol{\Sigma}_3=\begin{pmatrix}0.25 & 0.13125\\ 0.13125 & 0.5625\end{pmatrix}$. Na základě trénovacích dat byla zkonstruována jednotlivá diskriminační pravidla. Na obrázku 1 jsou vykresleny výsledné hraniční

křivky, které dělí $\mathbb{R}^2$ na tři oblasti, dále vrstevnicové grafy příslušných hustot (teoretických či odhadnutých) spolu se simulovanými daty (symboly $\circ$, $\times$ a $\diamond$ značí postupně body z první, druhé a třetí skupiny). Podle toho, ve které oblasti bod leží, je klasifikován. Pomocí této *resubstituce* lze posoudit účinnost klasifikačního pravidla např. díky % chybně klasifikovaných prvků.



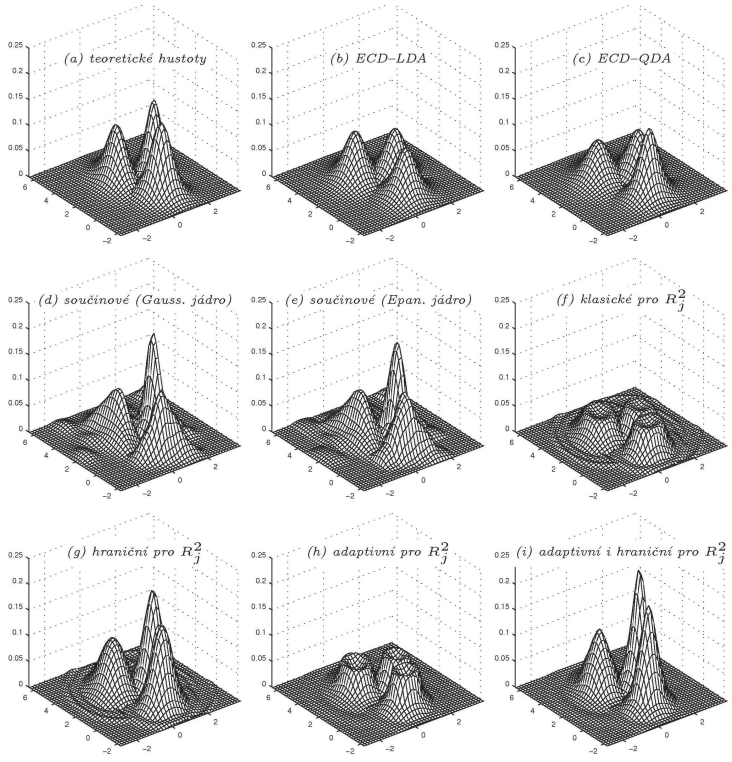
Obrázek 1: Hraniční křivky spolu s vrstevnicovými grafy podmíněných hustot.

Poznámky k obrázku 1:

(1) Porovnáme-li teoretické výsledky klasifikace s parametrickými klasifikačními postupy, jako jsou *ECD–LDA* (je použita i když nejsou splněny předpoklady o rovnosti kovariančních matic), tak *ECD–QDA*, pak se v tomto konkrétním případě výsledky příliš neliší, viz obrázky (a), (b) a (c).

(2) I když obecně na kvalitu jádrového odhadu má větší vliv volba vyhlazovacího parametru než volba jádra, ze tvaru hraničních křivek na obrázcích (d) a (e) je zřejmé, že při použití součinových jader je vhodnější volit místo polynomického jádra s konečným nosičem (což je např. Epanechnikovo jádro) raději jádro mající neomezený nosič (např. Gaussovo jádro).

(3) Při semiparametrickém přístupu odhadu hustoty nezáporných náhodných veličin R_j^2 $(j=1,2,3)$ se ukázalo, že problémy spojené s hraničními efekty neovlivnily tvar hraničních křivek, viz obrázky (f) a (g), kdežto použití odhadů s proměnlivou šířkou oken se projevilo na jejich hladkosti, viz obrázky (h) a (i). Při odhadech hustot R_j^2 $(j=1,2,3)$ bylo vždy použito polynomické, a to Epanechnikovo jádro.



Obrázek 2: Parametrické, neparametrické a semiparametrické odhady hustot.

Pro názornou představu o rozdílu mezi teoretickými hustotami a jejími odhady je připraven obrázek 2 a tabulka 1.

Poznámky k obrázku 2 a tabulce 1:

(1) Nejmenší odchylky vykazují neparametrické metody odhadu pomocí součinových jader a semiparametrický odhad s lineárním hraničním jádrem, viz metody (d), (e) a (g).

(2) Největší vychýlení, které je vždy v okolí bodů $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ a $\boldsymbol{\mu}_3$, způsobují především hraniční efekty při odhadu hustot nezáporných náhodných veličin R_j^2 $(j = 1, 2, 3)$, viz metody (f) a (h). Mezi horší metody se překvapivě dostala i metoda (i) s proměnlivou šířkou okna s lineárním hraničním jádrem.

Tabulka 1: Maximální absolutní odchylky od teoretických hustot.

	1. sk.	2. sk.	3. sk.
(b)	0.0383	0.0502	0.0601
(c)	0.0457	0.0277	0.0615
(d)	0.0263	0.0394	0.0631
(e)	0.0259	0.0394	0.0532
(f)	0.0730	0.0929	0.1029
(g)	0.0287	0.0415	0.0471
(h)	0.0701	0.0871	0.0965
(i)	0.0289	0.0563	0.0842

Účinnost jednotlivých klasifikačních postupů lze vyhodnotit pomocí tzv. *konfusních matic*, % chybné klasifikace (získaného při resubstituci) a také času (v sekundách) potřebného k vytvoření klasifikačního pravidla a pro resubstituci.

(a) teoretické výsledky

Classification Results

Original group	Classified group 1	2	3	Total
1	30	0	0	30
2	0	26	4	30
3	3	0	27	30
Misclass. in %	0.00	13.33	10.00	7.78

(b) ECD–LDA

CPU–time = 0.01 s

Original group	Classified group 1	2	3	Total
1	30	0	0	30
2	0	27	3	30
3	3	1	26	30
Misclass. in %	0.00	10.00	13.33	7.78

(c) ECD–QDA

CPU–time = 0.02 s

Original group	Classified group 1	2	3	Total
1	29	0	1	30
2	0	27	3	30
3	3	0	27	30
Misclass. in %	3.33	10.00	10.00	7.78

(d) součinové (Gauss. jádro)

CPU–time = 0.531 s

Original group	Classified group 1	2	3	Total
1	30	0	0	30
2	0	27	3	30
3	1	0	29	30
Misclass. in %	0.00	10.00	3.33	4.44

(e) součinové (Epan. jádro)

CPU–time = 0.571 s

Original group	Classified group 1	2	3	Total
1	30	0	0	30
2	0	27	3	30
3	3	1	26	30
Misclass. in %	0.00	10.00	13.33	7.78

(f) klasické pro R_j^2

CPU–time = 0.261 s

Original group	Classified group 1	2	3	Total
1	30	0	0	30
2	0	28	2	30
3	3	0	27	30
Misclass. in %	0.00	6.67	10.00	5.56

(g) hraniční pro R_j^2

CPU–time = 0.32 s

Original group	Classified group 1	2	3	Total
1	30	0	0	30
2	0	28	2	30
3	3	0	27	30
Misclass. in %	0.00	6.67	10.00	5.56

(h) adaptivní pro R_j^2

CPU–time = 0.301 s

Original group	Classified group 1	2	3	Total
1	29	0	1	30
2	0	27	3	30
3	3	0	27	30
Misclass. in %	3.33	10.00	10.00	7.78

(i) adaptivní i hraniční pro R_j^2

CPU–time = 3.625 s

Original group	Classified group 1	2	3	Total
1	29	0	1	30
2	0	27	3	30
3	3	0	27	30
Misclass. in %	3.33	10.00	10.00	7.78

Tabulka 2: Konfusní matice, % chybné klasifikace a CPU čas.

Poznámky k tabulkám 2(a)-(i):

(1) Nejnižší % chybné klasifikace vykazuje neparametrická metoda se součinovým gaussovským jádrem, následuje semiparametrická metoda s klasickým i hraničním jádrem.

(2) Podle očekávání nejrychlejšími byly parametrické metody, pak semiparametrické (s výjimkou metody (i)) a nejpomalejšími byly parametrické metody se součinovým jádrem. Absolutně nejhůře dopadla metoda (i) s proměnlivou šířkou okna a s lineárním hraničním jádrem. Stálo by za zvážení, zda pro odhady hustot nezáporných náhodných veličin nepoužít místo symetrických jader jádra asymetrická, jako jsou gamma jádra, popř. IG (*Inverse Gaussian*) nebo RIG (*Reciprocal Inverse Gaussian*) jádra, jak se doporučuje v pracích [1] a [9].

Reference

[1] Chen S.X. (2000). *Probability density functions using Gamma kernels.* Ann. Inst. Stat. Math. **52**, 471 – 480.

[2] Cooper P.W. (1963). *Statistical classification with quadratic forms.* Biometrika **50**, 439 – 448.

[3] Forbelská M. (2001). *Neparametrická diskriminační analýza*. Proceedings ROBUST'2000, Nečtiny, 50 – 58.

[4] Forbelská M. (2002). *A Comparison of some parametric and nonparametric discrimination procedures.* In Summer School DATASTAT'01. FOLIA, Masaryk University, Faculty of Science, Mathematica **11**, 53 – 81.

[5] Forbelská M. (2003). *Parametric and nonparametric discrimination.* PhD Thesis. University of Ostrava, Fac. of Science, Dpt. of Mathematics, Czech R.

[6] Gupta A.K., Varga T. (1993). *Elliptically contoured models in statistics.* Kluwer Academic Publishers, Dordrecht.

[7] Jones M.C, Foster P.J. (1996). *A simple nonnegative boundary correction method for kernel density estimation.* Statistika Sinica, 1005 – 1013.

[8] Randles R.H, Broffitt J.D., Ramberg J.S., Hogg R.V. (1978). *Generalized linear and quadratic discriminant functions using robust estimates.* Journal of the American Statistical Association **73**, no. 563, 564 – 568.

[9] Scaillet O. (2004). *Density estimation using inverse and reciprocal inverse Gaussian kernels.* J. of Nonpar. Statist. **16**, 217 – 226.

[10] Silverman B.W. (1993). *Density estimation for statistics and data analysis.* Chapman & Hall, New York.

Poděkování: Příspěvek vznikl s podporou výzkumného záměru MSM: J07/98:143100001.

Adresa: M. Forbelská, Katedra aplikované matematiky Přírodovědecké fakulty Masarykovy university v Brně, Janáčkovo nám. 2a, 662 95 Brno

E-mail: `forbel@math.muni.cz`

NEPARAMETRICKÉ BAYESOVSKÉ ODHADY V KOZIOLOVĚ-GREENOVĚ MODELU NÁHODNÉHO CENZOROVÁNÍ

Michal Friesl

Klíčová slova: Funkce spolehlivosti, neparametrické bayesovské odhady, náhodné cenzorování, Koziolův-Greenův model, gama proces.

Abstrakt: V příspěvku je odvozen neparametrický bayesovský odhad funkce spolehlivosti při Koziolově-Greenově modelu náhodného cenzorování, jako apriorní rozdělení kumulativní intenzity poruch se předpokládá gama proces.

1 Úvod

Mezi bayesovskými metodami jsou jako neparametrické označovány ty, které počítají s apriorními rozděleními nikoli pro konečný počet parametrů určitého rozdělení, ale pro obvykle nekonečněrozměrné parametry, jako je např. distribuční funkce. Ferguson [2] představil jako apriorní model na prostoru pravděpodobnostních měr rozdělení Dirichletova procesu a od té doby byla zkoumána jako apriorní celá řada procesů (beta, gama, smíšené, zprava neutrální) s uplatněním v různých modelech analýzy dat o přežití či teorie spolehlivosti, včetně cenzorování. V následujícím textu se budeme věnovat odhadu funkce přežití v Koziolově-Greenově modelu náhodného cenzorování [7]. V tomto modelu je rozdělení cenzoru svázáno s rozdělením cenzorované veličiny a nelze proto použít tradiční odhady.

Ještě předtím ale stručně přiblížíme princip neparametrických bayesovských metod, některá apriorní rozdělení a uplatnění těchto metod v analýze dat o přežití obecně. Z přehledných článků shrnujících tuto problematiku jmenujme [4] či [11].

2 Neparametrická apriorní rozdělení

Uvažujme pozorování X s hodnotami v měřitelném prostoru $(\mathcal{X}, \mathcal{A})$, jejichž rozdělení Q neznáme. Při parametrickém přístupu předpokládáme, že rozdělení Q je určitého typu, že pochází z určité úzké rodiny rozdělení parametrizované konečným počtem parametrů, např. že je normální, $Q = Q_{\mu,\sigma^2} = \mathrm{N}(\mu, \sigma^2)$. V bayesovské statistice považujeme parametry za náhodné veličiny, apriorní informace o nich je vyjádřena apriorním rozdělením na množině možných hodnot parametrů (v uvedeném příkladu na $\mathbf{R} \times \mathbf{R}^+$).

Při "neparametrickém" přístupu se v úvahách o Q neomezujeme, třídou možných rozdělení pozorování mohou být potenciálně všechna rozdělení na $(\mathcal{X}, \mathcal{A})$, neznámým parametrem je sama pravděpodobnostní míra Q. Pravděpodobnosti $Q(A)$ jednotlivých množin z $\mathcal{A}$ považujeme za náhodné veličiny

a celou náhodnou pravděpodobnost $Q = (Q(A), A \in \mathcal{A})$ můžeme chápat jako náhodný proces indexovaný prvky z $\mathcal{A}$. Apriorní informace o něm je popsána apriorním rozdělením na množině pravděpodobnostních měr, resp. jejich charakteristik, např. distribučních funkcí v případě pozorování s reálnými hodnotami.

Asi nejznámějším neparametrickým apriorním rozdělením je rozdělení *Dirichletova procesu* [2].

Definice 2.1. *Řekneme, že proces Q je Dirichletův s parametrem $\alpha = n_0 Q_0$, kde $n_0 \in \mathbf{R}^+$ a Q_0 je pravděpodobnostní míra na $(\mathcal{X}, \mathcal{A})$, pokud pro libovolný rozklad $A_1, \dots, A_k \in \mathcal{A}$, $\bigcup A_i = \mathcal{X}$ disjunktně, je*

$$(Q(A_1), \dots, Q(A_k)) \sim \mathrm{D}(\alpha(A_1), \dots, \alpha(A_k)),$$

kde D *značí Dirichletovo rozdělení. Značíme $Q \sim \mathcal{D}(\alpha)$.*

Je $\mathrm{E}\, Q(A_i) = Q_0(A_i)$ a $\operatorname{var} Q(A_i) = Q_0(A_i)(1 - Q_0(A_i))/(n_0 + 1)$, rozdělení procesu je tedy soustředěno kolem pravděpodobnostní míry Q_0, zatímco n_0 udává stupeň koncentrace. Dirichletův proces pokrývá ve smyslu nosiče širokou třídu rozdělení na $(\mathcal{X}, \mathcal{A})$, na druhou stranu Q je s pravděpodobností 1 diskrétním rozdělením. Dirichletovým procesem je např. $Q(A) = \sum p_n \delta_{Y_n}(A)$ (δ_x značí Diracovu míru soustředěnou v bodě x), kde body $Y_1, Y_2, \dots \in \mathcal{X}$, na nichž je hodnota Q soustředěna, se generují jako náhodný výběr z rozdělení Q_0, a to nezávisle na příslušných pravděpodobnostech $p_n = \theta_n \prod_{j<n}(1 - \theta_j)$, $n \in \mathbf{N}$, kde $\theta_1, \theta_2, \dots$ jsou nezávislé s beta rozdělením $\mathrm{B}(1, n_0)$ [9]. [2] ukazuje jinou volbu bodů a skoků.

Mnohem širší třídou apriorních rozdělení pro rozdělení na $\mathcal{X} = \mathbf{R}$ jsou procesy *neutrální zprava* [1]. O zprava neutrálním procesu hovoříme, když normalizované přírůstky distribuční funkce $F(t) = Q(-\infty, t\rangle$

$$F(t_1), \quad \frac{F(t_2) - F(t_1)}{1 - F(t_1)}, \quad \dots \quad , \frac{F(t_n) - F(t_{n-1})}{1 - F(t_{n-1})}$$

jsou nezávislé pro libovolná $t_1 < t_2 < \dots < t_n$. Totéž můžeme vyjádřit také pomocí příslušné "kumulativní intenzity" $\Lambda(t) = -\ln(1 - F(t))$.

Definice 2.2. *Řekneme, že Q, resp. Λ je zprava neutrální proces, jestliže Λ je neklesající, zprava spojitý proces s nezávislými přírůstky a $\Lambda(-\infty) = 0$, $\Lambda(\infty) = \infty$.*

Nemá-li proces Λ nenáhodnou složku, s pravděpodobností 1 intenzita Λ opět přísluší diskrétnímu rozdělení. Tak jak intenzita Λ generuje na $(\mathbf{R}, \mathcal{B}(\mathbf{R}))$ míru, značíme stručně např. $\Lambda(s, t) = \Lambda(t-) - \Lambda(s)$, $\Lambda\{t\} = \Lambda(t) - \Lambda(t-)$, apod. Speciálním případem zprava neutrálního procesu je gama proces.

Definice 2.3. *Mají-li přírůstky zprava neutrálního procesu Λ gama rozdělení, $\Lambda(s, t) \sim \mathrm{G}(n_0, n_0 \Lambda_0(s, t))$, kde $n_0 > 0$ a Λ_0 je nějaká kumulativní intenzita, nazveme ho gama procesem a značíme $\Lambda \sim \mathcal{G}(n_0, \Lambda_0)$.*

Parametry n_0 a Λ_0 mají podobný význam jako parametry Dirichletova procesu, je $\mathrm{E}\,\Lambda(s,t) = \Lambda_0(s,t)$, $\mathrm{var}\,\Lambda(s,t) = \Lambda_0(s,t)/n_0$. Rozdělením $\mathrm{G}(n_0, 0)$ rozumíme rozdělení degenerované v 0. Podobně jako Dirichletův proces, i gama proces může být definován na obecném prostoru, nejen na $(\mathbf{R}, \mathcal{B}(\mathbf{R}))$.

Podívejme se nyní, jak vypadají neparametrické bayesovské odhady odpovídající uvedeným apriorním procesům. Jako pozorovaná data uvažujme jednak úplný náhodný výběr $X_1, \ldots, X_n$ z rozdělení Q a jednak výběr s cenzorováním. Připouštíme tedy, že i-tý čas X_i může být zprava cenzorován veličinou Y_i na něm nezávislou. Skutečně pozorované hodnoty pak tvoří náhodný výběr dvojic

$$\text{data} = (Z_1, \delta_1), \ldots, (Z_n, \delta_n), \quad \text{kde } Z_i = \min(X_i, Y_i) \text{ a } \delta_i = I_{[X_i \leqq Y_i]} \tag{1}$$

jsou pozorovaný čas (skutečné pozorování X_i nebo čas cenzorování) a indikátor cenzorování. Značme jako $N_t = \#\{i, X_i > t\}$, resp. $N_t = \#\{i, Z_i > t\}$, počet pozorování překračujících t.

Je-li apriorním rozdělením Q Dirichletův proces $\mathcal{D}(\alpha)$, $\alpha = n_0 Q_0$, pak aposteriorní rozdělením $(Q \mid X_1, \ldots, X_n)$ je opět Dirichletův proces, $\mathcal{D}(\alpha + \sum \delta_{X_i})$. V případě reálných pozorování tak např. při kvadratické ztrátové funkci máme jako bayesovské odhady distribuční funkce $F(x) = Q(-\infty, x\rangle$, resp. střední hodnoty rozdělení Q, příslušné aposteriorní střední hodnoty

$$\widehat{F}(x) = \frac{(\alpha + \sum \delta_{X_i})(-\infty, x\rangle}{n_0 + n} = \frac{n_0 F_0(x) + n F_n(x)}{n_0 + n}, \quad \widehat{\mathrm{E}\,Q} = \frac{n_0\,\mathrm{E}\,Q_0 + n\bar{X}}{n_0 + n},$$

kde F_0 je distribuční funkce příslušná Q_0 a F_n je empirická distribuční funkce pozorování. V případě cenzorovaných pozorování (1) už konjugovanost systému Dirichletových měr neplatí, aposteriorní rozdělení spadá mezi beta-Stacyovy procesy [12]. Odhad funkce spolehlivosti $S = 1 - F$ je ale dobře znám a má tvar [10]

$$\widehat{S}(t) = \frac{n_0 S_0(t) + N_t}{n_0 + n} \prod_{x \in M_t^0} \frac{n_0 S_0(x-) + N(x-) - u(x)}{n_0 S_0(x-) + N_x}$$

kde $S_0(t) = 1 - F_0(t)$, $u(t)$ je počet necenzorovaných pozorování v okamžiku t a M_t^0 jsou okamžiky s (alespoň jedním) cenzorovaným pozorováním, $M_t^j = \{x \leqq t, \exists_i\, Z_i = x, \delta_i = j\}$, $j = 0, 1$. Dokonce rozdělení Y_i se pro různá i mohou lišit a může jít také o degenerované náhodné veličiny, cenzorování časem. Při zmenšující se míře počáteční informace $n_0 \to 0$ je $\widehat{F} \to F_n$ a $\widehat{\mathrm{E}\,Q} \to \bar{X}$ a podobně v případě cenzorování dostaneme v limitě neparametrický Kaplanův-Meierův odhad, $\widehat{S}(t) \to \prod_{x \in M_t^1} (1 - u(x)/N(x-))$.

Podobná "konjugovanost" platí i pro zprava neutrální procesy, ve třídě zprava neutrálních procesů (při daných hodnotách cenzorů) zůstaneme i v případě cenzorování, viz [3]. Zajímavou vlastností je, že i když proces kumulativní intenzity Λ nemá za apriorního rozdělení skoky v pevných bodech, v aposteriorním rozdělení se takové skoky v okamžicích pozorování vždy objeví. Ve třídě zprava neutrálních procesů zůstaneme dokonce i při jiném cenzorování, např. $X_i \geqq Y_i$. Zápis odhadů je v obecné šíři komplikovaný, uveďme

aspoň speciální případ gama procesu. Je-li apriorně $\Lambda \sim \mathcal{G}(n_0, \Lambda_0)$, kde Λ_0 je spojitá, a jsou-li pozorování navzájem různá, pak

$$\widehat{F}(t) = 1 - \Big(\frac{n_0 + N_t}{n_0 + N_t + 1}\Big)^{n_0\Lambda_0(t)} \prod_{i, X_i \leqq t} \Big(\frac{n_i + 1}{n_i}\Big)^{n_0\Lambda_0(t)} \frac{\ln((n_i+2)/(n_i+1))}{\ln((n_i+1)/n_i)},$$

kde $n_i = n_0 + N_{X_i}$. Aposteriorní rozdělení Λ ale gama procesem není. V případě cenzorování (1) má (při Z_i navzájem různých a spojité Λ_0) odhad opět stejný tvar, součin nyní probíhá pouze přes necenzorovaná pozorování.

Existuje a používá se řada dalších apriorních rozdělení, např. směsi Dirichletových či gama procesů nebo naopak Dirichletův proces, jehož parametr je směsí rozdělení. Jiné apriorní procesy mohou vést, na rozdíl od popsaných, ke spojitým rozdělením, nebo přímo modelovat náhodné hustoty či (nekumulativní) intenzity poruch. V určitých situacích může být vhodnější zabývat se místo kumulativní intenzity Λ její variantou $\Lambda^*(x) = \int_{\langle 0,x\rangle}(dF(t)/S(t-))$, $x > 0$, viz [5].

3 Odhady v Koziolově-Greenově modelu

V Koziolově-Greenově modelu se u pozorování (1) předpokládá, že distribuční funkce F dob života X_i je s distribuční funkcí F_Y časů cenzorování Y_i svázána podmínkou $1-F_Y = (1-F)^\gamma$ s nějakou konstantou $\gamma > 0$, tzn. že kumulativní intenzity jsou si úměrné, $\Lambda_Y = \gamma\Lambda$. Rozdělení X_i a Y_i tak mají společný parametr a i v pozorování samotné cenzorující veličiny ($Y_i = t$ nebo $Y_i \geqq t$) je obsažena informace o kumulativní intenzitě Λ. S tím uvedené příklady odhadů nepočítají. Podobný problém řeší v modelu konkurujících si rizik [8] zavedením "vícerozměrného" Dirichletova procesu.

Budeme předpokládat, že "parametry" Λ a γ jsou při apriorním rozdělení nezávislé, Λ je gama proces $\mathcal{G}(n_0, \Lambda_0)$ a rozdělení γ má hustotu $\pi(\gamma)$, $\gamma > 0$. Pro jednoduchost zápisu budeme dále předpokládat, že Λ_0 je kumulativní intenzita spojitého rozdělení, že pozorování Z_i jsou navzájem různá a již uspořádaná vzestupně. Označme ještě dále jako $\mathrm{G}_0(a, b)$, $a > 0, b > 0$ rozdělení na $\mathbf{R}^+$ s hustotou a momentovou vytvořující funkcí

$$f(x) = \frac{1}{x}\frac{\mathrm{e}^{-ax} - \mathrm{e}^{-bx}}{\ln(a/b)},\ x > 0, \quad M(\theta) = \frac{\ln((a-\theta)/(b-\theta))}{\ln(a/b)},\ \theta < \min(a, b),$$

které vyplyne jako aposteriorní rozdělení skoků u Λ, a nakonec

$$C_j(\gamma) = \Big(\frac{n_0}{n_0 + N_{Z_{j-1}}(1+\gamma)}\Big)^{n_0\Lambda_0(Z_{j-1}, Z_j)}, \tag{2}$$

$$D_j(\gamma) = \begin{cases} -\ln \frac{n_0 + N_{Z_j}(1+\gamma)}{n_0 + N_{Z_j}(1+\gamma)+1}, & \delta_j = 1 \\ -\ln \frac{n_0 + N_{Z_j}(1+\gamma)+1}{n_0 + N_{Z_j}(1+\gamma)+1+\gamma}, & \delta_j = 0, \end{cases}$$

kde N_t stále značí počet pozorování Z_j překračujících t, za našich speciálních předpokladů tedy $N_{Z_j} = n - j$.

Nyní můžeme zformulovat tvrzení o aposteriorním rozdělení parametrů. Přestože při apriorním rozdělení proces Λ nemá skoky v pevně daných bodech, v aposteriorním rozdělení takové skoky jsou, a to v bodech pozorování. Skutečnost, že rozdělení těchto skoků nezávisí (v našem případě) na konkrétních časech pozorování je dána homogenitou gama procesu jakožto Lévyova procesu (viz [3]).

Tvrzení 3.1. *Je-li Λ apriorně rozdělena jako gama proces $\mathcal{G}(n_0, n_0\Lambda_0)$, kde Λ_0 je absolutně spojitá funkce, a γ má apriorní hustotu $\pi(\gamma)$, $\gamma > 0$, a je nezávislé s Λ, pak při daných pozorováních* (1) *s uspořádáním $Z_1 < \cdots < Z_n$ pro aposteriorní rozdělení platí*

- $(\Lambda \mid \text{data}, \gamma)$ *je rozdělení procesu s nezávislými přírůstky, přičemž pro $(s,t) \subset (Z_i, Z_{i+1})$, $i = 0, \ldots, n$ (volíme $Z_0 = 0$, $Z_{n+1} = \infty$) je*

$$(\Lambda(s,t) \mid \text{data}, \gamma) \sim \mathrm{G}(n_0 + N_s(1+\gamma), n_0\Lambda_0(s,t))$$

  a pro $t = Z_i$ má Λ v t skok s rozdělením

$$(\Lambda\{Z_i\} \mid \text{data}, \gamma) \sim \begin{cases} \mathrm{G}_0(N_{Z_i}(1+\gamma), N_{Z_i}(1+\gamma)+1), & \delta_i = 1, \\ \mathrm{G}_0(N_{Z_i}(1+\gamma)+1, N_{Z_i}(1+\gamma)+1+\gamma), & \delta_i = 0, \end{cases}$$

- $(\gamma \mid \text{data})$ *má rozdělení s hustotou*

$$\pi(\gamma \mid \text{data}) \propto \Big(\prod_{j=1}^{n} C_j(\gamma)\, D_j(\gamma)\Big)\pi(\gamma), \quad \gamma > 0.$$

Důkaz. Aposteriorní rozdělení získáme postupnými aktualizacemi po jednotlivých pozorováních (Z, δ). Každé takové pozorování zachycuje buď dvojici $X = t, Y \geqq t$, když $Z = t, \delta = 1$, nebo $X > t, Y = t$, když $Z = t, \delta = 0$. Naznačíme důkaz aktualizace po prvním pozorování pro případ s $\delta = 0$, tj. výpočet aposteriorního rozdělení po pozorování dvojice $X > t, Y = t$.

Zvolme libovolně dělení $0 = t_0 < t_1 < \cdots < t_k < \infty$ s dělícími body různými od t a nechť i je ten index, pro nějž $t \in (t_{i-1}, t_i)$. Postupujeme ve dvou krocích. Nejprve určíme aposteriorní rozdělení veličin γ a $\lambda = (\lambda_1, \ldots, \lambda_k)$, kde $\lambda_j = \Lambda(t_{j-1}, t_j)$, $j = 1, \ldots, k$, při daném $Y = t$. Jejich momentová vytvořující funkce má v bodě $(\theta_0, \theta_1, \ldots, \theta_k)$ hodnotu

$$M(t) = \mathrm{E}(U \mid Y = t) = \int_0^\infty \mathrm{e}^{\gamma\theta_0} M_{(\lambda|\gamma, Y=t)}(\theta_1, \ldots, \theta_k)\pi(\gamma \mid Y = t)\,\mathrm{d}\gamma, \quad (3)$$

$U = \mathrm{e}^{\gamma\theta_0}\mathrm{e}^{\sum \lambda_j\theta_j}$. Spočteme $I(x) = \int_x^\infty M(t) f_Y(t)\,\mathrm{d}t$, $x \in (t_{i-1}, t_i)$, kde f_Y značí nepodmíněnou hustotu Y, a odtud pak $M(t) = (-1/f_Y(t))(\,\mathrm{d}I(t)/\,\mathrm{d}t)$.

Z definice podmíněné střední hodnoty máme

$$\begin{aligned} I(x) &= \int_{Y>x} \mathrm{E}(U \mid Y)\,\mathrm{d}\,\mathrm{P} = \int_{Y>x} U\ \,\mathrm{d}\,\mathrm{P} = \mathrm{E}(U \cdot \mathrm{P}(Y > x \mid \Lambda, \gamma)) \\ &= \mathrm{E}(U\mathrm{e}^{-\gamma\Lambda(x)}) = \mathrm{E}\Big(\mathrm{e}^{\gamma\theta_0} \prod_{j<i} \mathrm{e}^{-\lambda_j(\gamma-\theta_j)} \cdot \mathrm{e}^{-\lambda_{i1}(\gamma-\theta_i)}\mathrm{e}^{\lambda_{i2}\theta_i} \cdot \prod_{j>i} \mathrm{e}^{\lambda_j\theta_j}\Big) \\ &= \int_0^\infty \mathrm{e}^{\gamma\theta_0} \prod_{j<i} \Big(\frac{n_0}{n_0+\gamma-\theta_j}\Big)^{n_0\lambda_j^0} \cdot \Big(\frac{n_0}{n_0+\gamma-\theta_i}\Big)^{n_0\lambda_{i1}^0(x)} \\ &\qquad \cdot \Big(\frac{n_0}{n_0-\theta_i}\Big)^{n_0\lambda_{i2}^0(x)} \prod_{j>i} \Big(\frac{n_0}{n_0-\theta_j}\Big)^{n_0\lambda_j^0} \pi(\gamma)\,\mathrm{d}\gamma \end{aligned}$$

při značení $\lambda_j^0 = \mathrm{E}\,\lambda_j = \Lambda_0(t_{j-1}, t_j)$ a podobně $\lambda_{i1}^0(x) = \Lambda_0(t_{i-1}, x)$, $\lambda_{i2}^0(x) = \Lambda_0(x, t_i)$. Ve výsledném výrazu pro $M(t)$ (po zmíněném zderivování) si při srovnání s (3) přečteme hustotu $\pi(\gamma \mid Y = t)$ a vytvořující funkci $M_{(\lambda|\gamma,Y=t)}$. Ta se rozpadá na součin vytvořujících funkcí veličin λ_j, veličiny $\lambda_1, \ldots, \lambda_k$ jsou tedy při daném γ (a $Y = t$) nezávislé. Kromě λ_i jde o vytvořující funkce gama rozdělení, veličina $\lambda_i = \Lambda(t_{i-1}, t_i)$ má vytvořující funkci jako součet nezávislých veličin s rozděleními $\mathrm{G}(n_0 + \gamma, n_0\Lambda_0(t_{i-1}, t))$, $\mathrm{G}_0(n_0, n_0 + \gamma)$ a $\mathrm{G}(n_0, n_0\Lambda_0(t, t_i))$. Ve skutečnosti popisuje tato trojice rozdělení veličin $\Lambda(t_{i-1}, t)$, $\Lambda\{t\}$ a $\Lambda(t, t_i)$.

Nakonec přidáme ještě pozorování $X > t$, aktualizovaná aposteriorní hustota např. pro $\lambda = (\lambda_j, j \neq i, \lambda_{i1}, \lambda_{\{t\}} = \Lambda(\{t\}), \lambda_{i2})$ a γ bude

$$\begin{aligned} \pi(\lambda, \gamma \mid X > t, Y = t) &\propto \mathrm{P}(X > t \mid \lambda, \gamma, Y = t)\pi(\lambda, \gamma \mid Y = t) \\ &= \mathrm{P}(X > t \mid \lambda, \gamma)\pi(\lambda, \gamma \mid Y = t) = \mathrm{e}^{-(\sum_{j<i}\lambda_j) - \lambda_{i1} - \lambda_{\{t\}}}\pi(\lambda, \gamma \mid Y = t). \end{aligned}$$

V případě s $\delta = 1$ bychom stejným postupem nejprve určili rozdělení parametrů při $X = t$ a následně přidali $Y \geqq t$. Podobně bychom aktualizovali aposteriorní rozdělení při dalších pozorováních, do dělení bychom zařadili také všechny časy předchozích pozorování. □

V důkazu jsme využili, jak radí [6], konkrétního tvaru apriorního rozdělení. Obecnější důkazovou techniku, založenou na Lévyově míře neklesajícího procesu Λ, nabízí [1]. Ke správné aposteriorní hustotě ale vede i jednoduchý intuitivní přístup — zkombinovat apriorní hustotu parametrů s věrohodnostní funkcí danou pozorováními. Do apriorní hustoty ovšem musíme pro (apriorně) nulové veličiny $\lambda_{\{j\}} = \Lambda\{t\}$, $t = Z_j$, formálně zapsat "hustotu" $\lambda_{\{j\}}^{-1}\mathrm{e}^{-n_0\lambda_{\{j\}}}$, $\lambda_{\{j\}} > 0$, jakožto hustotu "gama" rozdělení $\mathrm{G}(n_0, 0)$. Máme tedy

$$\pi(\lambda) \propto \prod \lambda_j^{\lambda_j^0 - 1}\mathrm{e}^{-n_0\lambda_j} \cdot \lambda_{\{j\}}^{-1}\mathrm{e}^{-n_0\lambda_{\{j\}}}$$

a do věrohodnostní funkce s každým pozorováním přidáváme člen typu

$$\mathrm{e}^{-\sum_{j\leqq i}\lambda_j(1+\gamma)}\mathrm{e}^{-\sum_{j<i}\lambda_{\{j\}}(1+\gamma)}p_i$$

pro pozorování v čase t_i, kde $p_i = 1 - \mathrm{e}^{-\lambda_{\{i\}}}$ u pozorování necenzorovaného a $p_i = \mathrm{e}^{-\lambda_{\{i\}}}(1 - \mathrm{e}^{-\lambda_{\{i\}}\gamma})$ u cenzorovaného. Při daném γ rozeznáme pak hustotu λ a snadno zjistíme i normovací konstanty $C_j(\gamma)$ a $D_j(\gamma)$.

Uveďme konečně tvar bayesovského odhadu funkce spolehlivosti pro dobu života při kvadratické ztrátové funkci, tj. aposteriorní střední hodnotu veličiny $\exp(-\Lambda(t))$.

Tvrzení 3.2. *Za předpokladů předchozího tvrzení máme pro funkci spolehlivosti S bayesovský odhad*

$$\widehat{S}(t) = \frac{\displaystyle\int \Big(\prod_{j<i} C_j^+(\gamma) D_j^+(\gamma)\Big) C_{i1}^+(\gamma) C_{i2}(\gamma) D_i(\gamma) \Big(\prod_{j>i} C_j(\gamma) D_j(\gamma)\Big) \pi(\gamma)\,\mathrm{d}\gamma}{\displaystyle\int \Big(\prod_{j\neq i} C_j(\gamma)\Big) C_{i1}(\gamma) C_{i2}(\gamma) \Big(\prod_{j} D_j(\gamma)\Big) \pi(\gamma)\,\mathrm{d}\gamma}$$

$t \in (Z_{i-1}, Z_i)$, kde $C_{i1}(\gamma)$, resp. $C_{i2}(\gamma)$ jsou jako $C_i(\gamma)$ v (2)*, ale s exponenty $n_0\Lambda_0(Z_{i-1}, t)$, resp. $n_0\Lambda_0(t, Z_i)$ a C^+, resp. D^+ jsou jako C a D, ale s $N_t(1+\gamma)+1$ místo $N_t(1+\gamma)$.*

Důkaz. Použijeme značení jako v důkazu předchozího tvrzení, ale s dělením $0 = Z_0 < Z_1 < \cdots < Z_n$ a s $t \in (Z_{i-1}, Z_i)$. Počítáme

$$\widehat{S}(t) = \mathrm{E}\,\mathrm{e}^{-\Lambda(t)} = \mathrm{E}\Big[\Big(\prod_{j<i} \mathrm{E}(\mathrm{e}^{-\lambda_j} \mid \gamma)\,\mathrm{E}(\mathrm{e}^{-\lambda_{\{j\}}} \mid \gamma)\Big)\,\mathrm{E}(\mathrm{e}^{-\lambda_{i1}} \mid \gamma)\Big],$$

kde všechny střední hodnoty rozumíme navíc jako podmíněné pozorováními (1), tj. při aposteriorním rozdělení. Ve výrazu uvedené střední hodnoty $\mathrm{E}(.|\gamma)$ jsou postupně $C_j^+(\gamma)/C_j(\gamma)$, $D_j^+(\gamma)/\,\mathrm{D}_j(\gamma)$ a $C_{i1}^+(\gamma)/C_{i1}(\gamma)$. □

Omezující předpoklady jsme zavedli jen z důvodu přehlednosti, v jejich uvolnění nám nic nebrání. Pokud např. připustíme skoky v "apriorním odhadu" Λ_0, pak v případě shody některého pozorování s bodem skoku se jako aposteriorní rozdělení skoku v takovém bodě místo rozdělení G_0 objeví rozdělení s hustotou úměrnou rozdílu gama hustot a místo "normovací konstanty" $D(\gamma)$ rozdíl "konstant" $C(\gamma)$. Podobně, lze dospět k aposteriorním rozdělením i v případě shod časů některých pozorování. V případě dvou cenzorovaných pozorování ve stejném čase $t = t_i$ to znamená v příslušném dělícím bodě člen $\mathrm{e}^{-2\lambda_{\{i\}}}(1 - \mathrm{e}^{-\lambda_{\{i\}}\gamma})^2$, když byla obě cenzorovaná, nebo $\mathrm{e}^{-\lambda_{\{i\}}}(1 - \mathrm{e}^{-\lambda_{\{i\}}\gamma})(1 - \mathrm{e}^{-\lambda_{\{i\}}})$, když jedno bylo cenzorované a jedno nikoliv. Tomu pak odpovídají i normovací konstanty, místo $D_i(\gamma)$ dostaneme $\ln \frac{(n_0+2)(n_0+2+2\gamma)}{(n_0+2+\gamma)^2}$, resp. $\ln \frac{(n_0+1)(n_0+2+\gamma)}{(n_0+1+\gamma)(n_0+2)}$. Podobně můžeme uvážit i jiné typy cenzorovaných pozorování jako jsou $X = Y = t$ nebo $X \geqq Y, Y = t$.

Podobný postup jako pro gama proces je možné uplatnit také pro jiná apriorní rozdělení. Např. v případě apriorního Dirichletova procesu $1-\mathrm{e}^{-\Lambda} \sim \mathcal{D}(\alpha)$ v normovacích konstantách vyjdou rozdíly digama funkcí.

Reference

[1] Doksum K. (1974). *Tailfree and neutral random probabilities and their posterior distributions*. Ann. Probability **2**, No. 2, 183 – 201.

[2] Ferguson T.S. (1973). *A Bayesian analysis of some nonparametric problems*. Ann. Statist. **1**, No. 2, 209 – 230.

[3] Ferguson T.S., Phadia E.G. (1979). *Bayesian nonparametric estimation based on censored data*. Ann. Statist. **7**, No. 1, 163 – 186.

[4] Ferguson T.S., Phadia E.G., Tiwari R.C. (1992). *Bayesian nonparametric inference*. In Current issues in statistical inference: essays in honor of D. Basu (Ghosh M., Pathak P.K., eds.), IMS Lecture Notes Monogr. Ser. **17**, Inst. Math. Statist., Hayward, 127 – 150.

[5] Hjort, N.L. (1990). *Nonparametric Bayes estimators based on beta processes in models for life history data*, Ann. Statist. **18**, No. 3, 1259 – 1294.

[6] Kalbfleisch, J.D. (1978). *Non-parametric Bayesian analysis of survival time data*, J. Roy. Statist. Soc. Ser. B **40**, No. 2, 214 – 221.

[7] Koziol J.A., Green S.B. (1976). *A Cramér-von Mises statistic for randomly censored data*. Biometrika **63**, No. 3, 465 – 474.

[8] Salinas-Torres V.H., Pereira C.A.B., Tiwari R.C. (2002). *Bayesian nonparametric estimation in a series system or a competing-risks model*. J. Nonparametr. Stat. **14**, No. 4, 449 – 458.

[9] Sethuraman J. (1994). *A constructive definition of Dirichlet priors*. Statist. Sinica **4**, No. 2, 639 – 650.

[10] Susarla V., Van Ryzin J. (1976). *Nonparametric Bayesian estimation of survival curves from incomplete observations*, J. Amer. Statist. Assoc. **71**, No. 356, 897 – 902.

[11] Walker S.G., Damien P., Laud P.W., Smith A.F.M. (1999). *Bayesian nonparametric inference for random distributions and related functions*. With discussion and a reply by the authors, J. R. Stat. Soc. Ser. B Stat. Methodol. **61**, No. 3, 485 – 527.

[12] Walker S., Muliere P. (1997). *Beta-Stacy processes and a generalization of the Pólya-urn scheme*. Ann. Statist. **25**, No. 4, 1762 – 1780.

Poděkování: Tato práce vznikla za podpory výzkumného záměru MSM 235200001.

Adresa: Katedra matematiky, Fakulta aplikovaných věd, Západočeská univerzita v Plzni, Univerzitní 22, 306 14 Plzeň

E-mail: friesl@kma.zcu.cz

ODHAD RIZIKOVĚ NEUTRÁLNÍ HUSTOTY ZALOŽENÝ NA CENÁCH EVROPSKÝCH OPCÍ

Zdeněk Hlávka

Klíčová slova: Oceňování opcí, rizikově neutrální hustota, metoda nelineárních nejmenších čtverců.

Abstrakt: Cílem tohoto článku je navrhnout jednoduchý odhad rizikově neutrální hustoty založený na pozorovaných cenách evropských opcí. Navržená metoda je použita na ceny kupních opcí na index německého akciového trhu v lednu 1995.

1 Úvod

Vlastník kupní opce evropského typu získá v čase T částku $(S_T - K)_+ = \max(S_T - K, 0)$, kde S_T označuje cenu příslušné akcie (v čase T) a K realizační cenu předem dohodnutou v čase $t < T$. V čase t lze cenu $C_t(K,T)$ takové opce, tj. práva koupit danou akcii v čase T za cenu K, zapsat jako střední hodnotu zisku násobenou diskontním faktorem zohledňujícím bezrizikovou úrokovou míru r:

$$C_t(K,T) = \exp\{-r(T-t)\} \int_0^{+\infty} (S_T - K)_+ f(S_T) dS_T, \tag{1}$$

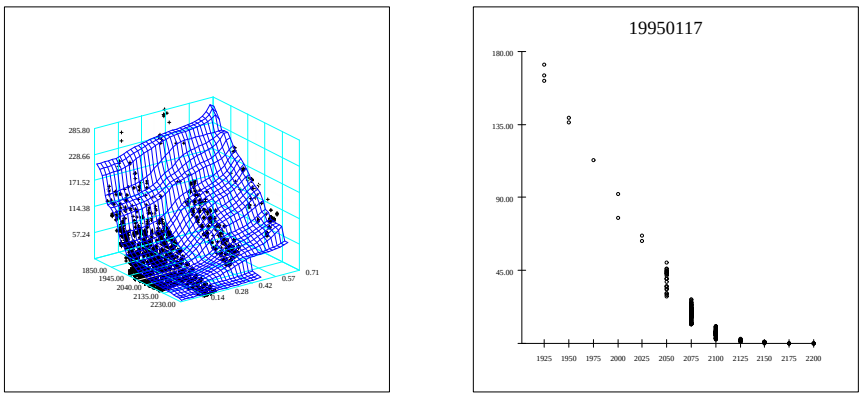
kde $f(\bullet)$ je hustota náhodné veličiny S_T. Hustota $f(\bullet)$ odpovídající pozorovaným cenám opcí se nazývá rizikově neutrální hustota (RNH).

Zajímavá je závislost ceny opcí na realizační ceně. Jednoduše se dá ukázat, že cena evropské kupní opce jako funkce realizační ceny musí být kladná, klesající a konvexní. Ceny kupních opcí na index německého akciového trhu DAX v lednu 1995 jsou graficky znázorněny na obrázku 1.

Cena $C_t(K,T)$ kupní opce evropského typu, jako funkce realizační ceny K v pevném čase t a s pevnou dobou splatnosti T, dovoluje vyjádřit RNH $f(\bullet)$ v následujícím tvaru [5]:

$$f(K) = \exp\{r(T-t)\} \frac{\partial^2 C_t(K,T)}{\partial K^2}. \tag{2}$$

Vyjádření (2) se často s úspěchem používá k odhadům RNH pomocí neparametrických jádrových odhadů [1], [2], [3]. Jiný zajímavý přístup založený na metodě neparametrických nejmenších čtverců a umožňující splnění všech požadovaných podmínek je navržen v článku [8]. Další parametrické i neparametrické metody jsou shrnuty v článku [6].



Obrázek 1: Ceny kupních opcí v závislosti na realizační ceně a na době splatnosti v lednu 1995 (levý obrázek) a ceny opcí v závislosti na realizační ceně pro nejkratší dobu splatnosti 17. ledna 1995 (19950117).

Důležitou aplikací odhadu RNH je studium rozdílů mezi RNH a skutečnou hustotou veličiny S_T, tj. studium vztahu investorů k riziku, např. [2]. Předpokládáme-li totiž rizikově neutrální chování investorů, pak RNH přesně odpovídá hustotě náhodné veličiny S_T. V praxi se ukazuje, že tento předpoklad bývá porušen zvláště pro opce, které umožňují získat malý zisk s velice malým rizikem nebo, na druhé straně, velký zisk s velkým rizikem.

Cílem tohoto článku je konstrukce odhadu RNH, který bude možné jednoduše zobecnit i pro heteroskedastická a závislá data. Dalším cílem je jednoduchá konstrukce konfidenčních intervalů založená na asymptotické normalitě získaných odhadů. Navrženou metodu použijeme na ceny opcí na index německého akciového trhu v lednu 1995.

2 Označení a jednoduchý lineární model

Označme i-tou pozorovanou cenu $C_i = C_{t,i}(K_i, T)$. Symbol K_i bude označovat realizační cenu odpovídající i-tému pozorování. V našich datech se vyskytuje pouze malé množství rozdílných realizačních cen, zatímco počet pozorování je mnohonásobně větší. Proto je užitečné zavést následující označení. Nechť $\mathcal{C} = (C_1, \ldots, C_n)^\top$ označuje vektor pozorovaných cen opcí. Předpokládejme, že vektor odpovídajících realizačních cen má následující strukturu:

$$\mathcal{K} = \begin{pmatrix} K_1 \\ K_2 \\ \vdots \\ K_n \end{pmatrix} = \begin{pmatrix} k_1 \mathbf{1}_{n_1} \\ k_2 \mathbf{1}_{n_2} \\ \vdots \\ k_p \mathbf{1}_{n_p} \end{pmatrix},$$

kde $k_1 < k_2 < \cdots < k_p$, $n_j = \sum_{i=1}^{n} I(K_i = k_j)$ a $\mathbf{1}_n$ označuje sloupcový vektor jedniček délky n.

Pro pevný čas t a pevnou dobu do splatnosti $\tau = T - t$, nyní vyjádříme i-tou pozorovanou cenu opce, odpovídající realizační ceně K_i, jako

$$C_{t,i}(K_i, T) = \mu(K_i) + \varepsilon_i, \tag{3}$$

kde ε_i jsou nezávislé stejně rozdělené náhodné veličiny s rozdělením $N(0, \sigma^2)$.

2.1 Existence a jednoznačnost

V našich datech i ve skutečném životě pozorujeme místo spojitých funkcí pouze několik funkčních hodnot. Proto formulujeme podmínky kladené na cenu opcí pouze pro diskrétní případ, kdy jsou ceny opcí $C(K_i)$, $i = 1, \ldots, n$ pozorované pouze pro několik různých realizačních cen $k_1 < \cdots < k_p$, $p \ll n$.

Ceny opcí $C(\bullet)$ splňují následující podmínky:

1. $C(k_i) \geq 0$, $i = 1, \ldots, p$,
2. $k_i < k_j$ implikuje $C(k_i) \geq C(k_j)$,
3. $k_i < k_j < k_l$ implikuje $-1 \leq C^{(1)}_{k_i,k_j} \leq C^{(1)}_{k_j,k_l} \leq 0$,

kde $C^{(1)}_{k_i,k_j} = \{C(k_i) - C(k_j)\}/\{k_i - k_j\}$ označuje přirozený odhad první derivace funkce $C(\bullet)$.

Regresní funkce $\hat{C}$ je jednoznačně určena svými funkčními hodnotami v bodech $k_1, \ldots, k_p$ a její druhé diference aproximují RNH. Podobně jako v [7] budeme považovat soubor funkcí $\mathcal{C}$, splňujících podmínky 1–3, za podmnožinu p-rozměrného Euklidovského prostoru. Regrese $\hat{C}$ splňující podmínky 1–3 je nejbližší bod množiny $\mathcal{C}$ k pozorovanému C.

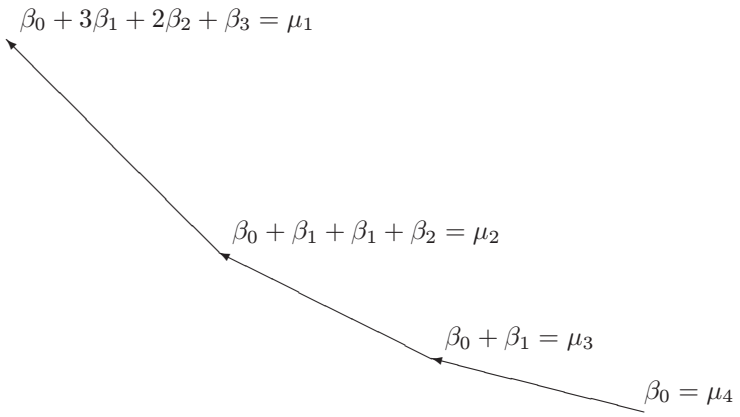
Množina $\mathcal{C}$ funkcí splňujících podmínky 1–3 má následující vlastnosti:

- $\mathcal{C}$ je uzavřená v topologii indukované Euklidovskou metrikou se vzdálenostmi $d(f, g) = \sum_{i=1}^{n} \{f(K_i) - g(K_i)\}^2$,
- $\mathcal{C}$ je konvexní, t.j., pokud $f, g \in \mathcal{C}$ a $0 \leq a \leq 1$, pak $af + (1 - a)g \in \mathcal{C}$.

Lemma 1 *Je-li $\hat{C}$ regrese funkce $C(\bullet)$ na $K_1, \ldots, K_n$ za podmínek 1–3 a jsou-li a a b konstanty takové, že $a \leq C(k_i) \leq b$, $i = 1, \ldots, p$, pak $a - (k_p - k_1) \leq \hat{C}(k_i) \leq b + (k_p - k_1)$.*

Důkaz: Není možné, aby $\hat{C}(k_i)$ bylo menší než a nebo větší než b pro všechna k_i, protože v takovém případě by ke zmenšení součtu čtverců stačilo posunout celou regresní funkci. Meze nyní implikuje podmínka 3. □

Věta 1 *Existuje regresní funkce $\hat{C} = \arg\min_{f \in \mathcal{C}} d(C, f)$, splňující podmínky 1–3.*



Obrázek 2: Grafické znázornění dummy proměnných pro ceny kupních opcí.

Důkaz: Lemma 1 říká, že $\hat{C}$ patří do množiny $\mathcal{C} \in \mathcal{M}$, ohraničené zespodu $a - (k_p - k_1)$ a zeshora $b + (k_p - k_1)$. Díváme-li se na funkce jako na body v Euklidovském prostoru, je jasné, že spojitá funkce $d(m, f)$ nabývá svého minima na uzavřené a omezené množině $\mathcal{C}$. □

Poznámka 1 *Předpokládejme, že $\mathcal{C}$ je konvexní množina funkcí a C je daná funkce. Je-li $\hat{C} = \arg\min_{f \in \mathcal{C}} d(f, C)$, pak pro každou $f \in \mathcal{C}$ platí*

$$\sum_{i=1}^{n} \{C(K_i) - \hat{C}(K_i)\}^\top \{\hat{C}(K_i) - f(K_i)\} \geq 0. \tag{4}$$

Existuje nejvýše jedna funkce $\hat{C}$ splňující (4).

Důkaz: Viz Věta 1.3.1 v [7]. □

Důsledek *Regresní funkce $\hat{C}$ splňující podmínky 1–3 existuje a je jednoznačně určená.*

Důkaz: Tvrzení plyne z věty 1 a z poznámky 1. □

2.2 Lineární model

Označme střední hodnotu ceny opce při dané realizační ceně k_j jako $\mu_j = EC(k_j)$. Nyní vyjádříme podmíněné střední hodnoty μ_j, $j = 1, \ldots, p$, pomocí koeficientů β_i, $i = 1, \ldots, (p-1)$ jako

$$\begin{aligned} \mu_p &= \beta_0, \\ \mu_{p-1} &= \beta_0 + \beta_1, \\ \mu_{p-2} &= \beta_0 + 2\beta_1 + \beta_2, \end{aligned}$$

$$\begin{aligned}\mu_{p-3} &= \beta_0 + 3\beta_1 + 2\beta_2 + \beta_3,\\ &\vdots\\ \mu_1 &= \beta_0 + (p-1)\beta_1 + (p-2)\beta_2 + \cdots + \beta_{p-1}.\end{aligned}$$

Na obrázku 2 je tento jednoduchý lineární model graficky znázorněn pro čtyři realizační ceny.

Z obrázku 2 je také dobře vidět interpretace regresních koeficientů β_j. Parametr β_0 je průměrná cena opce v bodě 4. Podle podmínky 1 musí být tento koeficient kladný. β_1 je rozdíl mezi cenami opcí v bodech 4 a 3 a podle podmínky 2 musí být tento koeficient také kladný. Další koeficient, β_2, můžeme popsat jako změnu první derivace v bodě 3 a tedy jako odhad druhé derivace v tomto bodě. Podobně, β_3 interpretujeme jako odhad druhé derivace funkce $C(K)$ v bodě 2. Podle podmínky 3 musí být β_2 i β_3 větší než nula. Podmínka 3 znamená také, že $\beta_1 + \beta_2 + \beta_3 \leq 1$.

Interpretace koeficientů nakreslených na obrázku 2 je zjednodušená díky předpokladu, že vzdálenost sousedních realizačních cen (bodů na horizontální ose) se rovná jedné. V praxi tento předpoklad nebývá splněn a pro zachování interpretace parametrů β_j musíme použít matici experimentu

$$\Delta = \begin{pmatrix} 1 & \Delta_p^1 & \Delta_{p-1}^1 & \Delta_{p-2}^1 & \cdots & \Delta_3^1 & \Delta_2^1 \\ 1 & \Delta_p^2 & \Delta_{p-1}^2 & \Delta_{p-2}^2 & \cdots & \Delta_3^2 & 0 \\ \vdots & & & & & & \vdots \\ 1 & \Delta_p^{p-2} & \Delta_{p-1}^{p-2} & 0 & \cdots & 0 & 0 \\ 1 & \Delta_p^{p-1} & 0 & 0 & \cdots & 0 & 0 \\ 1 & 0 & 0 & 0 & \cdots & 0 & 0 \end{pmatrix}, \tag{5}$$

kde $\Delta_j^i = \max(k_j - k_i, 0)$ označuje kladnou část rozdílu mezi k_i a k_j, t.j., i-tou a j-tou ($1 \leq i \leq j \leq p$) nejmenší pozorovanou hodnotou realizační ceny.

Vektor průměrných cen opcí μ lze zapsat pomocí parametrů β jako

$$\begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} = \mu = \Delta\beta = \Delta \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix}. \tag{6}$$

Podmínky 1–3 můžou být vyjádřeny pomocí parametrů lineárního modelu (6). Postačuje předpokládat, že $\beta_i > 0$, $i = 0, \ldots, p-1$ a $\sum_{j=2}^{p-1} \beta_j \leq 1$.

3 Nelineární metoda nejmenších čtverců

Kvůli snadnějšímu výpočtu je výhodná následující reparametrizace. Parametry β_i, $i = 0, \ldots, p-1$, nahradíme parametry ξ_j, $j = 0, \ldots, p$ následujícím způsobem

$$\beta_0(\xi) = \exp(\xi_0),$$

$$\begin{aligned}\beta_1(\xi) &= \frac{\exp(\xi_1)}{\sum_{j=1}^{p}\exp(\xi_j)},\\ &\vdots\\ \beta_{p-1}(\xi) &= \frac{\exp(\xi_{p-1})}{\sum_{j=1}^{p}\exp(\xi_j)}.\end{aligned}$$

Rovnost

$$\exp(\xi_p)\left\{\sum_{j=1}^{p-1}\exp(\xi_j)\right\}^{-1} = 1-\left\{\sum_{j=1}^{p-1}\beta_j(\xi)\right\}^{-1}$$

ukazuje význam parametru ξ_p. Pokud by byl tento parametr roven $-\infty$, pak by $\sum_{j=2}^{p-1}\beta_j(\xi) = 1$. Větší hodnoty tohoto parametru by znamenaly, že pozorované realizační ceny nepokrývají celý nosič hustoty RNH.

3.1 Inverzní transformace parametrů modelu

Při výpočtu je užitečné vědět, jak lze spočítat hodnoty parametrů ξ ze zadaných parametrů β.

Lemma 2 *Splňuje-li vektor* $\beta = (\beta_1,\ldots,\beta_p)^\top$ *podmínku* $\beta_p = 1-\sum_{i=1}^{p-1}\beta_i$, *pak odpovídající vektor* $\xi = (\xi_1,\ldots,\xi_p)^\top$ *splňuje systém rovnic*

$$\left(\beta\mathbf{1}_p^\top - \mathbf{I}_p\right)\exp\xi^\top = \mathcal{A}\exp\xi^\top = 0, \tag{7}$$

kde $\mathbf{1}_p$ *je p-rozměrný vektor jedniček a* $\mathbf{I}_p$ *je* $(p\times p)$ *jednotková matice. Dále platí, že* $\operatorname{rank}\mathcal{A} = p-1$. *Systém (7) má nekonečně mnoho řešení, která mohou být vyjádřena jako* $\exp(\xi) = (\mathcal{A}^-\mathcal{A}-\mathbf{I}_p)\,z$, *kde* $\mathcal{A}^-$ *značí zobecněnou inverzi matice* $\mathcal{A}$ *a kde z je libovolný vektor z* $\mathbf{R}^p$ *takový, že výraz na pravé straně je kladný.*

Důkaz: Vztah (7) a hodnost matice $\mathcal{A}$ lze odvodit z definice $\beta(\xi)$ použitím jednoduché algebry (řádkové součty matice $\mathcal{A}$ se rovnají nule). Řešení systému rovnic (7) plyne např. z Věty IV.18 v [4]. □

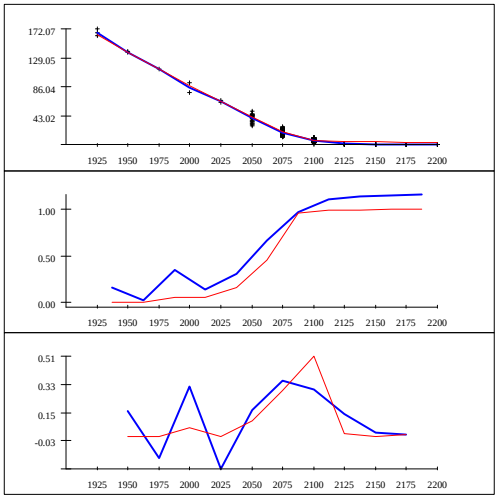
Zbývá už jenom volba vhodné pseudoinverze $\mathcal{A}^-$, tj., volba vhodného vektoru z v Lemma 2.

Lemma 3 *Hodnost matice* $\mathcal{A}^-\mathcal{A}-\mathbf{I}_p$ *je 1. Z toho plyne, že jakékoli řešení systému rovnic (7) je násobkem sloupcového součtu matice* $\mathcal{A}^-\mathcal{A}-\mathbf{I}_p$. *Vektor z v Lemma 2 může být zvolen jako* $z = \pm\mathbf{1}_p$, *kde znaménko je zvoleno tak, aby získané řešení bylo kladné.*

Důkaz: Z definice pseudoinverzní matice víme, že

$$\mathcal{A}\mathcal{A}^-\mathcal{A}-\mathcal{A} = \mathcal{A}(\mathcal{A}^-\mathcal{A}-\mathbf{I}_p) = 0. \tag{8}$$

Lemma 2 říká, že $\operatorname{rank}\mathcal{A} = p-1$. Vztah (8) implikuje $\operatorname{rank}(\mathcal{A}^-\mathcal{A}-\mathbf{I}_p)\le 1$. Jelikož zřejmě $\mathcal{A}^-\mathcal{A}\neq\mathbf{I}_p$, tedy $0\neq\operatorname{rank}(\mathcal{A}^-\mathcal{A}-\mathbf{I}_p) = 1$. □



Obrázek 3: Porovnání odhadu bez a s podmínkami 1–3. Odshora: odhad závislosti cen opcí na realizační ceně a odhady první a druhé derivace. Kupní opce na DAX 17. ledna 1995.

3.2 Algoritmus

Navržený algoritmus se skládá z těchto kroků:

- získání počátečního odhadu $\hat{\beta}^S$, např. pomocí izotonické regrese [7] použité na odhady první derivace získané metodou v odstavci 2.2,
- transformace počátečního odhadu $\hat{\beta}^S$ na počáteční odhad parametru $\hat{\xi}^S$ metodou popsanou v odstavci 3.1,
- minimalizace součtu čtverců a získání odhadu parametrů $\hat{\xi}$ a $\hat{\beta} = \beta(\hat{\xi})$ popsaných v odstavci 3 numerickými metodami.

4 Aplikace

Z obrázku 1 je dobře vidět, že opce se obchodují pouze pro několik realizačních cen. Pokud se omezíme pouze na nejkratší dobu splatnosti, dostaneme graf na pravé straně, kde pozorujeme 410 obchodů pouze pro $p = 12$ různých realizačních cen pravidelně rozmístěných mezi 1925DM a 2300DM. Pouze osm obchodů přitom proběhlo s opcemi s realizační cenou nižší než 2000DM.

Použitím navržené metody a aplikací algoritmu popsaného v odstavci 3.2 získáme odhady nakreslené na obrázku 3. Horní graf obsahuje pozorované ceny opcí, odhad ceny opcí pomocí lineárního modelu bez jakýchkoliv omezení a odhad ceny opcí splňující podmínky 1–3. Na tomto grafu oba odhady téměř splývají.

Prostřední a spodní graf na obrázku 3 ukazují postupně odhady první a druhé derivace. Odhad nesplňující podmínky 1–3 je označen tmavší čarou a je zřejmé, že tady se oba odhady již velice liší. Největší rozdíly pozorujeme pro odhad druhé derivace (RNH) pro realizační ceny nižší než 2000DM. To je způsobeno tím, že v této oblasti máme k dispozici pouze malý počet pozorování.

5 Závěr

Z obrázku 3 je zřejmé, že ignorování podmínek 1–3 může vést k velice špatným odhadům derivací regresní funkce, zvláště pro malý počet pozorování. Proto je důležité zabývat se metodami, které tyto podmínky umožní splnit.

Výhodou navržené metody je i její jednoduchost, která umožňuje snadnou konstrukci konfidenčních intervalů a další zobecnění pro korelovaná data nebo pro ceny prodejních opcí.

Reference

[1] Aït-Sahalia Y., Lo A.W. (1998). *Nonparametric estimation of state-price densities implicit in financial asset prices.* Journal of Finance **53**, 499 – 547.

[2] Aït-Sahalia Y., Lo A.W. (2000). *Nonparametric risk management and implied risk aversion.* Journal of Econometrics **94**, 9 – 51.

[3] Aït-Sahalia Y., Wang Y., Yared F. (2000). *Do option markets correctly price the probabilities of movement of the underlying asset?.* Journal of Econometrics **102**, 67 – 110.

[4] Anděl J. (1985). *Mathematical statistics (in Czech).* SNTL/Alfa, Prague.

[5] Breeden D., Litzenberger R. (1978). *Prices of state-contingent claims implicit in option prices.* Journal of Business **51**, 621 – 651.

[6] Jackwerth J.C. (1999). *Option-implied risk-neutral distributions and implied binomial trees: a literature review.* Journal of Derivatives **7**, 66 – 82.

[7] Robertson T., Wright F.T., Dykstra R.L. (1988). *Order restricted statistical inference.* Wiley, Chichester.

[8] Yatchew A., Härdle W. (2004). *Nonparametric state price density estimation using constrained least squares and the bootstrap.* Journal of Econometrics, to appear.

Poděkování: Děkuji MŠMT ČR za podporu projektu 1K04018 "Dynamické semiparametrické a neparametrické modely s aplikacemi ve financích" programu "Podpora začínajících pracovníků výzkumu" (1K) a výzkumnému záměru MSM 113200008.

Adresa: Z. Hlávka, Karlova Univerzita v Praze, Matematicko-fyzikální fakulta, Katedra pravděpodobnosti a matematické statistiky, Sokolovská 83, 186 75 Praha 8

E-mail: `hlavka@karlin.mff.cuni.cz`

DVOUROZMĚRNÁ ROZDĚLENÍ CHARAKTERISTIK SFÉROIDŮ; EXTRÉMY A STEREOLOGIE

Daniel Hlubinka

Klíčová slova: Stereologie sféroidů, výběrový extrém, dvourozměrná rozdělení, normalizační konstanty.

Abstrakt: V článku popíšeme dvourozměrné modely pravděpodobnostních rozdělení vhodné pro parametrickou a semiparametrickou analýzu extrémů sféroidů pomocí stereologických nástrojů. Ke stabilitě oblasti přitažlivosti maxim při stereologické transformaci potřebujeme určitou ekvivalenci chování chvostů. Náš článek věnujeme popisu konstrukce vhodných pravděpodobnostních modelů splňujících požadovanou ekvivalenci.

Úvod

Stereologie je matematická discipína zabývající se rekonstrukcí vlastností prostorových objektů na základě jejich projekcí či řezů nižší dimenze. Aplikace statistiky ve stereologii jsou významné například v materiálových vědách. Typickou ukázkou je zkoumání mikroskopických trhlin v kovech, které lze pozorovat pouze pomocí řezu materiálu, takzvaných *profilů*. Specifickým problémem zde je výzkum extrémně velkých, či extrémně deformovaných částic.

Dlouhou historii má takzvaný Wicksellův problém, viz [15] a [16]. V neprůhledném materiálu se uvažují malé kulové částice, jejichž poloměr je spojitou náhodnou veličinou. Na řezu materiálem lze pozorovat kružnice, jejichž poloměr má hustotu rozdělení danou takzvanou Wicksellovou transformací. Na základě odhadnuté hustoty poloměrů kružnic odhadujeme hustotu rozdělení poloměrů původních částic.

Zobecnění koulí na jiné typy částic nemusí vést k řešitelnému problému. Běžným zobecněním je přechod od koulí ke zploštěným (ve tvaru čočky) či protáhlým (ve tvaru doutníku) sféroidům. Ty jsou charakterizované tím, že mají pouze dvě různé délky poloos, u zploštěných sféroidů jsou stejné dvě delší (hlavní) poloosy, u protáhlých jsou naopak stejně dlouhé dvě kratší (vedlejší) poloosy. Řešení tohoto problému lze najít například v článcích [2], [3]. Uvedené články obsahují také zdůvodnění, proč nelze uvažovat zcela obecné sféroidy.

V posledních letech se začala v souvislosti se stereologií rozvíjet oblast stereologie extrémů. První na řadě byl pochopitelně Wicksellův problém, jehož zkoumání lze nalézt například v [4], či v sérii [10], [11], [12], [13]. Otázka zní, jak na základě pozorování extrémně velkých kružnic na řezu materiálem získáme představu o extrémech poloměrů původních koulí.

V následujícím příspěvku se budeme zabývat stereologickým problémem extrémů sféroidů. Navazujeme zde na práce [6], [7] a [1], kde byl daný problém postupně studován. Vzhledem k rozsahu článku se omezíme na popis vhodných pravděpodobnostních modelů. Začneme připomenutím hlavních výsledků ze stereologie a teorie výběrových extrémů. Poté ukážeme, že za určitých podmínek stejnoměrnosti (ekvivalence chvostů) lze dokázat shodu oblasti přitažlivosti maxim pro sféroidy a jejich řezy – až na jejich případný parametr, jehož změna je ovšem jednoznačně dána. Několik vhodných konstrukcí dvourozměrných pravděpodobnostních rozdělení splňujících podmínku stejnoměrnosti je předmětem druhé, hlavní, části příspěvku.

1 Stereologie a extrémy

Připomeňme základní pojmy. Obecný sféroid je částice charakterizovaná třemi poloosami o různé délce. Ve stereologii se však uvažují dvě užší třídy sféroidů a to *zploštěné* a *protáhlé*. V našem článku se omezíme jenom na zploštěné sféroidy, analýza protáhlých je dosti podobná. Náhodný zploštěný sféroid je charakterizován dvěma stejně dlouhými hlavními poloosami, zde značenými X a jednou kratší poloosou V. Přednost se dává charakterizaci pomocí *velikosti* X a *tvaru*, který je definován jako $T = X^2/V^2 - 1$. Triviálně tvar $T = 0$ určuje kouli, čím větší má tvar hodnotu, tím plošší sféroid je.

Řezem náhodného sféroidu je náhodná elipsa s hlavní poloosou Y a tvarem $Z = Y^2/W^2 - 1$, kde W je délka vedlejší poloosy. Pro řez sféroidu vždy platí $X \geq Y$ a $T \geq Z$.

Věta 1.1. *Předpokládejme, že existuje sdružená hustota $g(x,t)$ vektoru (X,T). Za předpokladu isotropního rozložení sféroidů v daném materiálu lze odvodit sdruženou hustotu $f(y,z)$ velikosti a tvaru řezu. Je dána*

$$f(y,z) = \frac{y\sqrt{1+z}}{2M} \int_y^\omega \int_z^\eta \frac{g(x,t)dtdx}{\sqrt{t}\sqrt{1+t}\sqrt{t-z}\sqrt{x^2-y^2}}, \tag{1}$$

kde M je polovina průměrné projekční velikosti v populaci částic.

Mějme na paměti, že částice jsou v materiálu prostorově náhodně orientované, v případě isotropního modelu jde o rovnoměrné rozdělení orientace.

Teorie výběrových extrémů je velmi rozvinutá oblast asymptotické statistiky, viz například [5]. Rozhodli jsme se zkoumat extrémy velikosti, případně tvaru, sféroidu podmíněné vždy zbývající charakteristikou, tedy vlastně extrémy jednorozměrných rozdělení. Modely pro vícerozměrné extrémy jsou, kvůli problémům s uspořádáním, hůře aplikovatelné. Jejich popis lze najít například v [8].

Připomeňme, že existují tři jednorozměrná rozdělení stabilní vůči maximům a to rozdělení Fréchetovo, Weibullovo a Gumbelovo s distribučními funkcemi (postupně)

$$\Lambda_i = \begin{cases} \exp(-v^{-\alpha}), & v \geq 0,\ \alpha > 0 \\ \exp(-(-v)^{\alpha}), & v \leq 0,\ \alpha > 0, \\ \exp(-\mathrm{e}^{-v}), & v \in \mathbb{R}, \end{cases} \tag{2}$$

kde $i = 1, 2, 3$ pro Fréchetovo, Weibullovo a Gumbelovo rozdělení. Dále připomeňme, že náhodná veličina U patří do oblasti přitažlivosti maxim pro Λ_i, existuje-li posloupnost *normalizačních konstant* (a_n, b_n) taková, že pro výběrový extrém $U_{(n)}$ platí $\mathrm{P}[(U_{(n)} - b_n)/a_n \leq x] \to \Lambda_i(x)$ pro n rostoucí nade všechny meze.

Věta 1.2. *Předpokládáme-li pro distribuční funkci K existenci hustoty k, pak postačující podmínkou pro to, aby $K \in \mathrm{MDA}(\Lambda_i)$ je*

$$\begin{aligned} \lim_{s\to\infty} \frac{k(xs)}{k(s)} &= x^{-(\alpha+1)}, \quad x > 0,\ \omega = +\infty, \textit{ Fréchetova MDA} \\ \lim_{s\searrow 0} \frac{k(\omega - xs)}{k(\omega - s)} &= x^{\alpha-1}, \quad x > 0,\ \omega < +\infty \textit{ Weibullova MDA} \\ \lim_{s\nearrow\omega} \frac{k(s + x\beta(s))}{k(s)} &= \mathrm{e}^{-x}, \quad x \in \mathbb{R} \qquad\qquad \textit{Gumbelova MDA} \end{aligned} \tag{3}$$

kde $\omega = \sup\{x : k(x) > 0$. Pomocná funkce β může být volena z funkce přežití

$$\beta(x) = \frac{\int_x^\omega 1 - K(t)dt}{1 - K(x)}.$$

2 Stabilita oblasti přitažlivosti maxim

Označme si pro potřeby následujícího textu

- $g_x(t)$ a $g_t(x)$ jsou pro charakteristiky sféroidu po řadě podmíněné hustoty tvaru T při dané velikosti $X = x$ a naopak velikosti při daném tvaru.
- $f_y(z)$ and $f_z(y)$ jsou definovány podobně jako g_x a g_t ovšem pro profily.
- $g_1(x), g_2(t), f_1(y), f_2(y)$ jsou marginální hustoty po řadě velikosti a tvaru sféroidu a profilu.

Uveďme nyní výsledek umožňující zkoumání extrémů sféroidů na základě pozorování extrémů jejich řezů. K dispozici máme následující

Věta 2.1. *[Stabilita MDA] Mějme sdruženou hustotu $g(x, t)$ rozdělení velikosti a tvaru sféroidu. Pak*

1. *Pokud podmíněná hustota tvaru při dané velikosti $g_x(t)$ splňuje některou z podmínek (3)* stejnoměrně v *x pak obě hustoty, podmíněná hustota tvaru řezu při dané velikosti řezu $f_y(z)$ a marginální hustota tvaru řezu $f_2(z)$, splňují stejnou podmínku. Parametr se změní na $\gamma = \alpha$ pro Fréchetovo rozdělení a $\gamma = \alpha + 1/2$ pro Weibullovo.*

2. *Pokud podmíněná hustota velikosti při daném tvaru $g_t(x)$ splňuje některou z podmínek (3)* stejnoměrně v x, *a $\alpha > 1$ pro Fréchetovo rozdělení pak obě hustoty, podmíněná hustota velikosti řezu při daném tvaru řezu $f_z(y)$ a marginální hustota velikosti řezu $f_1(y)$, splňují stejnou podmínku. Parametr se změní na $\gamma = \alpha - 1$ pro Fréchetovo rozdělení a $\gamma = \alpha + 1/2$ pro Weibullovo.*

Pozastavme se nad požadavkem stejnoměrnosti. Podstatné zde je, že všechna podmíněná rozdělení mají stejnou oblast přitažlivosti maxim, navíc limitní rozdělení mají stejný parametr a rychlost konvergence levých stran v (3) nezávisí na podmínce. Co nás k takovému požadavku vede?

Nejprve si uvědomme, že pozorujeme-li profil sféroidu o dané velikosti a tvaru, pak velikost i tvar sféroidu mohou být libovolné větší hodnoty než napozorované. Jinými slovy, podmiňujeme-li pozorovaným tvarem (velikostí) profilu, pak do těchto pozorování mohou přispívat všechny sféroidy s nejméně tak velkým tvarem (velikostí). Chceme-li něco říci o extrémech profilů, musíme uvažovat u extrémů tvaru (velikosti) sféroidů v nějakém smyslu stejné chování vůči podmínce velikosti (tvaru).

3 Modely s ekvivalentními chvosty

K sestrojení vhodného dvourozměrného rozdělení můžeme použijeme cestu kopulí, nebo cestu integrace.

3.1 Kopule

Cesta kopulí v našem případě propojuje dvě marginální hustoty do hustoty dvourozměrného náhodného vektoru pomocí funkce zvané kopule. Konstruujeme

$$G(x,t) = C\big(G_1(x), G_2(t)\big)$$
$$g(x,t) = g_1(x)g_2(t)\frac{\partial^2}{\partial x \partial t} C\big(G_1(x), G_2(t)\big)$$

kde *kopule* $C : [0,1]^2 \to [0,1]$ je vhodná diferencovatelná funkce. V takovém případě tedy začínáme dvěma marginálními rozděleními, pro velikost a pro tvar, a závislost mezi X a T, tedy i požadovaná ekvivalence chvostů je dána volbou kopule.

Uvažujme kopule konstruované pomocí funkce C, jejíž řezy jsou dány kvadratickou či kubickou funkcí druhého argumentu. Kopuli ve tvaru

$$C(u,v) = uv + \psi(u)v(1-v),$$

kde $\psi : [0,1] \to \mathbb{R}$ je absolutně spojitá, $|\psi'(u)| \le 1$ a $|\psi(u)| \le u \wedge (1-u)$, nazýváme kopulí s *kvadratickými řezy*. Kopule s *kubickými řezy* má tvar

$$C(u,v) = uv + \psi(v)u(1-u)^2 + \xi(v)u^2(1-u),$$

kde ψ a ξ jsou opět absolutně spojité a $1+\psi'(v)(1-4u+3u^2)+\xi'(v)(2u-3u^2)\geq 0$.

Již z tvaru kopulí je zřejmé, že zásadní roli pro ekvivalenci chvostů hrají funkce ψ a ξ. Shrňme si tyto předpoklady do tvrzení.

Věta 3.1. *Nechť je sdružená hustota $g(x,t)$ dána marginálními hustotami $g_1(x)$ a $g_2(t)$ a kopulí C. Pak*

1. *Je-li C kopule s kvadratickými řezy a navíc $|\psi'(u)|\leq\lambda<1$, pak pro model*
$$g_x(t)=g_2(t)\big[1+\psi'\big(G_1(x)\big)\big[1-2G_2(t)\big]\big]$$
platí ekvivalence chvostů.

2. *Je-li C kopule s kubickými řezy a navíc*
$$|1+2\psi'(v)-\xi'(v)|\geq\lambda>0,\ \forall\ v\ \in[0,1]$$
pak pro model
$$\begin{aligned}g_x(t)=g_2(t)\big[&1+\psi'\big(G_1(x)\big)\big(1+G_2(t)\big)-\xi'\big(G_1(x)\big)G_2(t)\\&+3\big\{\psi'\big(G_1(x)\big)+\xi'\big(G_1(t)\big)\big\}G_2(t)\big(1-G_2(t)\big)\big]\end{aligned}$$
platí ekvivalence chvostů.

Uvedená věta platí samozřejmě i se zaměněnými symboly x a t. Důkaz této věty se provádí ověřením podmínek (3) z věty 1.2, stejnoměrnost v podmínce je zaručena požadavky na ψ a ξ.

3.2 Stochastická jádra

Cesta integrace vede přes podmíněnou strukturu a jedno marginální rozdělení. Hustota dvourozměrného náhodného vektoru je pak určena

$$g(x,t)=\int_{\mathbb{R}}g_x(t)g(x)\mathrm{d}x,$$

kde $g_x(t)$ je hustota pro s.v. x a jde o zobrazení měřitelné v x, tedy *stochastické jádro*. Lze též definovat opačně $g(x,t)=\int g_t(x)g(t)\mathrm{d}t$. V tomto případě tedy rovnou popisujeme závislost zkoumané proměnné na podmínce a přidáváme marginální rozdělení podmiňovací veličiny.

Vezměme několik typů parametrických tvarů chvostů, kde parametr závisí na podmínce a zkoumejme, jak moc mohou dané parametry na podmínce záviset. Zejména se podíváme na rozdělení s polynomickými chvosty s nosičem shora omezeným i neomezeným a na rozdělení s exponenciálními chvosty se shora neomezeným nosičem.

Označme symbolem $h\approx k$ pro dvě hustoty h,k a jejich distribuční funkce H,K fakt, že

$$\lim_{x\to\omega}\frac{\int_x^\omega h(y)dy}{\int_x^\omega k(y)dy}=\lim_{x\to\omega}\frac{1-H(x)}{1-K(x)}=1.$$

Nyní můžeme shrnout do následující věty:

Věta 3.2. *Uvažujme sdružené rozdělení $g(x,t)$ získané vyintegrováním jádra podmíněných hustot $g_t(x)$ přes marginální hustotu $g_2(t)$. Platí*

1. *Nechť*
$$g_t(x) \approx a(t)\big(\log x\big)^{b(t)} x^{-c(t)-1} \text{ pro velká } x.$$
Je-li $c(t) \equiv c$ a $b(t)$ omezená funkce, pak model $g_t(x)$ splňuje podmínku pro konvergenci k Fréchetovu rozdělení stejnoměrně.

2. *Nechť*
$$g_t(x) \approx a(t)\left(\frac{x}{\omega}\right)^{b(t)-1}(\omega - x)^{c(t)-1}.$$
Je-li $c(t) \equiv c$ a $b(t)$ omezená funkce, pak model $g_t(x)$ splňuje podmínku pro konvergenci k Weibullovu rozdělení stejnoměrně.

3. *Nechť*
$$g_t(x) \approx a(t)x^{b(t)}\exp\{-c(t)x^{d(t)}\}.$$
Je-li $c(t) \equiv c$, $d(t) \equiv d$ a $b(t)$ omezená funkce, pak model $g_t(x)$ splňuje podmínku pro konvergenci ke Gumbelovu rozdělení stejnoměrně.

Uvedená věta platí samozřejmě i se zaměněnými symboly x a t. Důkaz této věty je velmi snadný, jde jen o přímé a snadné ověření podmínek (3) z věty 1.2. Jak si lze snadno povšimnout, některé parametry musí být na podmínce nezávislé. Na druhou stranu nám stačí jen asymptotický parametrický tvar chvostu k našim účelům.

3.3 Sdružené rozdělení poloos

Prozatím jsme používali sdruženou hustotu velikosti a tvaru $g(x,t)$. Lze však vyjít i ze sdružené hustoty obou poloos $h(x,v)$ a tuto transformovat na $g(x,t)$. Uvažujme jednoduchý model s použitím integrační konstrukce i pro tuto alternativu.

Hledáme sdružené rozdělení dvourozměrného vektoru (X,V). Všimněme si, že pro nosič takového vektoru platí

$$\operatorname{supp}(X,V) \subset \{(x,v); 0 \le v \ \le x\}.$$

Smysluplné se zdá být modelování rozdělení delší poloosy X a podmíněného rozdělení V při daném $X = x$. Za podmíněné rozdělení volme Beta rozdělení zobecňující rovnoměrné rozdělení na intervalu $[0,1]$, ale poznamenejme, že by nám opět stačilo předpokládat asymptotické chování sdružené hustoty v pro v blízká x. V našem případě samozřejmě upravíme nosič rozdělení na $[0,x]$.

Uvažujme sdružené rozdělení délek poloos (X,V) ve tvaru

$$h(x,v) = g_1(x)\frac{1}{x\mathrm{B}\big(a(x),b(x)\big)}\left(\frac{v}{x}\right)^{a(x)-1}\left(1-\frac{v}{x}\right)^{b(x)-1} \tag{4}$$

odkud ze substituce plyne $g(x,t) = h\big(x, x(1+t)^{-1/2}\big)x/\big(2(1+t)^{3/2}\big)$ a tedy

$$g(x,t) = g_1(x)\frac{x^{b(x)}}{2\mathrm{B}\big(a(x),b(x)\big)}(1+t)^{-\left(a(x)+b(x)+1\right)/2}\big((1+t)^{1/2}-1\big)^{b(x)-1}$$

a tedy podmíněné chvosty tvarového faktoru při velikosti budou polynomické. Všimněme si, že tímto způsobem nelze modelovat podmíněné rozdělení velikosti při daném tvarovém faktoru. Při substituci $(x,v) \mapsto (x, x^2/v^2 - 1)$ lze snadno podmínit druhou složku pomocí první, ale naopak stojíme před neřešitelným problémem.

Věta 3.3. *Nechť je dáno sdružené rozdělení velikostí poloos sféroidů předpisem (4). Pak pro podmíněné rozdělení tvaru sféroidu při dané velikosti platí ekvivalence chvostů, je-li $a(x) \equiv a$ a $b(x)$ je omezená funkce.*

4 Poznámky

Poznamenejme, že i pro konstrukci pomocí kopulí vlastně stačí popsat parametricky pouze asymptotický tvar $g_1(t)$ (nebo $g_2(t)$). Otázkou je, zda musíme předpokládat parametrický tvar ψ, ξ a g_2 (nebo g_1), abychom mohli odhadnout hodnoty $\psi'\big(G_2(x)\big)$ a $\xi'\big(G_2(x)\big)$. Bohužel na těchto hodnotách budou normalizační konstanty podstatně záviset a proto vhodné neparametrické vyhlazovací metody pro odhad funkcí $\psi'\big(G_2(\cdot)\big)$ a $\xi'\big(G_2(\cdot)\big)$ jsou vítány.

Jak hodláme využít předchozí výsledky k odhadu extrémních velikostí či tvarových faktorů? Teď již víme, že za jistých podmínek stejnosti nám extrémy profilů konvergují ke stejnému limitnímu rozdělení jako extrémy sféroidů – až na parametr. Představme si, že z pozorování extrémně velkých profilů s daným tvarem z zjistíme, že limitní rozdělení je Fréchetovo. Odhadneme, například metodou maximální věrohodnosti, [14], normalizační konstanty $\widehat{a}_n^p, \widehat{b}_n^p$. Pro předpokládaný asymptotický tvar chvostu lze spočítat, jak mají vypadat normalizační konstanty $(a_n^p, b?n^p)$ profilů a (a_n, b_n) původních částic, viz například [5], [1] a [9].

Dalším krokem je výpočet $\widehat{a}_n, \widehat{b}_n$ z hodnot $\widehat{a}_n^p, \widehat{b}_n^p$ a pomocí těchto odhadů pak odhadnout rozdělení maxima tvarového faktoru (pro danou velikost) či velikosti (pro daný tvar) sféroidu. Jednoduchý model spolu se simulacemi je možné najít v [1].

Reference

[1] Beneš V., Bodlák, K., Hlubinka D. (2003). *Stereology of extremes; Bivariate models and computation.* Methodology and Computing in Applied Probability **5**, 289 – 308.

[2] Cruz-Orive, L.-M. (1976). *Particle size-shape distributions; The general spheroid problem. I. Mathematical model.* Journal of Microscopy **107**, 235 – 253.

[3] Cruz-Orive, L.-M. (1978). *Particle size-shape distributions; The general spheroid problem. II. Stochastic model and practical guide.* Journal of Microscopy **112**, 153 – 167.

[4] Drees H., Reiss R.-D. (1992) *Tail behavior in Wicksell's corpuscle problem.* Probability Theory and Applications, Kluwer, Dordrecht, 205 – 220.

[5] Embrechts P., Klüppelberg C., Mikosh T. (1997). *Modelling Extremal Events.* Springer, Berlin.

[6] Hlubinka D. (2003). *Stereology of extremes; Shape factor of spheroids.* Extremes **6**, 5 – 24.

[7] Hlubinka, D. (2003). *Stereology of extremes; Size of spheroids.* Mathematica Bohemica **128**, 419 – 438.

[8] Kotz, S., Nadarajah, S. (2000). *Extreme value distributions; theory and practise.* Imperial College Press, London.

[9] Takahashi R. (1987). *Normalizing constants of a distribution which belongs to the domain of attraction of the Gumbel distribution.* Stat. Probab. Lett. **5**, 197 – 200.

[10] Takahashi R., Sibuya M. (1996). *The maximum size of the planar sections of random spheres and its application to metalurgy.* Ann. Inst. Stat. Math. **48**, 127 – 144.

[11] Takahashi R., Sibuya M. (1998). *Prediction of the maximum size in Wicksell's corpuscle problem.* Ann. Inst. Stat. Math. **50**, 361 – 377.

[12] Takahashi R., Sibuya M. (2001). *Prediction of the maximum size in Wicksell's corpuscle problem. II.* Ann. Inst. Stat. Math. **53**, 647 – 660.

[13] Takahashi R., Sibuya M. (2002). *Maximum size prediction in Wicksell's corpuscle problem for the exponential tail data.* Extremes **5**, 55 – 70.

[14] Weissman I. (1978). *Estimation of parameters and large quantiles based on the k largest observations.* JASA **73**, 812 – 815.

[15] Wicksell S. D. (1925) *The corpuscle problem I.* Biometrika **17**, 84 – 99.

[16] Wicksell S. D. (1926) *The corpuscle problem II.* Biometrika **18**, 152 – 172.

Poděkování: Tato práce vznikla za podpory výzkumného záměru MŠMT ČR MSM 113200008 *Matematické metody ve stochastice* a grantu GAČR 201/03/0946 *Modely stochastické geometrie a prostorová statistika.*

Adresa: D. Hlubinka, Matematicko-fyzikální fakulta UK, Katedra pravděpodobnosti a matematické statistiky, Sokolovská 83, 186 75, Praha 8

E-mail: daniel.hlubinka@mff.cuni.cz

LINEARIZÁCIA NELINEÁRNEJ REGRESIE A OBLASTI SPOĽAHLIVOSTI

Klára Hornišová

Kľúčové slová: Nelineárna regresia, linearizácia podmodelu, funkcie parametrov.

Abstrakt: V čiastočne lineárnych regresných modeloch sú pre niektoré funkcie parametrov známe presné inferencie založené na taylorovskej semilinearizácii modelu. Zostrojíme alternatívne presné inferencie využívajúce iné spôsoby linearizácie.

1 Linearizácia regresného modelu

Uvažujme nelineárny regresný model $\mathfrak{M}$

$$y = \eta(\theta) + \varepsilon \ , \quad \varepsilon \sim N(0, \sigma^2\Sigma) \ , \quad \theta \in \Theta \subseteq \mathbb{R}^p, \tag{1}$$

kde $\eta(.) : \Theta \to \mathbb{R}^N$ je merateľné zobrazenie, $y \in \mathbb{R}^N$ je vektor pozorovaní, $\varepsilon \in \mathbb{R}^N$ je vektor náhodných chýb, θ sú neznáme parametre, Σ je známa kladne definitná matica, $\sigma > 0$ je neznáme. Nech $g(\theta)$ je užitočná funkcia parametrov, kde pre $r \leq p$ je $g(.) : \Theta \mapsto \mathbb{R}^r$ merateľné zobrazenie.

Na zjednodušenie inferencií často model (1) linearizujeme. Najzvyčajnejšia je taylorovská linearizácia modelu (1) v nejakom bode $\theta^0 \in \Theta$:

$$y = \eta(\theta^0) + \frac{\partial\eta(\theta^0)}{\partial\theta^\top}(\theta - \theta^0) + \varepsilon =: A\theta + a + \varepsilon, \ \ \varepsilon \sim N(0, \sigma^2\Sigma), \tag{2}$$

kde θ^0 býva buď maximálne vierohodným alebo apriórne daným odhadom θ.

Pre prípad daného apriórneho rozdelenia π na Θ sa navrhli viaceré spôsoby linearizácie (1). Pri linearizácii vyhladzovaním z [8] sa linearizácia

$$Y = A\theta + a + \varepsilon \ , \quad \varepsilon \sim N(0, \sigma^2\Sigma),$$

vyberá podľa kritéria minimálnej strednej kvadratickej chyby (MSE)

$$MSE := \min_{\substack{A\in\mathbb{R}^{N\times p}\\ a\in\mathbb{R}^N}} E_\pi[\|\eta(\theta) - (A\theta + a)\|^2_\Sigma],$$

kde $\|z\|^2_\Sigma := z^\top\Sigma^{-1}z$ pre $\forall z \in \mathbb{R}^N$. Riešenie tejto minimalizačnej úlohy je

$$A = Cov_\pi(\eta, \theta)(Var_\pi\theta)^-, \qquad a = E_\pi\eta - AE_\pi\theta. \tag{3}$$

Podľa rovnakého kritéria možno hľadať aj najlepšiu aproximáciu modelu (1) vnútorne lineárnym modelom, t.j. (pozri [7]) modelom tvaru

$$y = A\beta(\theta) + a + \varepsilon \ , \quad \varepsilon \sim N(0, \sigma^2\Sigma),$$

kde $A_{N\times k}$, $k \leq p$, $h(A) = k$, $a \in \mathbb{R}^N$ a $\beta(.) : \Theta \mapsto \mathbb{R}^k$ je parametrizácia (1).

Optimálna vnútorná linearizácia modelu (1) má podľa [5] tvar

$$\begin{aligned} &k = \max\{i; i \leq p, \lambda_i > 0\} \ , \ A = \Sigma^{1/2}(u_1, \ldots, u_k), \\ &a = E_\pi\eta \ , \ \beta(\theta) = (A^\top\Sigma^{-1}A)^{-1}A^\top\Sigma^{-1}(\eta(\theta) - a), \end{aligned} \tag{4}$$

kde $u_1, \ldots, u_N$ sú ortonormálne vlastné vektory, porade zodpovedajúce vlastným hodnotám $\lambda_1 \geq \cdots \geq \lambda_N \geq 0$ matice $Var_\pi(\Sigma^{-1/2}\eta)$.

V [3] sa pri znalosti združeného apriórneho rozdelenia π pre (θ, σ^2), nelinearizuje model (1), ale priamo sa odhaduje funkcia parametrov $g(\theta, \sigma) \in \mathbb{R}^r$ explicitnou linearizáciou, t.j. odhadom $A^\top y + a$, ktorý minimalizuje priemernú strednú kvadratickú chybu (AMSE)

$$\min_{A\in\mathbb{R}^{N\times r}, a\in\mathbb{R}^r} E_\pi E_{f(.|\theta,\sigma)}\|g(\theta, \sigma) - (A^\top y + a)\|^2,$$

kde $f(.|\theta, \sigma)$ je podmienená hustota y. Riešením je

$$A = [Var_\pi\eta + E_\pi(\sigma^2)\Sigma]^{-1}Cov_\pi(\eta, g) \ , \ a = E_\pi g - A^\top E_\pi\eta. \tag{5}$$

2 Linearizácia podmodelu a oblasti spoľahlivosti pre funkciu parametrov

V modeli (1) sa používajú rôzne druhy oblastí spoľahlivosti pre $g(\theta)$ so spoľahlivosťou približne $(1 - \alpha)$, (napr. [4]). Ďalej sa budeme zaoberať len rozličnými variantmi oblasti spoľahlivosti založenej na projektoroch:

$$\begin{aligned} &\{g \in g(\Theta); F(g, y) := \\ &\frac{(N - p)\|(P(g, y) - P_2(g, y))(y - \eta(\tilde{\theta}_g))\|^2_\Sigma}{r\|(I - P(g, y))(y - \eta(\tilde{\theta}_g))\|^2_\Sigma} < F_{r,N-r-q}(1 - \alpha)\}, \end{aligned} \tag{6}$$

ktorej zodpovedá nasledujúca kritická oblasť testu hypotézy $H_0 : g(\theta) = g^0 \in g(\Theta) \subseteq \mathbb{R}^r$ oproti $H_1 : g(\theta) \neq g^0$ s hladinou približne $(1 - \alpha)$:

$$\{y; F(g^0, y) < F_{r,N-r-q}(1 - \alpha)\}.$$

Tu $\hat{\theta}$ a $\tilde{\theta}_g$ označujú porade maximálne vierohodné odhady θ v modeli $\mathfrak{M}$ a v podmodeli $\mathfrak{M}_g$ (implicitne) danom hypotézou $H_g : g(\theta) = g$, a pre $\forall y$ sú $P(g, y)$ a $P_2(g, y)$ také $\Sigma-$ortogonálne projektory v $\mathbb{R}^n$, že pre podpriestory

generované ich stĺpcami platí $\mathcal{M}(P_2)\subseteq\mathcal{M}(P)$, $\mathrm{h}(P)=r+q\le p$, $\mathrm{h}(P-P_2)=r$. Potom pre $\forall y$ existuje $D_{N\times r}=D(g,y)$ taká, že (pozri [4])

$$P=P_2+(I-P_2)D(D^\top\Sigma^{-1}(I-P_2)D)^{-1}D^\top(I-P_2)^\top\Sigma^{-1}.$$

Za P_2 možno vybrať napr. projektor na taylorovskú linearizáciu $\mathfrak{M}_g$ v $\tilde\theta_g$:

$$P_2:=P_A:=A(A^\top\Sigma^-A)^{-1}A^\top\Sigma^{-1},\qquad\text{kde}$$

$$A:=\left.\frac{\partial\eta(\theta)}{\partial\theta^\top}\right|_{\theta=\tilde\theta_g}(I_p-P_B)\qquad\text{, kde}\qquad B:=\left.\frac{\partial g^\top(\theta)}{\partial\theta}\right|_{\theta=\tilde\theta_g}.$$

Pri danom $P_2(g,y)$ treba zvoliť $D(y)$ tak, aby spĺňala

$$\mathrm{h}((I_N-P_2(g,y))D(g,y))=r. \tag{7}$$

Pre všeobecné $g(.)$, D a P_2 má (6) asymptoticky hladinu $1-\alpha$. Podľa [2] je oblasť (6) presná, ak súčasne platí (7), ďalej ak pre nejaké maticové funkcie premennej $g(\theta)$, $V(.):g(\Theta)\to\mathbb{R}^{N\times(p-r)}$ a $v(.):g(\Theta)\to\mathbb{R}^N$ spĺňajúce podmienku $\mathrm{h}(V(g))=q$ pre každé $g\in g(\Theta)$, a pre každé θ platí:

$$\eta(\theta)=V(g(\theta))\nu(\theta)+v(g(\theta)), \tag{8}$$

kde $\nu(.):\Theta\to\mathbb{R}^{p-r}$ je také, že $(g^\top(\theta),\nu^\top(\theta))^\top$ je regulárna reparametrizácia θ, a ak nakoniec D závisí od y iba prostredníctvom $V^\top(g)(y-v(g))$:

$$D(g,y)=D(g,V^\top(g)(y-v(g))). \tag{9}$$

Ak platí (8), tak $P_2(g)=P_{V(g)}$ a maximálne vierohodný odhad pre $\nu(\theta)$ pri podmienke $g(\theta)=g$ je

$$\tilde\nu(g,y)=[V^\top(g)\Sigma^{-1}V(g)]^{-1}V^\top(g)\Sigma^{-1}(y-v(g)),$$

takže podmienky (7), (8), (9) spĺňajú napr. prvé dve taylorovské semilinearizácie (t.j. taylorovské linearizácie modelu (8) parametrizovaného parametrom g pri konštantnom ν, do ktorých sa dosadí $\tilde\nu(g,y)$ za ν):

$$D(g,y)=D(g,V^\top(g)(y-v(g)))= \tag{10.a, b, c}$$

$$\sum_{i=0}^{p-r}\frac{\partial V_{.i}(g)}{\partial g^\top}w_i(g),\qquad\sum_{i=0}^{p-r}\left.\frac{\partial V_{.i}(g)}{\partial g^\top}\right|_{g=g^0}w_i(g),\qquad\sum_{i=0}^{p-r}\left.\frac{\partial V_{.i}(g)}{\partial g^\top}\right|_{g=\hat g}w_i(g),$$

kde $V_{.i}(g)$, $i=1,\dots,p-r$, sú stĺpce matice $V(g)$, $V_{.0}:=v(g)$, $w(g):=\tilde\nu(g,y)$, $w_0(g):=1$, g^0 je apriórne daný bod (pri voľbe c) pravdaže nedostaneme presnú oblasť). V [2] uvažovali voľbu a).

Ak poznáme apriórne rozdelenie $\pi(.)$ funkcie $g(\theta)$, či $g(\theta,\sigma)$, okrem taylorovských semilinearizácií možno v modeli (8) zostrojiť $D(g,y)$ tak, aby

zodpovedali prirodzeným linearizáciám (3), (4) či (5), a pritom spĺňali podmienky (7) a (9) postačujúce podľa [2] na presnosť inferencií (6):

1.) (semilinearizácia vyhladzovaním):

$$D = (Cov_{\pi(g)}(\eta, g) Var^{-1}_{\pi(g)}(g))\Big|_{\nu=\tilde{\nu}(g,y)} = \tag{11}$$
$$= [\sum_{i=0}^{p-r} Cov_{\pi(g)}(V_{.i}(g), g) w_i(g)] Var^{-1}_{\pi(g)}(g)$$

2.) (explicitná semilinearizácia):

$$D = ([Var_{\pi(g,\sigma^2)}(\eta) + E_{\pi(g,\sigma^2)}(\sigma^2)\Sigma]^{-1} Cov_{\pi(g,\sigma^2)}[\eta, g])\Big|_{\nu=\tilde{\nu}(g,y)} = \tag{12}$$
$$= [\sum_{i=0}^{p-r}\sum_{j=0}^{p-r} Cov_{\pi(g,\sigma^2)}(V_{.i}(g), V_{.j}(g)) w_i(g) w_j(g) + E_{\pi(g,\sigma^2)}(\sigma^2)\Sigma]^{-1} \cdot$$
$$\cdot [\sum_{i=0}^{p-r} Cov_{\pi(g,\sigma^2)}(V_{.i}(g), g) w_i(g)]$$

3.) (vnútorná semilinearizácia):

$$D = \Sigma^{1/2}(u_1, \ldots, u_r),$$

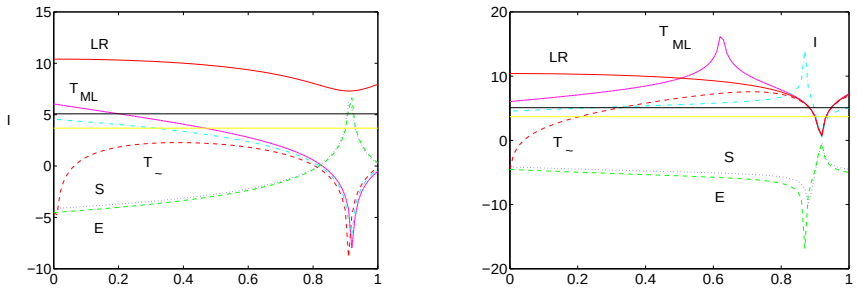
kde $u_1, \ldots, u_N$ sú $\Sigma-$ortonormálne vlastné vektory porade zodpovedajúce vlastným hodnotám $\lambda_1 \geq \cdots \geq \lambda_N \geq 0$ matice

$$(Var_{\pi(g)}[\Sigma^{-1/2}\eta])\Big|_{\nu=\tilde{\nu}(g,y)} = Var_{\pi(g)}[\Sigma^{-1/2}[V(g).\nu + v(g)]]\Big|_{\nu=\tilde{\nu}(g,y)} = \tag{13}$$
$$= \sum_{i=0}^{p-r}\sum_{j=0}^{p-r} Cov_{\pi(g)}(\Sigma^{-1/2}V_{.i}(g), \Sigma^{-1/2}V_{.j}(g)) w_i(g) w_j(g).$$

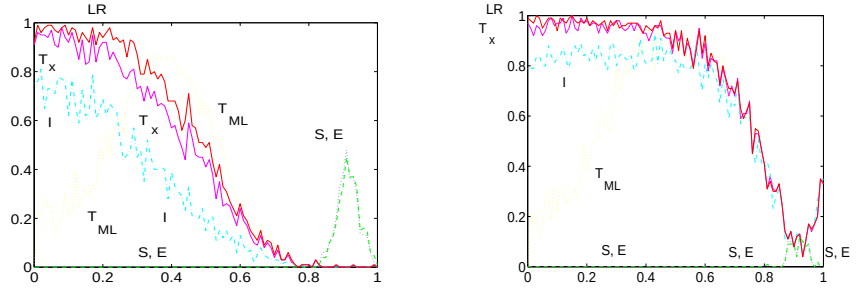
Teda model (8) pri konštantnom ν parametrizovaný parametrom g sa zlinearizuje spôsobom (3), (4) alebo (5) a do výslednej linearizácie sa potom dosadí $\tilde{\nu}(g, y)$ za ν.

Namiesto marginálneho apriórneho rozdelenia $\pi(g(\theta))$ pre $g(\theta)$ možno použiť podmienené rozdelenie $\pi(g(\theta)|\nu(\theta))$.

Poznámka. Iný postup je v [3] aj pre $g(\theta)$, ktoré nespĺňajú (7), (8) a (9). Pre θ sa zostrojí presná oblasť spoľahlivosti Ω tvaru (6) s $P_2 = 0$ a D zodpovedajúcou explicitnej linearizácii funkcie $g(\theta)$. Potom $g(\Omega)$ je konzervatívna oblasť spoľahlivosti pre $g(\theta)$ na tej istej hladine. Ak $g(\Omega)$ nie je uspokojivá, D sa dopĺňa stĺpcami zodpovedajúcimi explicitným linearizáciám ďalších skusmo vyberaných funkcií parametrov.



Obrázok 1: Funkcie $F(g, y)$ pre rôzne D.



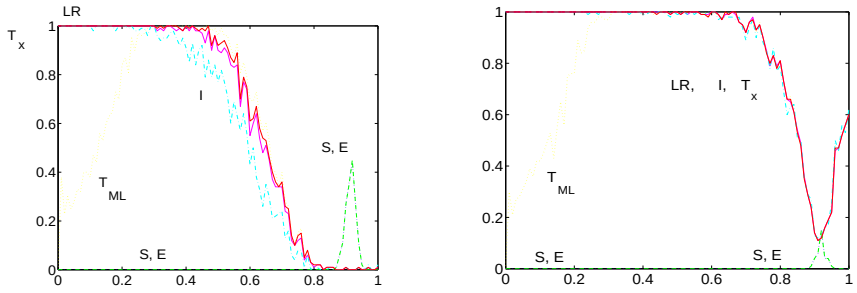
Obrázok 2: Silofunkcie pre $c = 10$, $\theta^0 = 0,916$.

3 Príklady

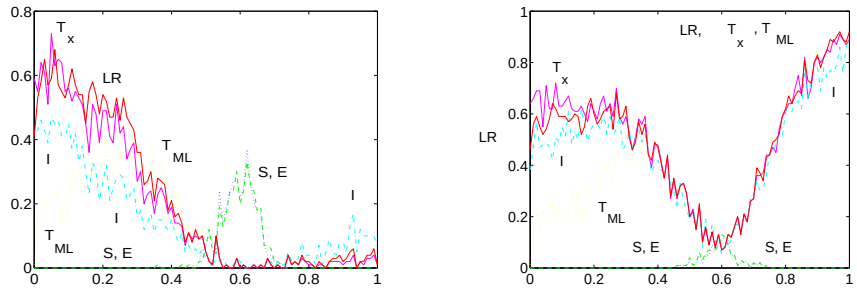
Príklad 1 Pre $g(\theta) = \theta$ a model so známym c

$$y \sim N((\theta, c\theta^2)^\top, \sigma^2 I_{2\times 2}),\ \theta \in \Theta =< -1, 1 >,$$

pre π_1 rovnomerné na Θ a nezávislé rozdelenie π_2 pre σ^2 s $E_{\pi_2}(\sigma^2) = 0,3^2$, $y = (0,87; 8,4)^\top$, kde ML-odhad $\hat{\theta} \doteq 0,916$, sú na obr. 1 pre $c = 10$ priebehy funkcií $\ln(L(g, y))$, kde $L(g, y)$ je pomer vierohodností (označenie LR), a $\ln(F(g, y))$ z (6) s D ako v (10c), (10a), (11), (12), (13) (porade označené $\mathrm{T_{ML}}$, $\mathrm{T_\sim}$, S, E, I), spolu s kritickými hodnotami $\ln(F_{1,1}(1-\alpha))$ pre $\alpha = 0,9$ a $0,95$, pre $-\theta \in <\ 0; 1\ >$ (časť a)) a $\theta \in <\ 0; 1\ >$ (časť b)). V a) je na $x-$ovej osi $-\theta$, v b) θ. Najužšie intervaly spoľahlivosti sú LR a $\mathrm{T_{ML}}$ s vhodným obmedzením tvaru $\|\eta(\theta) - \eta(\hat{\theta})\|_\Sigma < \rho$ (pozri [7]). Ak by sme takéto obmedzenie uvažovali aj pri ostatných druhoch, bol by rovnako vyhovujúci aj $\mathrm{T_\sim}$ a pri $1 - \alpha = 0,9$ aj I. Na obr. 2, 3, 4 sú silofunkcie príslušných testov hypotézy $H_0 : \theta = \theta^0$ oproti $H_1 : \theta \neq \theta^0$ pre $c = 10$ & $\theta^0 = 0,916$, $c = 20$ & $\theta^0 = 0,916$, $c = 10$ & $\theta^0 = 0,6$. Namiesto $\mathrm{T_\sim}$ uvažujeme $\mathrm{T_x}$, čo zodpovedá (6) s D ako v (10b) pre $g^0 = \theta^0$. Grafy sú dosť kostrbaté, lebo pre každú použitú hodnotu θ sa počítalo len 100 simulácií, no na porovnanie rôznych postupov to stačí. Najlepšie sú LR, $\mathrm{T_x}$ a sčasti I, a to tým výraznejšie, čím sú c a $|\theta^0|$ väčšie. V malom okolí θ_0 je dobrá aj $\mathrm{T_{ML}}$.



Obrázok 3: Silofunkcie pre $c = 20$, $\theta^0 = 0,916$.



Obrázok 4: Silofunkcie pre $c = 10$, $\theta^0 = 0,6$.

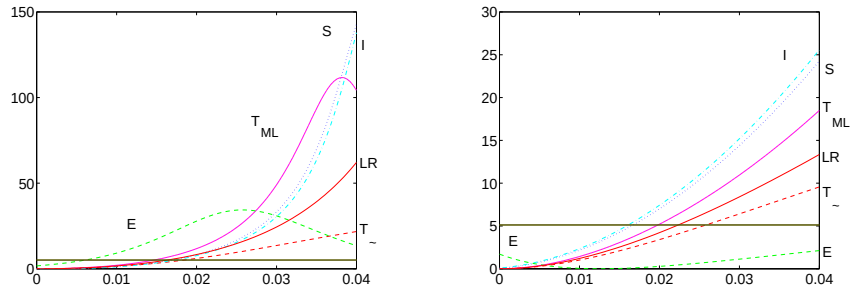
Príklad 2 Pre údaje z [4] a model s malou vnútornou krivosťou ([1])

$$y_i = \theta_1 x_i/(\theta_2 + x_i) + \varepsilon_i, \qquad \varepsilon \sim N[\bar{0}; \sigma^2 I_n],$$

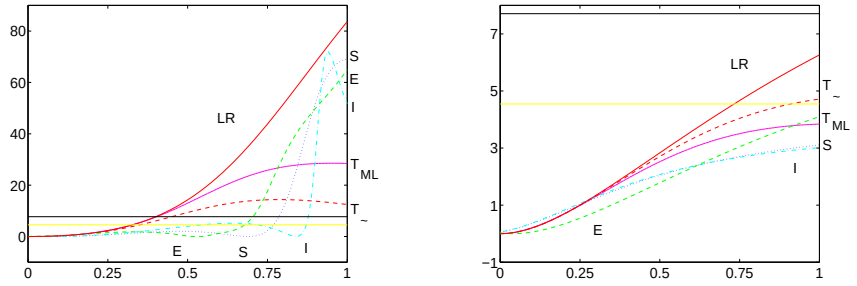
kde $\hat{\theta} = (160,28; 0,047709)$, $s^2 = 95,51 = E_\pi(\sigma^2)$ a pre $\pi(\theta_2)$ - rovnomerné na $0,047709 \pm 0,04$, sú intervaly spoľahlivosti pre $g(\theta) = \theta_2$ na hladine $0,95$:

$$\begin{aligned} T_{ML} &=< 0,0331; 0,0675 > \text{ , kde } \theta_2^0 = \hat{\theta}_2 \\ T_\sim &=< 0,0293; 0,0735 > \text{ , rovnako v [4]} \\ S &=< 0,0309; 0,0643 > \\ I &=< 0,0304; 0,0637 > \\ E &=< 0,0418; 0,1237 > \text{ , (bez obmedzenia daného nosičom } \pi) \\ LR \doteq Trub &=< 0,0314; 0.0701 > \text{ , podľa [6]} \end{aligned}$$

$Trub$ je konzervatívna oblasť spoľahlivosti nájdená trubicovou metódou ([6]). Na obr. 5 sú priebehy funkcií $L(g, y)$ a $F(g, y)$ pre θ_2 a kritická hodnota $F_{1,9}(0.95)$. Na a) je na $x-$ovej osi $\hat{\theta}_2 - \theta_2$ a na b) $\theta_2 - \hat{\theta}_2$. Najlepšie sú S, I a $\mathrm{T_{ML}}$, ktoré dávajú zhruba rovnako široké intervaly. Kým však nevezmeme do úvahy ohraničenosť nosiča $\pi(.)$, tieto intervaly typu (6) sú iba komponentom súvislosti presných oblastí spoľahlivosti, ktorý obsahuje $\hat{\theta}_2$.



Obrázok 5: Funkcie $F(g, y)$ pre rôzne D.



Obrázok 6: Funkcie $F(g, y)$ pre rôzne D.

Príklad 3 Pre údaje BOD I z [1] a model s veľkou vnútornou krivosťou

$$y_i = \theta_1(1 - \exp(-x_i\theta_2)) + \varepsilon_i \ , \qquad \varepsilon \sim N(0, \sigma^2 I_n),$$

kde $\hat{\theta}_2 = 0,5311$, $s^2 = 6,498 = E_\pi(\sigma^2)$, sú na obr. 6 pre $g(\theta) = \theta_2$ a $\pi(\theta_2)$ - rovnomerné na $\hat{\theta}_2 \pm 1$ priebehy funkcií $L(g, y)$ a $F(g, y)$ a kritické hodnoty $F_{1,4}(1 - \alpha)$ pre $1 - \alpha = 0,9$ a $0,95$. V a) je na $x-$ovej osi $\hat{\theta}_2 - \theta_2$ a v b) $\theta_2 - \hat{\theta}_2$. Najlepšia je LR, no i tá dáva pre $1-\alpha = 0,95$ interval, ktorý je zhora ohraničený len hranicou nosiča π, či hranicou parametrického priestoru. Bez tohto obmedzenia (ak $1 - \alpha = 0,95$):

$$LR =< 0,132; 1,77 >=< \hat{\theta}_2 - 0,3991; \hat{\theta}_2 + 1,2389 > \ , \text{podľa [1]}$$
$$Trub =< 0,130; 0,41 >=< \hat{\theta}_2 - 0,4011; \hat{\theta}_2 + 0,8789 > \ , \text{podľa [6]}.$$

Reference

[1] Bates D.M., Watts D.G. (1988). *Nonlinear regression analysis and its applications*. Wiley, New York.

[2] El-Shaarawi A., Shah K.R. (1980). *Interval estimation in non linear models*. Sankhya B **42**, 229 – 232.

[3] Gallant A.R. (1980). *Explicit estimators of parametric functions in nonlinear regression*. J. Am. St. Assoc. **75**, 182 – 193.

[4] Hamilton D. (1986). *Confidence regions for parameter subsets in nonlinear regression*. Biometrika **73**, 57 – 64.

[5] Hornišová K. (2004). *Intrinsic linearization of nonlinear regression by principal components method*. AMUC **LXXIII**, 207 – 216.

[6] Knowles M., Siegmund D., Zhang H. (1991). *Confidence regions in semilinear regression*. Biometrika **78**, 15 – 31.

[7] Pázman A. (1993). *Nonlinear statistical models*. Kluwer, Dordrecht.

[8] Pázman A. (2001). *Linearization of nonlinear regression models by smoothing*. Tatra Mt. Math. Publ. **22**, 13 – 25.

Poďakovanie: Na výskum prispela agentúra Vega grantom č. 2/4026/04.

Adresa: K. Hornišová, Ústav merania SAV, Dúbravská 9, 841 04 Bratislava

E-mail: `umerhorn@savba.sk`

SHLUKOVÁNÍ A TEXTOVÉ DOKUMENTY

Dušan Húsek, Hana Řezanková, Václav Snášel

Klíčová slova: Výpočetní statistika, shlukování, rozsáhlé datové soubory, dokumentografické informační systémy.

Abstrakt: V práci jsou zhodnoceny některé možnosti využití shlukovacích metod pro vyhledávání v textových dokumentech. Využití těchto metod je umožněno geometrizací vyhledávacího problému na základě vektorového modelu. Přestože tato problematika patří v oblasti vyhledávání mezi klasické, není v současné době uspokojivě vyřešena. Jedním z hlavních problémů při shlukování je vysoká dimenzionalita vstupních dat. V příspěvku jsou charakterizovány speciální postupy navržené pro shlukování takových vysoce dimenzionálních dat.

1 Úvod

Třída programových nástrojů určených pro zpracovávání, úschovu a výběr dat, kterými jsou textové dokumenty, je reprezentována textovými (dokumentografickými) informačními systémy (DIS). V takovém systému je vhodné s texty pracovat na více úrovních abstrakce, tj. vedle textů je vhodné popsat i jejich schéma pomocí odpovídajícího modelu. Toto schéma je ve svých funkcích podobné schématu klasických databází. Model DIS je definován jako soubor pojmů či nástrojů pro reprezentaci dokumentu (tvoří formální popis informace obsažené v dokumentu), reprezentaci dotazu (umožňuje specifikovat formálně požadavek na informace), reprezentaci pravidel a procedur umožňujících určit shodu mezi požadavkem uživatele na informace a dokumenty, které vyhovují tomuto požadavku. Mezi prominentní modely DIS patří v současnosti rozšířený Booleovský model, vektorový model a model založený na pravděpodobnostním výběru.

V tomto příspěvku se budeme zabývat vektorovým modelem. V něm je textový dokument reprezentován vektorem, jehož prvky charakterizují výskyt slov (resp. termů) obsažených v dokumentech. Geometrizace vyhledávacích problémů, viz [20], dává vznik mnoha zajímavým problémům spojeným se shlukováním dat ve vysoce dimenzionálních prostorech. Vektor může být buď binární (slovo se v dokumentu vyskytuje nebo ne), nebo může obsahovat četnosti výskytu, případně váhy založené na důležitosti slov v celé kolekci dokumentů. Pro analýzu pak máme k dispozici matici n x m, kde n je počet dokumentů a m je počet slov. Pro uvedenou matici je typické, že je velkých rozměrů, a to zejména pokud jde o počet sloupců, a že je velmi řídká (uvádí se, že nenulových prvků je obvykle pouze kolem dvou procent). Další podrobnosti viz [4].

Základní úlohou při analýze takových dat je shlukování dokumentů. Pomocí hierarchické shlukové analýzy lze nalézt různé úrovně skupin dokumentů. Cílem shlukové analýzy je nalézt shluky tak, aby dokumenty uvnitř

shluku si byly co nejvíce podobné a aby jejich podobnost s dokumenty z jiných shluků byla menší.

Pomocí shlukování dokumentů tak můžeme zjistit témata, která se vyskytují ve sledované kolekci textových dokumentů [7]. Dále mohou být na základě zjištěných skupin navrženy modely, pomocí nichž jednak může být nový dokument zařazen do některé ze skupin, jednak může být vyhledána skupina dokumentů, které nejvíce vyhovují zadanému dotazu.

Protože rozsah datové matice je obvykle značný, jsou využívány jednak metody redukce dimenzionality, jednak speciální postupy pro shlukování. Tyto postupy jsou buď přímo zaměřeny na problematiku textových dokumentů nebo jde o obecné metody určené pro analýzu rozsáhlých datových souborů (data mining).

Jako příklad prvního typu je shlukování náhodně vybraných dokumentů opakované pro různé výběry, které může vést ke stanovení množiny slov vhodné pro charakterizování sledované kolekce dokumentů (viz [25]). Shlukování dokumentů a případné vytváření modelů pro přiřazování dokumentů či dotazů ke zjištěným shlukům je pak prováděno s redukovaným počtem slov.

Jiné využití shlukování při analýze textových dokumentů vychází z toho, že při aplikaci metod strojového učení pro řešení klasifikačních úloh je potřeba najít vhodnou tréninkovou množinu, tj. takovou, která by neobsahovala příliš mnoho podobných dokumentů. Toho lze docílit tím, že jsou dokumenty rozděleny do shluků a do tréninkové množiny jsou vybírání zástupci těchto shluků. V [12] je navrženo použít metodu k-průměrů a z každého vytvořeného shluku vybrat dokument, který je nejbližší centroidu.

2 Měření podobnosti mezi textovými dokumenty

Nejčastějším ohodnocením výskytu slov v dokumentech je výpočet vah. Ty mohou být počítány různými způsoby (rozsáhlé experimenty s vážením termů jsou popsány v [21]), dva z nich jsou popsány v následující části.

2.1 Výpočet vah pro vstupní datovou matici

Políčka vstupní datová matice obsahují váhy w_{ij}, kde i označuje dokument ($i = 1, \ldots, n$) a j označuje slovo, které se vyskytuje ve zkoumané kolekci dokumentů ($j = 1, \ldots, m$). Označme si TF_{ij} četnosti výskytu j-tého slova v i-tém dokumentu a IDF_j inverzní četnost slov ve všech dokumentech, která je počítána jako $IDF_j = \log\,(n\ /\ k_j) + 1$, kde k_j je počet dokumentů, v nichž se vyskytuje j-té slovo. Potřebnou váhu pak získáme jako součin četnosti a odpovídající inverzní četnosti příslušného slova, tj. $w_{ij} = TF_{ij}\ *\ IDF_j$.

Ibrahimov v [11] navrhuje tzv. kombinovanou váhu počítanou jako (pozměněno značení)

$$w_{ij} = \frac{(K+1) * CFW_j * TF_{ij}}{K * ((1-b) + b * NDL_i) + TF_{ij}}\ ,$$

přičemž $CFW_j = \log\left(\frac{n}{k_j}\right)$ a NDL_i je délka i-tého dokumentu normalizovaná průměrnou délkou dokumentu. Parametr b je uživatelem zadané číslo a jeho doporučovaná hodnota je 0,75 (řídí vliv délky dokumentu), K je diskontní parametr, který souvisí s četností slov, Ibrahimov používá hodnotu 2.

Dále můžeme přiřadit jednotlivým dokumentům váhu podle následujícího vzorce (viz Ibrahimov): $DW_i = \sum_{k \in D_i} w_{ik}$, kde D_i je množina slov, jejichž váha je větší než stanovená prahová hodnota. Vztah dokumentu X k dokumentu Y lze vyjádřit jako

$$INTER(X,Y) = \frac{DW_{X \cap Y | k \in D_Y}}{DW_Y} \ .$$

2.2 Míry podobnosti

Pro měření podobnosti dvou textových dokumentů používá Ibrahimov chí-kvadrát test. Nulová hypotéza vyjadřuje shodu rozdělení vah pro vybranou množinu slov. Je používána chí-kvadrát statistika počítaná jako

$$\chi^2 = \sum_{k \in D_X \cap D_Y} \frac{(w_{Xk} - w_{Yk})^2}{w_{Yk}}$$

neboli

$$\chi^2 = \sum_{j \in D_X \cap D_Y} \frac{(x_j - y_j)^2}{y_j} \ .$$

K této statistice můžeme zjistit minimální hladinu významnosti, od které zamítáme nulovou hypotézu. Jestliže si tuto hodnotu označíme jako δ, můžeme si podobnost mezi dvěma dokumenty vyjádřit podle vzorce

$$sim\,(X,Y) = \delta * INTER(X,Y) \ .$$

Tato míra podobnosti nabývá hodnot v intervalu od 0 do 1. Častěji než tato asymetrická míra jsou v praxi používány spíše míry symetrické. Za základ je považována kosinová míra, kterou pro objekty (dokumenty) X a Y můžeme zapsat jako

$$s(X,Y) = \frac{\sum_{j=1}^{m} x_j y_j}{\sqrt{\sum_{j=1}^{m} (x_j)^2 \sum_{j=1}^{m} (y_j)^2}} \ ,$$

kde m je počet proměnných (slov).

V některých případech mohou být dokumenty charakterizovány pouze jako binární vektory (slovo se v dokumenty vyskytuje nebo ne), případně

jako vektory četností výskytu. I pro tyto případy existují speciální míry. Pro binární data je možné podobnost vyjádřit například pomocí vzorce

$$s(X,Y) = \frac{\Theta \sum_{j=1}^{m} x_j y_j}{\Theta \sum_{j=1}^{m} x_j y_j + \sum_{j=1}^{m} |x_j - y_j|} \ .$$

Pokud $\Theta = 1$, dostáváme Jaccardův koeficient, v případě, že $\Theta = 2$, jde o Diceovu (Czekanowského) míru podobnosti.

Pro četnosti lze využít chí-kvadrát míru nepodobnosti, která je vyjádřena jako

$$d(X,Y) = \sqrt{\sum_{j=1}^{m} \frac{(x_j - E(x_j))^2}{E(x_j)} + \sum_{j=1}^{m} \frac{(y_j - E(y_j))^2}{E(y_j)}} \ ,$$

kde

$$E(x_j) = \frac{\left(\sum_{j=1}^{m} x_j\right) \cdot (x_j + y_j)}{\sum_{j=1}^{m} x_j + \sum_{j=1}^{m} y_j} \quad a \quad E(y_j) = \frac{\left(\sum_{j=1}^{m} y_j\right) \cdot (x_j + y_j)}{\sum_{j=1}^{m} x_j + \sum_{j=1}^{m} y_j} \ .$$

3 Vyhledávání dokumentů na základě shluků

Aby mohl být při vyhledávání v textových dokumentech uspořen čas potřebný pro nalezení odpovědi na zadaný dotaz, lze v procesu předzpracování dat identifikovat shluky dokumentů, které pokrývají podobná témata. Tato problematika je označována jako vyhledávání shluků (cluster retrieval) a je rozpracována v knize [5]. V procesu vyhledávání shluků jsou uživateli prezentovány pouze dokumenty obsažené v jednom nebo několika vybraných shlucích.

Jako zajímavé téma zkoumané při vývoji metod je rozpoznávání výskytu překrývajících se shluků. Jako speciální metody pro nalezení shluků dokumentů jsou v [5] uvedeny iterativní reškálování, dynamické reškálování založené na latentním sémantickém indexování (LSI) a dynamické reškálování založené na analýze kovarianční matice. První metodu navrhla Ando v roce 2000. Je určena k identifikování malých shluků v omezeném kontextu. Vstupními parametry jsou matice typu dokumenty x termy, konstantní škálovací faktor a dimenze k, do které má být vyhledávání informací mapováno.

Tento algoritmus má však řadu nedostatků,proto autoři Kobayashi a Aono navrhli jednak jeho vylepšení, jednak algoritmus založený na jiném principu. Dynamické reškálování založené na latentním sémantickém indexování má být zmíněným vylepšením. Může však být použito pouze na malé datové soubory.

Bylo navrženo v roce 2001. Je vhodné poznamenat, že ideou všech tří zde uvedených algoritmů je uchovat hlavní témata při výběru základních vektorů pro podprostor, do kterého bude úloha vyhledávání informací mapována. To je v navržené metodě ošetřeno zavedením vah ke snížení důležitostí atributů (slov, termů), které již jsou reprezentovány podprostorem již vypočítaných základních vektorů. Přiřazování vah je řízeno dynamicky, aby se zabránilo ztrátě informace jak ve velkých tak malých shlucích.

Dynamické reškálování založené na analýze kovarianční matice je určeno k identifikování malých shluků. Tento algoritmus je možné (dle jeho autorů) použít pro velké datové soubory. Vstupními parametry jsou kovarianční matice, reškálovací faktor (používaný pro přiřazování vah) a dimenze k, do níž má být úloha redukována. Při výpočtech jsou používány dvě matice reziduí, přičemž na počátku je jedna tato matice tvořena kovarianční maticí pro celou množinu dokumentů.

Závěrem lze poznamenat, že metody navrhované v oblasti literatury zabývající se vyhledáváním informací zřejmě nejsou stále vhodné pro použití v případě velkých souborů dat. Analýza by neměla vycházet z matice vzdáleností, neboť tento postup je velmi náročný, a to jak výpočetně, tak z hlediska uložení matice. Kovarianční matice je sice matice podobností, ale podstata zůstává stejná. Problém spočívá v tom že při vyhledávání informací jsou obvykle řešeny současně dvě úlohy, a to redukce počtu proměnných (například pomocí jejich shlukování, čímž jsou místo jednotlivých slov používána témata) a shlukování dokumentů.

4 Přístupy k řešení problému vysoké dimenzionality

Vývoj v oblasti shlukování se zaměřuje především na soubory s velkým počtem objektů. Méně pozornosti je věnováno problematice velkého počtu proměnných. Berkhin uvádí, že shlukovací algoritmy založené na vzdálenostech fungují efektivně do 16 proměnných. Je potřeba si uvědomit, že počet dimenzí, s nimiž pracujeme, se reálně pohybují ve stovkách tisíc viz [2]. Soubory obsahující více než 16 proměnných nazývá Berkhin data s vysokou dimenzionalitou. V takových případech se používá *redukce dimenzionality*, která se realizuje buď transformací proměnných nebo doménovou dekompozicí.

K prvnímu uvedenému přístupu lze zařadit *analýzu hlavních komponent*, která ovšem může vést k vytvoření shluků s obtížnou interpretovatelností. V oblasti shlukování dokumentů je používána metoda SVD (singular value decomposition).

V případě *doménové dekompozice* jsou data rozdělena do podsouborů (anglicky *canopies*). Dimenzinalita se tedy neredukuje, ale tento postup vede ke snížení nákladů.

Pro shlukování objektů charakterizovaných velkým počtem proměnných lze použít metody založené na shlukování podprostorů (*subspace clustering*). Místo vytváření redukované matice založené na nových proměnných (získaných například lineární kombinací původních proměnných) je problém vel-

kého počtu dimenzí řešen zkoumáním podprostorů původního prostoru. Tento přístup je výhodný tím, že jsou zachovány původní proměnné, které mají reálný význam, zatímco lineární kombinace původních proměnných může být někdy těžko interpretovatelná.

Základem pro shlukování podprostorů je analýza hustoty objektů v prostoru. Cílem je nalezení podmnožin proměnných tak, aby projekce dat zahrnovaly regiony s vysokou hustotou. Podstatou je rozdělení všech dimenzí do stejného počtu stejně dlouhých intervalů. Jsou-li nalezeny vhodné podprostory, úloha spočívá v nalezení shluků v odpovídajících projekcích. Shluky jsou oblasti navazujících jednotek s vysokou hustotou (v rámci určitého podprostoru).

Podrobný popis těchto metod je uveden například v [3]. Základní metodou uváděnou v literatuře je algoritmus CLIQUE (CLustering In QUEst), který pro numerické proměnné navrhli v roce 1998 Agrawal a kolektiv. Tento shlukovací algoritmus využívá jak principů metod založených na hustotě, tak principů metod založených na mřížce.

Algoritmus ENCLUS (ENntropy-based CLUStering), navržený v r. 1999 Chengem a kolektivem, je založen podobném principu jako CLIQUE, avšak používá rozdílné kritérium pro výběr podprostorů. Výpočetní náklady této metody jsou ale vysoké.

Metoda MAFIA (Merging of Adaptive Finite Intervals (And more than a CLIQUE)) je modifikací algoritmu CLIQUE, která funguje rychleji a nalézá shluky lepší kvality. Prezentovali ji v roce 1999 Goil a kolektiv a v roce 2001 Nagesh a kolektiv. Metoda v každé dimenzi konstruuje tzv. adaptivní mřížky. Její paralelní verze se nazývá pMAFIA.

Kromě výše uvedených uvádí Berkhin ještě tři algoritmy, a to OptiGrid (navrhli v roce 1999 Hinneburg a Keim), PROCLUS (PROjected CLUStering), navržený v roce 1999 Aggarwalem a kolektivem, a ORCLUS (ORiented projected CLUSter generation), který v roce 2000 navrhli Aggarwal a Yu.

V poslední době se objevily práce, které využívají k redukci dimenziionality náhodné projekce viz [26]. Ukazuje se, že metoda náhodných projekcí umožňuje redukovat dimenzi podstatně efektivnějším způsobem než ostatní metody. Experimenty ukazují, že výsledky dosažené touto metodou jsou velmi slibné, viz [2]. Další možností je kombinace náhodných projekcí s metodou SVD [14].

5 Závěr

V příspěvku jsme zhodnotili některé možnosti využití shlukovacích metod pro vyhledávání v textových dokumentech. Využití shlukovacích metod je umožněno geometrizací vyhledávacího problému na základě vektorového modelu. Přestože tato problematika patří v oblasti vyhledávání mezi klasické, není v současné době uspokojivě vyřešena. Mezi základní problémy patří:

- návrh datové struktury pro indexování, viz [23],
- redukce dimenzionality a řešení "prokletí" dimenzionality [26],

- vyhledávání témat a jejich automatická detekce [7],
- modifikace míry podobnosti [24].

V dalším výzkumu bychom se chtěli zaměřit na rozsáhlé experimenty, s jejichž pomocí bychom chtěli vybrat vhodné míry podobnosti pro tvorbu shluků tak, aby tyto shluky odpovídaly tématům obsaženým v dané kolekci.

Reference

[1] Anghelescu A., Muchnik I. (2002). *Combinatorial clustering for textual data representation in machine learning models.* `http://mms-01.rutgers.edu/Documents/CombinatorialClustering/Theoretical.v01.pdf.`

[2] Bingham E., Mannila H. (2001). *Random projection in dimensionality reduction: applications to image and text data.* KDD, San Francisko.

[3] Berkhin P. *Survey of clustering data mining techniques.* Accrue Software, Inc., San Jose. `www.ee.ucr.edu/∼barth/EE242/clustering_survey.pdf`

[4] Berry M.W., Browne M. (1999). *Understanding search engines: mathematical modeling and text retrieval.* SIAM Book Series: Software, Environments, and Tools.

[5] Berry M.W. (editor). (2004). *Survey of text mining: clustering, classification and retrieval.* Springer-Verlag, New York.

[6] Dobrynin V., Patterson D., Rooney N. (2004). *Contextual document clustering.* ECIR 2004, LNCS 2997, Springer-Verlag, Berlin, 167 – 180.

[7] Dvorský J., Martinovič J., Pokorný J., Snášel V. (2004). *Vyhledávání témat v kolekci dokumentů*, Znalosti 2004, Brno, 317 – 326.

[8] Gordon A.D. (1999). *Classification*, 2nd Edition. Chapman & Hall/CRC, Boca Raton.

[9] Hotho A., Staab S., Maedche A. *Ontology-based text clustering.* `http://www-2.cs.cmu.edu/∼mccallum/textbeyond/papers/ hotho.pdf`

[10] Chakrabarti S. (2003). *Mining the web: discovering knowledge from hypertext data.* Morgan Kaufmann Publishers, San Fransisco.

[11] Ibrahimov O., Pashayev R. (2003). *Measuring similarities of textual documents: An overview of challenges and solutions.* TAINN 2003 (Turkish Symposium on Artificial Intelligence and Neural Networks). `http://www.ijci.org/product/tainn/E07033.pdf`

[12] Kang J., Ryu K.R., Kwon H. M. (2004). *Using cluster-based sampling to select initial training set for active learning in text classification.* PAKDD 2004, LNAI 3056, Springer-Verlag, Berlin, 384 – 388.

[13] Mercer D.P. (2003). *Clustering large datasets.* Linacre College, `http://www.stats.ox.ac.uk/∼mercer/documents/Transfer.pdf`

[14] Moravec P., Snášel V. (2004). *Rychlý přibližný výpočet LSI předzpracováním náhodnou projekcí.* Znalosti 2004, Brno, 166 – 177.

[15] Mylonas P., Wallace M., Kollias S. (2004). *Using k-nearest neighbor and feature selection as an improvement to hierarchical clustering.* SETN 2004, LNAI 3025, Springer-Verlag, Berlin, 191 – 200.

[16] Peltonen J., Sinkkonen J., Kaski S. (2002). *Discriminative clustering of text documents.* ICONIP 2002. IEEE **4**, 1956 – 1960.
`http://lib.hut.fi/Diss/2003/isbn9512267977/article8.pdf`

[17] Řezanková, H. (2004). *Klasifikace pomocí shlukové analýzy.* In: Kupka, K. (ed.) Analýza dat 2003/II. TriloByte Statistical Software, Pardubice, 119 – 135.

[18] Řezanková H., Húsek D., Smid J, Snášel V. (2003). *Clustering of documents via similarity measures.* In: D'Auriol, B. J. (ed.). CIC'03. CSREA Press, Las Vegas, 292 – 299.

[19] Řezanková H., Húsek D., Snášel V. (2003). *Applications of clustering methods to textual documents.* In: Bulletin of the International Statistical Institute Volume LX. International Statistical Institute, Berlin, 322 – 323.

[20] Rijsbergen C.J. (2004). *The geometry of information retrieval.* Cambridge University Press.

[21] Salton G., Buckley C. (1988). *Term weighting approaches in automatic text retrieval.* Information Processing and Management **24**, 5, 513 – 523.

[22] Sinkkonen J., Kaski S. (2000). *Clustering by similarity in an auxiliary space.* IDEAL 2000, Springer-Verlag, London, 3 – 8.
`http://lib.hut.fi/Diss/2003/isbn9512267977/article2.pdf`

[23] Skopal T., Moravec P., Pokorný J., Snášel V. (2004). *Metric indexing for the vector model in text retrieval.* SPIRE 04, Padova, Springer Verlag, 183 – 195.

[24] Skopal T., Moravec P., Pokorný J., Krátký M., Snášel V. (2003). *Efficient implementation of vector model in information retrieval.* In Proceedings of the fifth National Russian Research Conference, RCDL'2003, Digital Libraries: Advanced Methods and Technologies, Digital Collections, St. Petersburg, 170 – 179.

[25] Volk D., Stepanov M.G. (2001). *Resampling methods for document clustering.* `http://arxiv.org/PS_cache/cond-mat/pdf/0109/0109006.pdf`

[26] Vempala S.S. (2004). *The random projection method.* DIMACS. 103-111.

[27] Zhang Y., Zincir-Heywood N., Milios E. (2004). *Term-based clustering and summarization of web page collections.* Canadian AI 2004, LNAI 3060, Springer-Verlag, Berlin, 60 – 74.

Poděkování: Tento výzkum je součástí projektu COST 274 (TARSKI).

Adresa: D. Húsek, Ústav informatiky AV ČR, Pod Vodárenskou věží 2, 182 07 Praha 8; H. Řezanková, Vysoká škola ekonomická v Praze, nám. W. Churchilla 4, 130 67 Praha 3; V. Snášel, VŠB-TU Ostrava, 17.listopadu 15, 708 33 Ostrava-Poruba

E-mail: `dusan@cs.cas.cz, rezanka@vse.cz, vaclav.snasel@vsb.cz`

SLABÁ KONVERGENCE SUPREMA NÁHODNÝCH PROCESŮ

Jana Husová

Klíčová slova: Náhodné procesy, slabá konvergence.

Abstrakt: Uvažujeme posloupnost náhodných procesů $(X_n(t), t \in T)$, o které víme, že konverguje v distribuci k nějakému jinému procesu. A studujeme procesy $(Y_n(A), A \in \mathcal{A})$, kde $Y_n(A_1) := \sup_{t\in A_1} X_n(t)$. Zkoumáme, pro které kolekce množin $\mathcal{A} \subset \mathcal{P}(T)$ procesy $(Y_n(A), A \in \mathcal{A})$ konvergují v distribuci. Prvky procesů $(X_n(t), t \in T)$ uvažujeme jako funkce v $C(T)$, $l^{+\infty}(T)$, $D(T)$.

1 Úvod do problému

Předpokládejme, že víme, že posloupnost náhodných procesů $(X_n(t),\ t \in T)_{n\in\mathbf{N}_0}$, kde (T, ρ) je obecný kompaktní metrický prostor, konverguje slabě k nějakému náhodnému procesu X, $X_n \xrightarrow{\mathcal{D}} X$.

Položme

$$Y_n(A) := \sup_{t\in A} X_n(t).$$

Zajímá nás, pro jaké $\mathcal{A} \subset \mathcal{P}(T)$ posloupnost náhodných procesů $Y_n(A)$ konverguje slabě.

Postupně se podívejme, jak je to v různých případech, závislých na tom na jakém prostoru je původní konvergence.

2 Prostor $C(T)$

Uvažujme náhodné procesy $(X_n(t), t \in T, n \in \mathbf{N}_0)$ jako prvky $C(T)$, kde $C(T)$ je prostor reálných spojitých funkcí na T se supremální metrikou generovanou supremální normou, tj.

$$||x|| = \sup_{t\in T} |x(t)|.$$

Nejprve se tedy podívejme na to, v jakém prostoru je náš zkoumaný proces. Náhodné procesy X_n, X jsou spojité na kompaktu, tudíž jsou omezené. A tedy i náhodné procesy Y_n, Y jsou omezené, neboť

$$||y||^{\mathcal{A}} = \sup_{A\in\mathcal{A}} |y(A)| = \sup_{A\in\mathcal{A}} |\sup_{t\in A} |x(t)|| \le ||x|| \quad \forall x \in C(T).$$

Obecně tedy budeme uvažovat o prostoru $l^{+\infty}(\mathcal{A})$, prostoru reálných omezených funkcí na $\mathcal{A}$ se supremální metrikou generovanou supremální normou.

Zkoumáme zobrazení f z $C(T)$ do $l^{+\infty}(\mathcal{A})$ takové, že

$$f(x(t))(A) := \sup_{t\in A} x(t). \tag{1}$$

Tvrzení 2.1. *Funkce f z prostoru $(C(T), ||\cdot||)$ do prostoru $(l^{+\infty}(\mathcal{A}), ||\cdot||^{\mathcal{A}})$ definovaná vztahem (1) je spojitá.*

Důkaz: Potřebujeme ukázat, že vzor otevřené množiny je otevřený a to plyne z vlastnosti zobrazení vůči zvoleným metrikám. Platí totiž:

$$d(x,y) \geq \ d^{\mathcal{A}}(f(x), f(y)) \ \ \forall x, y \in C(T). \tag{2}$$

Z tohoto vztahu plyne spojitost zobraní (1).

Q.E.D.

Teď tedy již můžeme formulovat tvrzení.

Věta 2.1. *Nechť náhodné procesy $(X_n(t), t \in T)_{n\in\mathbf{N}_0}$ mají spojité trajektorie a navíc $X_n \xrightarrow{\mathcal{D}} X$ jako proces v $C(T)$. Potom pro každou kolekci $\mathcal{A} \subset \mathcal{P}(T)$ posloupnost procesů $(Y_n(A), A \in \mathcal{A})$, kde*

$$Y_n(A) := \sup_{t\in A} X_n(t),$$

konverguje k procesu $(Y(A), A \in \mathcal{A})$, kde $Y(A) = \sup_{t\in A} X(t)$, jako proces v prostoru $l^{+\infty}(\mathcal{A})$.

Důkaz: Protože slabá konvergence se zachovává spojitým zobrazením, viz např. [1], je věta důsledkem tvrzení 2.1.

Q.E.D.

3 Prostor $l^{+\infty}(T)$

Předpokládejme, obecněji, že procesy $X_n \in l^{+\infty}(T)$. Prostor uvažujeme se supremální metrikou. Předpokládejme tedy, že v prostoru omezených funkcí se supremální metrikou posloupnost náhodných procesů (X_n) konverguje slabě k náhodnému procesu X. Pro jaké kolekce množin $\mathcal{A} \subset \mathcal{P}(T)$ konverguje posloupnost náhodných procesů $(Y_n(A), A \in \mathcal{A})$ slabě k nějakému náhodnému procesu $(Y(A), A \in \mathcal{A})$?

Procesy $Y_n(A)$ budou opět prvky $l^{+\infty}(\mathcal{A})$. Prvky prostoru $l^{+\infty}(T)$ obecně nemusí být měřitelné, připomeňme tedy definici slabé konvergence pro obecně ne nutně měřitelné nety, viz např. [2].

Definice 3.1 (Slabá konvergence). *Net (X_λ) konverguje **slabě** k měřitelnému a těsnému zobrazení X v $(l^{+\infty}(T), ||\cdot||)$, zn. $X_\lambda \xrightarrow{\mathcal{D}} X$, tehdy a jen tehdy pokud*

$$\liminf_{\lambda\in\Lambda} Pr_*(X_\lambda \in G) \geq Pr(X \in G) \ \ \forall G \in \mathcal{G}(l^{+\infty}(T)).$$

Všimněme si, že stejně jako slabá konvergence pro měřitelná zobrazení i tato konvergence se zachovává spojitým zobrazením. Na tomto prostoru také platí vztah (2), zobrazení (1) je spojité a tedy slabá konvergence se zachovává.

A tedy věta 2.1 platí i pro procesy s trajektoriemi v prostoru $l^{+\infty}(T)$.

4 Prostor D(T)

Prostor $D(T)$ se Skorochodovovou metrikou je jiný případ, nemůžeme použít dosavadní výsledky, protože prostor $D(T)$ je vybaven Skorochodovou metrikou a ne supremální.

Opět víme, že náhodné procesy $(Y_n(A), A \in \mathcal{A})$ budou náležet do $l^{+\infty}(\mathcal{A})$. Tady už se ale konvergence obecně nezachovává, protože zobrazení (1) není spojité. V následujícím příkladě ukážeme, že se obecně nezachovává ani konvergence jednorozměrných marginálů.

Příklad: Mějme

$$\begin{aligned} X_{2n}(t) &= \mathbf{I}_{(1/2-1/n;1]}(t) \ \ \forall\omega \\ X_{2n+1}(t) &= X(t) = \mathbf{I}_{[1/2;1]}(t) \ \ \forall\omega \end{aligned}$$

Potom $X_n \xrightarrow{\mathcal{D}} X$ v $(D([0,1]), d)$ (prostor tzv. cadlag funkcí) (dokonce v pravděpodobnosti), ale $Y_{2n}([0,1/2)) = 1$ *s.j.* a $Y_{2n+1}([0,1/2)) = 0$ *s.j.*, a tedy $Y_n([0,1/2))$ nekonverguje.

△

Abychom mohli mluvit o konvergenci konečně rozměrných marginálů, potřebujeme pro obrazy funkcí z $D(T)$ najít také vhodný prostor s vhodnou metrikou. Tj. zobecněný Skorochodův prostor.

Předpokládejme tedy, že prostor $D(T)$ je zobecněný Skorochodův prostor (podle [3]). To znamená, že existuje $(\Lambda, ||\cdot||)$ prostor bijekcí na T s normou, která splňuje:

(i) $||\lambda|| \geq 0, \ ||\lambda|| = 0 \Leftrightarrow \lambda = e,$

(ii) $||\lambda \circ \mu|| \leq ||\lambda|| + ||\mu||,$

(iii) $||\lambda^{-1}|| = ||\lambda||.$

A také existuje kolekce $\mathcal{D}$ konečných dělení T usměrněná zjemněním a invariantní vůči Λ, tj.

$$\lambda \in \Lambda \text{ a } \Delta = \{A_i\} \in \mathcal{D} \ \Rightarrow \lambda(\Delta) = \{\lambda(A_i)\} \in \mathcal{D}.$$

A prostor $D(T)$ je prostor všech omezených reálných funkcí na T, které leží v uzávěru (vůči Skorochodově metrice) množiny všech jednoduchých funkcí na $\mathcal{D}$. Skorochodova vzdálenost dvou funkcí je

$$d(x,y) = \inf\{\varepsilon > 0 : \ \exists\lambda \in \Lambda : \ ||\lambda|| < \varepsilon \ \wedge \ ||x \circ \lambda - y|| < \varepsilon\}. \qquad (3)$$

Potřebujeme tedy najít také prostor bijekcí s normou na $\mathcal{A}$ a vhodnou kolekci dělení $\mathcal{D}^{\mathcal{A}}$ tak, aby byli splněny podmínky Skorochodova prostoru.

A navíc, aby obrazy funkcí z $D(T)$ patřily do $\mathcal{D}(\mathcal{A})$ a aby naše zobrazení také zachovávalo slabou konvergenci.

Uvažujme tedy následující prostor bijekcí $(\Lambda^{\mathcal{A}}, ||\cdot||^{\mathcal{A}})$, kde $\lambda^{\mathcal{A}} \in \Lambda^{\mathcal{A}}$:

$$\lambda^{\mathcal{A}} : \quad \mathcal{A} \to \mathcal{A} \tag{4}$$

$$A \mapsto \lambda(A). \tag{5}$$

Všimněme si, že tím předpokládáme, že $\mathcal{A} \subset \mathcal{D}(T)$ a navíc, že $\mathcal{A}$ je invariantní vůči λ. A kolekci $\mathcal{D}^{\mathcal{A}}$ obdobně

$$\mathcal{D}^{\mathcal{A}} := \{\mathcal{P}(\Delta) : \Delta \in \mathcal{D}\}. \tag{6}$$

Takto definovaná kolekce $\mathcal{D}^{\mathcal{A}}$ je také usměrněná vůči zjemnění a invariantní vůči $\Lambda^{\mathcal{A}}$.

Teď je otázka, jestli platí

$$x \in D(T) \ \Rightarrow f(x) \in D(\mathcal{A}),$$

kde f je definováno v (1).

To záleží na normách, které jsou na Λ a $\Lambda^{\mathcal{A}}$. Předpokládejme, že na Λ máme následující normy (obvyklé):

$$\begin{aligned} ||\lambda||_s &:= \sup_{t\in T} \rho(\lambda(t), t) \\ ||\lambda||_t &:= \sup_{t,s\in T:\rho(t,s)\neq 0} \left|\log \frac{\rho(\lambda(t),\lambda(s))}{\rho(t,s)}\right| \\ ||\lambda||_m &:= ||\lambda||_s + ||\lambda||_t. \end{aligned}$$

A normy na $\Lambda^{\mathcal{A}}$ indukujme stejným způsobem, tj. místo ρ bude $\rho^{\mathcal{A}}$, což bude označovat Hausdorfovu pseudometriku na $\mathcal{A}$. Tyto normy budeme opět označovat stejně, jen s indexem $\mathcal{A}$ nahoře.

Potom lze nahlédnout, díky tomu, že $||\lambda|| = ||\lambda^{-1}||$, že

$$||\lambda^{\mathcal{A}}||_s^{\mathcal{A}} = \sup_{A\in\mathcal{A}} \rho^{\mathcal{A}}(\lambda(A), A) \leq ||\lambda||_s.$$

A podobně

$$||\lambda^{\mathcal{A}}||_t^{\mathcal{A}} = \sup_{A,B\in\mathcal{A}:\rho^{\mathcal{A}}(A,B)\neq 0} \left|\log \frac{\rho^{\mathcal{A}}(\lambda(A),\lambda(B))}{\rho^{\mathcal{A}}(A,B)}\right| \leq ||\lambda||_t.$$

A tedy také

$$||\lambda^{\mathcal{A}}||_m^{\mathcal{A}} \leq ||\lambda||_m.$$

Díky těmto vztahům také platí: $x \in D(T) \ \Rightarrow f(x) \in D(\mathcal{A})$, pro prostory, které jsou indukovány normami $||\cdot||_s$, resp. $||\cdot||_t$, $||\cdot||_m$ a $||\cdot||_s^{\mathcal{A}}$, resp. $||\cdot||_t^{\mathcal{A}}$, $||\cdot||_m^{\mathcal{A}}$.

Tvrzení 4.1. *Nechť $D(T)$ je zobecněný Skorochodův prostor, který je indukovaný normovaným prostorem $(\Lambda, ||\cdot||)$ a kolekcí $\mathcal{D}$. A nechť $D(\mathcal{A})$ je také zobecněný Skorochodův prostor indukovaný normovaným prostorem $(\Lambda^{\mathcal{A}}, ||\cdot||^{\mathcal{A}})$ z (4) a kolekcí $\mathcal{D}^{\mathcal{A}}$ podle (6), kde $||\cdot||^{\mathcal{A}}$ splňuje:*

$$\forall x \in D(T) : ||x|| \geq ||f(x)||^{\mathcal{A}}, \text{ kde } f \text{ je definováno v (1).}$$

Potom pro všechny $\mathcal{A} \subset \mathcal{D}$, které jsou invariantní vůči λ, platí:

$$\{y : \exists x \in D(T) : y = f(x)\} \subset D(\mathcal{A}),$$

tj. zobrazení f z (1) je měřitelné zobrazení z $D(T)$ do $D(\mathcal{A})$.

Důkaz: Buď $x \in D(T)$ libovolné. Položme $y(A) = \sup_{t\in A} x(t)$. A chceme ukázat, že $y \in D(\mathcal{A})$. Víme, že $y \in l^{+\infty}(\mathcal{A})$ a potřebujeme, že $d^{\mathcal{A}}(y, \mathbf{I}_{\mathcal{D}^{\mathcal{A}}}) = 0$. Víme, že $d(x, \mathbf{I}_{\mathcal{D}}) = 0$, tj. existuje posloupnost $(\lambda_n \in \Lambda)$ a jednoduchých funkcí (x_n) taková, že $||\lambda_n|| \searrow 0$ a $||x - x_n \circ \lambda_n|| \searrow 0$. Položme $\lambda_n^{\mathcal{A}}$ obraz λ_n podle (4) a y_n obraz x_n podle (1). Potom určitě $||\lambda_n^{\mathcal{A}}||^{\mathcal{A}} \searrow 0$ a

$$\sup_{A\in\mathcal{A}} |y(A) - y_n(\lambda_n(A))| = \sup_{A\in\mathcal{A}} |\sup_{t\in A} x(t) - \sup_{t\in A} x_n(\lambda_n(t))| \searrow 0.$$

Tedy $y \in D(\mathcal{A})$.

Q.E.D.

Za předpokladů z tvrzení 4.1 lze snadno ukázat, že funkce (1) je spojitá, a tedy slabá konvergence zůstane zachována.

Tvrzení 4.2. *Za předpokladů tvrzení 4.1 je funkce f z (1) spojitá.*

Důkaz: K důkazu spojitosti stačí ukázat, že pro všechny nety (x_α) takové, $x_\alpha \to x$ platí, že $f(x_\alpha) \to f(x)$. My konkrétně ukážeme, že platí:

$$d(x, y) \geq d^{\mathcal{A}}(f(x), f(y)).$$

Předpokládejme tedy, že $d(x, y) \leq \varepsilon$, kde d je Skorochodova metrika, viz (3). To znamená, že existuje $\lambda \in \Lambda$ takové, že $||\lambda|| \leq d(x, y)$ a zároveň $||x\circ\lambda - y|| \leq d(x, y)$.

Potom také podle předchozího $||\lambda^{\mathcal{A}}||^{\mathcal{A}} \leq d(x, y)$, kde $\lambda^{\mathcal{A}}$ je indukována λ podle (4). A ukažme, že také $||f(x) \circ \lambda^{\mathcal{A}} - f(y)||^{\mathcal{A}} \leq d(x, y)$.

$$\begin{aligned}
\sup_{A\in\mathcal{A}} |f(x) \circ \lambda(A) - f(y)| &= \sup_{A\in\mathcal{A}} |\sup_{t\in\lambda(A)} x(t) - y(t)| = \\
&= \sup_{A\in\mathcal{A}} |\sup_{t\in A} x(\lambda(t)) - y(t)| \leq \\
&\leq \sup_{A\in\mathcal{A}} ||x \circ \lambda - y|| \leq d(x, y).
\end{aligned}$$

A tedy $d^{\mathcal{A}}(f(x), f(y)) \leq d(x, y)$ a zobrazení f je spojité.

Q.E.D.

Věta 4.1. *Nechť posloupnost procesů* (X_n) *konverguje slabě k procesu* X *v* $D(T)$*, potom posloupnost procesů* (Y_n) *indukovaných zobrazením (1) konverguje slabě k procesu* Y *v* $D(\mathcal{A})$*, kde* $D(\mathcal{A})$ *je z tvrzení 4.1 a* $\mathcal{A} \subset \mathcal{D}$*.*

Důkaz: Spojité zobrazení zachovává slabou konvergenci, a tak je věta důsledkem tvrzení 4.2.

Q.E.D.

Reference

[1] Billingsley P. (1968): *Convergence of probability measures.* Wiley, New York.

[2] van der Vaart A.W., Wellner J.A. (1996): *Weak convergence and empirical processes.* Springer-Verlag, New York.

[3] Straf M.L.(1970): *Weak convergence of stochastic processes with several parameters.* Proc. 6th Berkley Symposium math. Statist. Probab., Univ. Calif. 187- 221 (1972).

Poděkování: Příspěvek vznikl za podpory výzkumného záměru MŠMT: MSM 113200008 a GA ČR grantu 201/03/1027.

Adresa: J. Husová, KPMS MFF UK, Sokolovská 83, Praha 8 - Karlín

E-mail: husova@zf.jcu.cz

EXTRÉMY GAUSSOVSKÝCH POSLOUPNOSTÍ A PROCESŮ

Daniela Jarušková

Klíčová slova: Extrémy závislých a nezávislých gaussovských posloupností, asymptotická teorie, Gumbelovo rozdělení, procesy Ornsteinova-Uhlenbeckova typu, derivovatelné procesy, extrémy na pevném intervalu, střední počet překročení úrovně, Riceova věta, extrémy na rostoucích intervalech, maxima milánských teplotních odchylek, detekce bodu změny.

Abstrakt: Článek se zabývá asymptotickou teorií extrémů gausovských posloupností a procesů. Pomocí simulací studuje rychlost konvergence k limitnímu rozdělení. Teorie je aplikovaná na rozdělení testových statistik maximálního typu pro testování změny v posloupnosti náhodných veličin.

1 Úvod

Ačkoliv stochastická teorie extrémů není zdaleka nová, její popularita v posledních letech stoupá. Zájem o ni projevují odborníci v oboru meteorologie, hydrologie a klimatologie, kteří by ji rádi využili k odhadu výskytu extrémních událostí, jako jsou nezvykle vysoké či nízké teploty, povodně apod. Extrémní události zajímají i ekonomy a matematiky v pojišťovnách. Kromě toho se během posledních let objevilo i mnoho teoretických výsledků týkajících se chování extrémů v jednorozměrném i vícerozměrném případě. Mnoho nových výsledků bylo publikováno v časopise Extremes, který začal vycházet před šesti lety. Jednou za dva roky se také pořádá mezinárodní konference věnovaná právě teorii extrémů. Té poslední v portugalském Aveiru (2004) se zúčastnili čtyři statistici z České republiky.

Já jsem se začala zajímat o teorii extrémů náhodných procesů při studiu limitního chování testových statistik maximálního typu, které se používají při detekci bodu změny.

Ve svém příspěvku se chci zabývat asymptotickým chováním extrémů gaussovských posloupností $\{X_i\}$ a gaussovských procesů $\{X(t)\}$, přesněji chováním

$$\begin{aligned} &P(\max_{i=1,\dots,n} X_i > u), && \text{jestliže} && n \to \infty \quad \text{a} \quad u \to \infty, \\ &P(\max_{0\le t\le T} X(t) > u), && \text{jestliže} && u \to \infty \\ & && \text{nebo jestliže} && T \to \infty \quad \text{a} \quad u \to \infty. \end{aligned}$$

Přestože matematické myšlenky, postupy i získané výsledky teorie extrémů jsou velmi elegantní, sklízí teorie i řadu výtek. Nejzávažnější z nich spočívá v tom, že konvergence je velmi pomalá a rozumné aproximace lze

získat až pro velmi vysoké rozsahy výběrů n nebo pro příliš vysoké, a tudíž prakticky nezajímavé, hranice překročení u.

I když v poslední době vyšlo několik zajímavých knih týkajících se extrémů, viz např. [3], můj výklad se opírá o klasickou knihu [6]. Snažila jsem se přitom získat představu o tom, jak limitní věty asymptotické teorie extrémů fungují, to znamená pro jaké rozsahy a jak vysoké úrovně překročení mají pro použití smysl.

2 Extrémy gaussovských posloupností

2.1 Extrémy nezávislých standardně normálně rozdělených veličin

Nejprve uvažujme nejjednodušší možný případ, to znamená posloupnost nezávislých stejně $N(0,1)$ rozdělených náhodných veličin $\{e_i\}$. Zřejmě pro každé n platí

$$P\left(\max_{i=1,\dots,n} e_i < u\right) = \prod_{i=1}^{n} P(e_i < u) = (\Phi(u))^n\,,$$

kde $\Phi(\cdot)$ značí distribuční funkci standardního normálního rozdělení, zatímco $\phi(\cdot)$ bude dále značit jeho hustotu.

Pro libovolné $n \in N$ tak známe nejen distribuční funkci maxima, ale i všechny momenty (alespoň numericky), a tedy i např. střední hodnotu, rozptyl, šikmost, špičatost apod.

Studujme dále limitní chování $\max_{i=1,\dots,n} e_i$, když $n \to \infty$. Vzhledem k tomu, že pro všechna $u \in R_1$ platí $\Phi(u) < 1$, pak zřejmě pro všechna $u \in R_1$ platí

$$P\left(\max_{i=1,\dots,n} e_i < u\right) \to 0, \quad \text{když} \quad n \to \infty.$$

To znamená, že $\max_{i=1,\dots,n} e_i$ jde s počtem pozorování nade všechny meze. Naší snahou je proto najít posloupnost $\{u_n\}$ ($u_n \to \infty$ pro $n \to \infty$) takovou, aby $P(\max_{i=1,\dots,n} e_i < u_n)$ konvergovala k hodnotě z intervalu $(0,1)$. Příjemné by bylo, kdyby se podařilo najít posloupnost $\{u_n\}$ ve tvaru $u_n = b_n + a_n x$ tak, aby

$$P\left(\frac{\max_{i=1,\dots,n} e_i - b_n}{a_n} < x\right)$$

měla limitu v intervalu $(0,1)$ pro každé $x \in R_1$. Následující věta říká, že je možné takové posloupnosti $\{a_n\}$ a $\{b_n\}$ nalézt, a udává také, jak mohou vypadat.

Věta 2.1. *Nechť $\{e_i\}$ je posloupnost nezávislých $N(0,1)$ rozdělených náhodných veličin. Nechť dále $\{a_n\}$ a $\{b_n\}$ jsou posloupnosti reálných čísel splňujících*

$$b_n = \sqrt{2\log n} - \frac{\frac{1}{2}\log\log n + \frac{1}{2}\log 4\pi}{\sqrt{2\log n}} \quad \text{a} \quad a_n = \frac{1}{\sqrt{2\log n}}. \tag{1}$$

Pro každé $x \in R_1$ platí

$$P\left(\max_{i=1,\dots,n} e_i < b_n + a_n x\right) \to e^{-e^{-x}} \quad \text{pro} \quad n \to \infty.$$

Poznámka 2.1. Tabulka 1 udává konkrétní hodnoty a_n a b_n pro některá vybraná n. Je patrné, že s rostoucím n se hodnoty obou posloupností mění jen velmi pomalu.

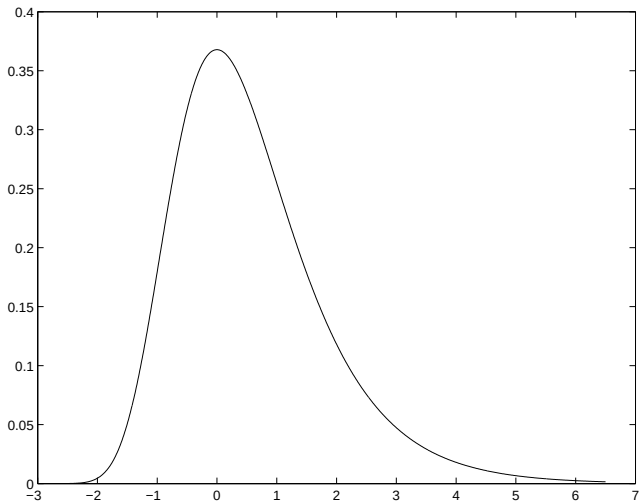
	$n = 100$	$n = 500$	$n = 1000$	$n = 5000$
b_n	2.3662	2.9074	3.1164	3.5611
a_n	0.3295	0.2836	0.2690	0.2423

Tabulka 1: Hodnoty konstant a_n a b_n pro několik vybraných hodnot n.

Poznámka 2.2. Věta 2.1 říká, že pro velká n je rozdělení maxima nezávislých standardních veličin přibližně Gumbelovo s parametrem posunutí b_n a parametrem měřítka a_n, to znamená, že platí

$$P\left(\max_{i=1,\dots,n} e_i < y\right) \doteq e^{-e^{-\frac{y-b_n}{a_n}}}.$$

Obrázek 1 ukazuje tvar hustoty Gumbelova rozdělení pro parametr posunutí rovný 0 a parametr měřítka rovný 1.



Obrázek 1: Hustota Gumbelova rozdělení.

	$n = 100$	$n = 500$	$n = 1000$	$n = 5000$
$\mathsf{E}\, Y_n$	2.556	3.071	3.272	3.701
$\mathsf{sd}\, Y_n$	0.423	0.364	0.345	0.311

Tabulka 2: Střední hodnota a směrodatná odchylka Gumbelova rozdělení, které aproximuje rozdělení maxima nezávislých $N(0,1)$ rozdělených veličin pro vybrané hodnoty n.

Pro náhodnou veličinu Y_n s Gumbelovým rozdělením s parametrem posunutí b_n a parametrem měřítka a_n platí:

$$\mathsf{E}\, Y_n \doteq b_n + 0.5722\, a_n, \quad \mathsf{sd}\, Y_n \doteq \sqrt{1.6449}\, a_n, \quad \text{šikmost}\, Y_n \doteq 1.2986.$$

Tabulka 2 udává střední hodnotu a směrodatnou odchylku pro různé hodnoty $n \in N$ a opět ilustruje pomalý vývoj těchto charakteristik.

Důkaz věty 2.1. Označme $u_n = b_n + a_n\, x$ pro $n \in N$. Zřejmě chceme dokázat

$$(\Phi(u_n))^n \to e^{-e^{-x}},$$

což je ekvivalentní s tím, že

$$n \log\left(1 - (1 - \Phi(u_n))\right) \to -e^{-x}. \tag{2}$$

Použijeme-li aproximaci logaritmu na okolí bodu 1 a známé aproximace pro pravděpodobnost překročení úrovně standardně normálně rozdělenou veličinou, tj. $1 - \Phi(x) \sim \phi(x)/x$ pro $x \to \infty$, je (2) ekvivalentní s tím, že

$$n\, (1 - \Phi(u_n)) \to e^{-x} \quad \Leftrightarrow \quad n\, \frac{\phi(u_n)}{u_n}\, e^x \to 1. \tag{3}$$

Po zlogaritmování obou stran docházíme k závěru, že chceme najít posloupnost $\{u_n\}$ ($u_n \to \infty$ pro $n \to \infty$) tak, aby splňovala

$$\log n + x - \log u_n - \tfrac{1}{2} \log 2\pi - \tfrac{u_n^2}{2} \to 0.$$

Odtud

$$u_n = \sqrt{2 \log n} + d_n, \quad \text{kde} \quad d_n = o(\sqrt{2 \log n}),$$

$$\log u_n = \log\left(\sqrt{2 \log n}\left(1 + \frac{d_n}{\sqrt{2 \log n}}\right)\right) \sim \log \sqrt{2 \log n} = \frac{1}{2} \log 2 + \frac{1}{2} \log \log n.$$

Posloupnost $\{d_n\}$ hledáme tedy tak, aby platilo

$$\log n + x - (1/2)\, \log 4\pi - (1/2)\, \log \log n - \left(\log n + d_n \sqrt{2 \log n} + d_n^2/2\right) \to 0.$$

Takovou posloupností je zřejmě

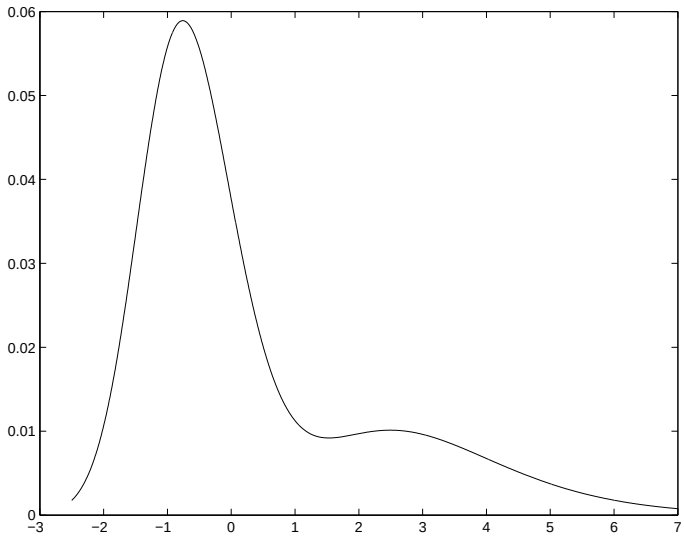
$$d_n = \frac{x - (1/2)\,\log 4\pi - (1/2)\,\log\log n}{\sqrt{2\log n}}.$$

□

Poznámka 2.3. Následující tvrzení ukazuje asymptotické chování rozdílu mezi skutečnou distribuční funkcí $\max_{i=1,\dots,n} e_i$ a distribuční funkcí Gumbelova rozdělení:

$$\left(P\left(\max_{i=1,\dots n} e_i < b_n + a_n\,x\right) - e^{-e^{-x}}\right) \sim \frac{e^{-e^{-x}}\,e^{-x}}{16}\,\frac{(\log\log n)^2}{\log n}.$$

Uvedená aproximace ilustruje pomalou konvergenci distribučních funkcí maxim. Pro odhad skutečného rozdílu mezi oběma distribučními funkcemi však nemá žádný velký význam. Na obrázku 2 vidíme rozdíl $(\Phi(b_n + a_n\,x))^n - e^{-e^{-x}}$ pro $n = 100$. Největší rozdíl je přibližně roven 0.06, přičemž pro různá x se rozdíly výrazně liší.



Obrázek 2: Rozdíl distribuční funkce maxima posloupnosti standardních normálních veličin o $n = 100$ členech počítané v argumentu $b_n + a_n\,x$ a distribuční funkce Gumbelova rozdělení.

2.2 Extrémy závislých normálně rozdělených náhodných posloupností

Nyní se zabývejme otázkou, jaké bude limitní rozdělení maxima posloupnosti standardně normálně rozdělených veličin, jestliže mezi těmito veličinami exis-

tuje závislost, která je charakterizována korelační maticí. Nejdůležitějším výsledkem, který zde uvedeme, je tvrzení, že za velmi obecných podmínek na autokorelační funkci má maximum z n členů stacionární gaussovské posloupnosti stejné limitní rozdělení jako maximum nezávislých normálních veličin. Tyto podmínky jsou splněny například pro stacionární gaussovské ARMA posloupnosti. Tvrzení o limitním chování maxima stacionární posloupnosti je jednoduchým důsledkem odhadu vzdálenosti mezi distribučními funkcemi maxim dvou stacionárních gaussovských vektorů s různými korelačními maticemi.

Předpokládejme, že vektor $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ má korelační matici $\boldsymbol{R}_1$ a distribuční funkci $F_1(x_1, \ldots, x_n)$, zatímco vektor $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ má korelační matici $\boldsymbol{R}_0$ a distribuční funkci $F_0(x_1, \ldots, x_n)$. Budeme hledat horní odhad rozdílu $F_1(x_1, \ldots, x_n) - F_0(x_1, \ldots, x_n)$, protože zřejmě platí

$$P\left(\max(X_1, \ldots, X_n) < u\right) - P\left(\max(Y_1, \ldots, Y_n) < u\right) = F_1(u, u, \ldots, u) - F_0(u, u, \ldots, u). \tag{4}$$

Pro každé $h \in [0, 1]$ definujme náhodný vektor $\boldsymbol{Z}_h = \sqrt{h}\,\boldsymbol{X} + \sqrt{1-h}\,\boldsymbol{Y}$. Vektor $\boldsymbol{Z}_h$ má standardně normálně rozdělené složky, přičemž jeho korelační matice $\boldsymbol{R}_h$ splňuje $\boldsymbol{R}_h = h\,\boldsymbol{R}_1 + (1-h)\,\boldsymbol{R}_0$. Označme $F_h(x_1, \ldots, x_n)$ distribuční funkci vektoru $\boldsymbol{Z}_h$ a $\phi_h(x_1, \ldots, x_n)$ jeho hustotu, tj.

$$F_h(x_1, \ldots, x_n) = \int_{\infty}^{x_n} \ldots \int_{\infty}^{x_1} \phi_h(y_1, \ldots, y_n)\, dy_1 \ldots dy_n.$$

Věta 2.2. *Platí*

$$\begin{aligned} &F_1(x_1, \ldots, x_n) - F_0(x_1, \ldots, x_n) \\ &= \sum_{i<j}\sum (r_{1ij} - r_{0ij}) \int_0^1 F_h(x_1, \ldots, x_n | Z_{hi} = x_i, Z_{hj} = x_j)\, \phi_{hij}(x_i, x_j)\, dh \\ &\le \sum_{i<j}\sum (r_{1ij} - r_{0ij}) \int_0^1 \phi_{hij}(x_i, x_j)\, dh, \end{aligned}$$

kde $\phi_{hij}(x_i, x_j)$ *je marginální hustota* $(i, j)-$*tých složek* Z_{hi} *a* Z_{hj} *vektoru* $\boldsymbol{Z}_h$.

Důkaz.

$$\begin{aligned} &F_1(x_1, \ldots, x_n) - F_0(x_1, \ldots, x_n) \\ &= \int_0^1 \frac{\partial F_h(x_1, \ldots, x_n)}{\partial h}\, dh = \int_0^1 \sum_{i<j}\sum \frac{\partial F_h(x_1, \ldots, x_n)}{\partial r_{ij}} \frac{\partial r_{ij}}{\partial h}\, dh \\ &= \sum_{i<j}\sum (r_{1ij} - r_{0ij}) \int_0^1 \int_{-\infty}^{x_n} \ldots \int_{-\infty}^{x_1} \frac{\partial \phi_h(y_1, \ldots, y_n)}{\partial r_{ij}}\, dy_1 \ldots dy_n\, dh \\ &= \sum_{i<j}\sum (r_{1ij} - r_{0ij}) \int_0^1 \int_{-\infty}^{x_n} \ldots \int_{-\infty}^{x_1} \frac{\partial^2 \phi_h(y_1, \ldots, y_n)}{\partial y_i\, \partial y_j}\, dy_1 \ldots dy_n\, dh \end{aligned}$$

$$
\begin{aligned}
&= \sum_{i<j}\sum (r_{1ij} - r_{0ij}) \cdot \\
&\qquad \int_0^1 \int_{-\infty}^{x_n} \dots \int_{-\infty}^{x_1} \phi_h(y_1, \dots, x_i, \dots, x_j, \dots, y_n) \frac{dy_1 \dots dy_n}{dy_i dy_j}\, dh \\
&= \sum_{i<j}\sum (r_{1ij} - r_{0ij}) \cdot \\
&\qquad \int_0^1 \int_{-\infty}^{x_n} \dots \int_{-\infty}^{x_1} \frac{\phi_h(\dots, x_i, \dots, x_j, \dots)}{\phi_{hij}(x_i, x_j)} \phi_{hij}(x_i, x_j) \frac{dy_1 \dots dy_n}{dy_i dy_j}\, dh \\
&= \sum_{i<j}\sum (r_{1ij} - r_{0ij}) \cdot \\
&\qquad \int_0^1 F_h(x_1, \dots, x_n | Z_{hi} = x_i, Z_{hj} = x_j) \phi_{hij}(x_i, x_j)\, dh \\
&\le \sum_{i<j}\sum (r_{1ij} - r_{0ij}) \cdot \int_0^1 \phi_{hij}(x_i, x_j)\, dh.
\end{aligned}
$$

V předchozím byla využita známá rovnost pro hustoty vícerozměrného standardního normálního rozdělení $\phi(x_1, \dots, x_n)$ s korelační maticí $\boldsymbol{R} = ||r_{ij}||$:

$$\frac{\partial \phi(x_1, \dots, x_n)}{\partial r_{ij}} = \frac{\partial^2 \phi(x_1, \dots, x_n)}{\partial x_i \partial x_j}.$$

□

Věta 2.3. *Nechť vektor* $\boldsymbol{X} = (X_1, \dots, X_n)^T$ *má korelační matici* $\boldsymbol{R}_1 = ||r_{1ij}||$ *a vektor* $\boldsymbol{Y} = (Y_1, \dots, Y_n)^T$ *má korelační matici* $\boldsymbol{R}_0 = ||r_{0ij}||$ *a nechť pro každou dvojici* (i, j) *platí* $r_{1ij} \ge r_{0ij}$. *Potom pro každé* $u \in R_1$ *platí*

$$P(\max\{X_1, \dots, X_n\} < u) \ge P(\max\{Y_1, \dots, Y_n\} < u).$$

Zjednodušeně by se předchozí tvrzení dalo říci slovy: *Posloupnosti s vyšší korelací mezi svými členy dosahují menších maxim.*

Důkaz.

$$
\begin{aligned}
&P(\max\{X_1, \dots, X_n\} < u) - P(\max\{Y_1, \dots, Y_n\} < u) \\
&\qquad = \sum_{i<j}\sum (r_{1ij} - r_{0ij})\, C_{ij}(\boldsymbol{R}_1, \boldsymbol{R}_0, u),
\end{aligned}
$$

kde $C_{ij}(\boldsymbol{R}, \boldsymbol{R}_0, u)$ jsou nezáporné konstanty, neboť jsou to integrály z nezáporné funkce (F_h je distribuční funkce a ϕ_h je hustota). □

Předchozí věta se dá zobecnit i na náhodné procesy.

Věta 2.4. *Nechť* $\{X(t), t \in [0,T]\}$ *a* $\{Y(t), t \in [0,T]\}$ *jsou standardizované gaussovské procesy s korelační funkcí* $R_1(s,t)$, *resp.* $R_2(s,t)$. *Nechť pro každé* $s < t$ *platí* $R_1(s,t) \geq R_2(s,t)$, *pak pro každé* $u \in R_1$

$$P\left(\max_{t\in[0,T]} X(t) < u\right) \geq P\left(\max_{t\in[0,T]} Y(t) < u\right).$$

Věta 2.5. *Pro shora uvažované vektory* $\boldsymbol{X}$ *a* $\boldsymbol{Y}$ *a* $u \in R_1$ *platí*

$$\begin{aligned} &P(\max\{X_1,\dots,X_n\} < u) - P(\max\{Y_1,\dots,Y_n\}) \\ &\leq \sum_{i<j}\sum |r_{1ij} - r_{0ij}| \int_0^1 \phi_{hij}(u,u)\,dh \\ &\leq \sum_{i<j}\sum |r_{1ij} - r_{0ij}| \,\frac{1}{2\pi}\,\frac{1}{\sqrt{1-\tilde{r}_{ij}^2}}\,\exp\left\{-\frac{u^2}{1+\tilde{r}_{ij}}\right\}, \end{aligned}$$

kde $\tilde{r}_{ij} = \max(|r_{1ij}|, |r_{0ij}|)$.

Věta 2.5 dává horní odhad pro vzdálenost distribučních funkcí maxim dvou vektorů standardních normálních veličin, které se liší korelací mezi svými členy. Pro praktické účely (jakmile u není opravdu velké) je uvedený odhad hrubý. V následující větě se odhaduje rozdíl mezi distribuční funkcí maxima stacionární závislé posloupnosti a distribuční funkcí maxima nezávislých veličin.

Věta 2.6. *Nechť* $X_1, X_2, \dots$ *je stacionární standardizovaná gaussovská posloupnost s korelační funkcí* $\{\rho(k)\}$, *kde* $|\rho(k)| < 1-\delta$ *pro všechna* $k = 1,\dots,$ *a* $Y_1, Y_2, \dots$ *je posloupnost nezávislých normálních náhodných veličin. Pak pro každé* $u \in R_1$ *a každé* $n \in N$ *platí*

$$\begin{aligned} &|P\left(\max\{X_1,\dots,X_n\} < u\right) - P\left(\max\{Y_1,\dots Y_n\} < u\right)| \\ &\leq n\sum_{i=1}^{n} |\rho(k)|\,\frac{1}{2\pi}\,\frac{1}{\sqrt{1-\rho(k)^2}}\,\exp\left\{-\frac{u^2}{1+|\rho(k)|}\right\} \\ &\leq C\,n\sum_{k=1}^{n} |\rho(k)|\,\exp\left\{-\frac{u^2}{2-\delta}\right\}. \end{aligned}$$

Z věty 2.6 plyne, že pro vysoké úrovně překročení, například $u_n \sim \sqrt{2\log n}$, je limitní chování pravděpodobnosti překročení pro stacionární posloupnosti s nepříliš pomalu klesající korelační funkcí stejné jako pro nezávislé veličiny.

Věta 2.7. *Nechť posloupnosti* $X_1, X_2, \dots$ *a* $Y_1, Y_2, \dots$ *jsou stejné jako ve větě 2.6 a* $\{u_n\}$ *je posloupnost reálných čísel taková, že* $u_n \sim \sqrt{2\log n}$ *(taková je např. posloupnost* $u_n = b_n + a_n x$ *z věty 2.1) a* $\sum_{k=1}^{\infty} |\rho(k)| < \infty$, *pak*

$$|P\left(\max\{X_1,\dots,X_n\} < u_n\right) - P\left(\max\{Y_1,\dots,Y_n\} < u_n\right)| \to 0 \quad \textit{pro} \quad n \to \infty. \tag{5}$$

Poznámka 2.4. Platí-li $\rho(n)\log n \to 0$ pro $n \to \infty$, pak (5) opět konverguje k 0.

Věta 2.8. *Nechť $\{X_i\}$ je stacionární gaussovská ARMA posloupnost. Nechť dále $\{a_n\}$ a $\{b_n\}$ jsou posloupnosti reálných čísel splňujících*

$$b_n = \sqrt{2\log n} - \frac{\frac{1}{2}\log\log n + \frac{1}{2}\log 4\pi}{\sqrt{2\log n}} \quad \text{a} \quad a_n = \frac{1}{\sqrt{2\log n}}.$$

Pro každé $x \in R_1$ platí

$$P\left(\max_{i=1,\dots,n} X_i < b_n + a_n x\right) \to e^{-e^{-x}} \quad \text{pro} \quad n \to \infty.$$

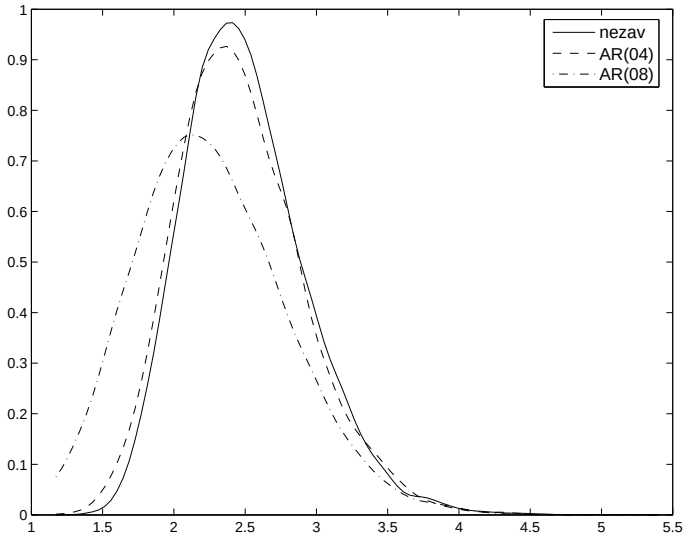
Předchozí věta říká, že se pro velká n maxima stacionárních gaussovských ARMA posloupností a maxima nezávislých posloupností chovají stejně. Podívejme se nyní konkrétněji na to, jak se liší rozdělení maxim pro AR posloupnosti a nezávislé posloupnosti. Tabulka 3 udává, jak se liší distribuční funkce $F(\cdot)$ maxima AR(0.4) posloupnosti (autoregresní posloupnosti 1. řádu s autoregresním koeficientem 0.4) od distribuční funkce maxima nezávislých $N(0,1)$ rozdělených veličin pro délku posloupnosti $n = 100$. Tabulka 4 udává základní momentové charakteristiky maxim obou posloupností. Poznamenejme, že všechny hodnoty byly získány simulacemi o rozsahu 100 000. Rozdíl mezi oběma distribučními funkcemi není příliš velký (největší je řádově roven 0.045) a také základní charakteristiky rozdělení se dobře shodují. Poměrně dobrou shodu potvrzuje i obrázek 3. Velice špatnou shodu však dostaneme, když uvažujeme rozdělení maxima posloupnosti AR(0.8) o délce $n = 100$, viz tabulka 5 a 6 a obrázek 3.

u	$F(u) - (\Phi(u))^{100}$	$(\Phi(u))^{100}$
1.5	0.00367	0.00099
2.0	0.03856	0.10013
2.14	0.04495	0.19574
2.5	0.02764	0.53639
3.0	0.00552	0.87365
3.5	0.00084	0.97700

Tabulka 3: Rozdíl mezi distribučními funkcemi maxima posloupnosti AR(0.4) a nezávisle rozdělených veličin pro $n = 100$.

	střední hodnota	směrodatná odch.	šikmost
AR(0.4)	2.4737	0.4467	0.5703
nezávislé	2.5081	0.4293	0.6601

Tabulka 4: Základní charakteristiky maxima posloupnosti AR(0.4) a nezávisle rozdělených veličin pro $n = 100$.



Obrázek 3: Hustota maxim nezávislých veličin (plná čára) a maxim autoregresní posloupnosti 1. řádu s autoregresním koeficientem 0.4 (čárkovaná čára), respektive 0.8 (čerchovaná čára) pro $n = 100$ získané simulací o rozsahu 10 000.

u	$F(u) - (\Phi(u))^{100}$	$(\Phi(u))^{100}$
1.5	0.07060	0.00099
2.0	0.24374	0.10013
2.11	0.25395	0.17234
2.5	0.16907	0.53639
3.0	0.04126	0.87365
3.5	0.00568	0.97700

Tabulka 5: Rozdíl mezi distribučními funkcemi maxima posloupnosti AR(0.8) a nezávisle rozdělených veličin pro $n = 100$.

	střední hodnota	směrodatná odch.	šikmost
AR(0.8)	2.2453	0.5351	0.3926
nezáv.	2.5081	0.4293	0.6601

Tabulka 6: Základní charakteristiky maxima posloupnosti AR(0.8) a nezávisle rozdělených veličin pro $n = 100$.

Situace se příliš nezlepší, ani když se délka posloupnosti prodlouží, např. na $n = 1000$, viz tabulka 7 a 8.

u	$F(u) - (\Phi(u))^{1000}$	$(\Phi(u))^{1000}$
2.0	0.00004	0.00000
2.5	0.03156	0.00197
3.0	0.15363	0.25903
3.5	0.04902	0.79243
4.0	0.00584	0.96882
4.5	0.00059	0.99661

Tabulka 7: Rozdíl mezi distribučními funkcemi maxima posloupnosti AR(0.8) a nezávisle rozdělených veličin pro $n = 1000$.

	střední hodnota	směrodatná odch.	šikmost
AR(0.8)	3.1225	0.3884	0.6617
nezáv.	3.2414	0.3509	0.7998

Tabulka 8: Základní charakteristiky maxima posloupnosti AR(0.8) a nezávisle rozdělených veličin pro $n = 1000$.

Z předchozího je zřejmé, že pokud nás zajímá rozdělení maxima posloupnosti délky sto až tisíc pozorování, dává aproximace Gumbelovým rozdělením alespoň přijatelné výsledky pro nezávislé či velmi slabě závislé posloupnosti. Jakmile však jsou veličiny silně kladně zkorelované (např. maximum autoregresní posloupnosti 1. řádu s autoregresním koeficientem 0.8, viz [7]), pak je rozdělení maxima takové posloupnosti výrazně odlišné od Gumbelova.

3 Gaussovské procesy

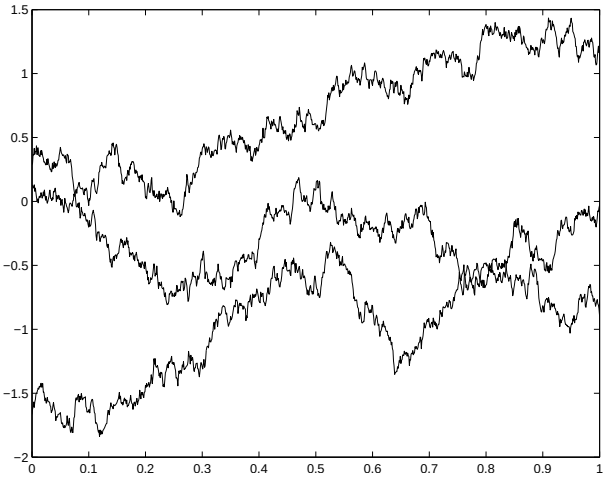
3.1 Stacionární gaussovské procesy

Náhodný proces $\{X(t)\, t \geq 0\}$ se nazývá gaussovský proces, jestliže pro každé $n \in N$ a každou posloupnost $0 \leq t_1 < \cdots < t_n$ má náhodný vektor $(X(t_1), \ldots, X(t_n))^T$ normální rozdělení. Nadále budeme uvažovat pouze standardizované procesy, tj. procesy, pro které platí $\mathsf{E}\,(X(t)) = 0$ a $\mathsf{E}\,(X(t))^2 = 1$. Gaussovský proces je stacionární, jestliže jeho korelační funkce $r(t,\, s) = \mathsf{E}\,(X(t)\, X(s))$ závisí pouze na rozdílu časových argumentů: $r(t, s) = r(t - s)$. Pro standardizovaný stacionární gaussovský proces platí, že veškeré jeho vlastnosti jsou určeny pouze jeho korelační funkcí $\{r(t),\, t \geq 0\}$, přičemž zřejmě $r(0) = 1$.

Při studiu extrémů gaussovských procesů se budeme zabývat pouze procesy se spojitou korelační funkcí, která má v bodě 0 jeden ze dvou následujících rozvojů:

$$r_1(t) = 1 - C_1\, t + o\,(t) \quad \text{pro} \quad t \to 0, \tag{6}$$

$$r_2(t) = 1 - \frac{1}{2} C_2\, t^2 + o\,(t^2) \quad \text{pro} \quad t \to 0. \tag{7}$$



Obrázek 4: Trajektorie procesu s korelační funkcí $r(t) = \exp(-t/2)$.

3.2 Procesy Ornsteinova – Uhlenbeckova typu

Procesy, pro které korelační funkce splňuje (6), se někdy nazývají procesy Ornsteinova – Uhlenbeckova typu. Nejjednodušším takovým procesem je pochopitelně sám Ornsteinův – Uhlenbeckův proces, tj. proces s korelační funkcí $r(t) = e^{-C\,t}$. Na obrázku 4 vidíme trajektorie takovéhoto procesu.
Ornsteinův – Uhlenbeckův proces často slouží jako model rychlosti částice vykonávající Brownův pohyb. Ornsteinův – Uhlenbeckův proces může vzniknout jako limita spojitých po částech lineárních procesů vytvořených ze stacionární autoregresní posloupnosti 1. řádu

$$Y_{k+1} = a\,Y_k + e_{k+1}, \quad \text{kde} \quad 0 < a < 1,$$

s autokorelační funkcí $r(k) = a^k, k = 0, 1, \ldots$. Předpokládejme, že autoregresní koeficient a je funkcí parametru Δ a že platí:

$$a(\Delta) = 1 - \alpha\,\Delta + o\,(\Delta) \quad \text{pro} \quad \Delta \to 0.$$

Definujme spojitý po částech lineární gaussovský proces $\{Y_\Delta(t),\, t \geq 0\}$ tak, aby $Y_\Delta(t) = Y_k$ pro $t = k\,\Delta$, $k = 0, 1, \ldots$
Pro korelační funkce $r_\Delta(s, s+t)$ procesu $\{Y_\Delta(t)\}$ platí

$$r_\Delta(s, s+t) \sim \mathsf{corr}\,(Y_\Delta(t),\, Y_\Delta(0)) \sim a^{t/\Delta} \sim (1 - \alpha\,\Delta)^{t/\Delta} \sim e^{-\alpha\,t},$$

a tedy $\{Y_\Delta(t)\}$ konverguje (v distribuci) pro $\Delta \to 0$ k Ornsteinově – Uhlenbeckově procesu s korelační funkcí $r(t) = e^{-\alpha\,t}$.

3.3 Derivovatelné procesy

Proces $\{X(t),\, t \geq 0\}$ je derivovatelný v L_2, jestliže existuje proces s konečným rozptylem $\{\dot{X}(t),\, t \geq 0\}$ takový, aby

$$\mathsf{E}\left(\frac{X(t+h)-X(t)}{h} - \dot{X}(t)\right)^2 \to 0 \quad \text{pro} \quad h \to 0.$$

Proces $\{\dot{X}(t)\}$ se nazývá derivací procesu $\{X(t)\}$. Dá se ukázat, že pokud je proces $\{X(t)\}$ standardizovaný stacionární gaussovský, pak i $\{\dot{X}(t)\}$ je stacionární gaussovský proces s nulovou střední hodnotou. Vzhledem k tomu, že platí

$$\mathsf{E}\left(\frac{X(t+h)-X(t)}{h}\right)^2 \to \mathsf{Var}\left(\dot{X}(t)\right) \quad \text{pro} \quad h \to 0,$$

musí korelační funkce $\{r(t)\}$ procesu $\{X(t)\}$ splňovat

$$\frac{2-2\,r(h)}{h^2} \to \mathsf{Var}\left(\dot{X}(0)\right) \quad (\equiv \dot{\sigma}_x^2) \quad \text{pro} \quad h \to 0,$$

a tedy mít rozvoj v bodě 0 typu (7):

$$r(h) = 1 - \frac{\dot{\sigma}_x^2}{2}\,h^2 + o\,(h^2) \quad \text{pro} \quad h \to 0. \tag{8}$$

Navíc platí

$$\mathsf{corr}\left(X(t), \dot{X}(t)\right) = 0$$

a mezi korelačními funkcemi $\{\dot{r}(t)\}$ procesu $\{\dot{X}(t)\}$ a $\{r(t)\}$ procesu $\{X(t)\}$je vztah $\dot{r}(t) = -r''(t)$. Platí-li shora uvedené podmínky, dá se ukázat, že existuje verze procesu s derivovatelnými trajektoriemi.

Ornsteinův – Uhlenbeckův proces není derivovatelný. Uveďme tedy několik příkladů procesů, které derivovatelné jsou:

1. Funkce kosinus s náhodnou amplitudou a posunutím:

$$X(t) = A\cos(\omega_0 t) + B\sin(\omega_0 t) = C\cos(\omega_0 t + \phi).$$

Předpokládáme-li, že veličiny A a B jsou nezávislé standardně normálně rozdělené a frekvence ω_0 je fixní a nenáhodná, pak $\{X(t)\}$ je standardní gaussovský proces. Snadno lze ukázat, že za těchto předpokladů má posunutí ϕ rovnoměrné rozdělení $R(0,\, 2\pi)$ a amplituda C rozdělení s hustotou

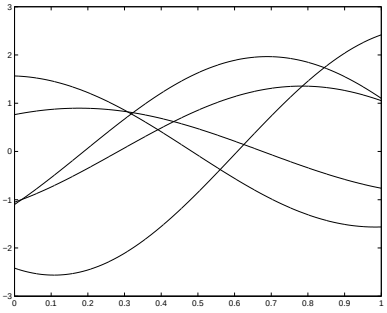
$$f(x) = xe^{-x^2/2} \quad \text{pro} \quad x \geq 0, \quad f(x) = 0 \quad \text{pro} \quad x < 0.$$

Korelační funkce

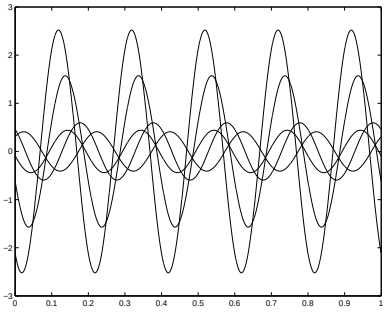
$$r(t) = \cos(\omega_0 t)$$

má rozvoj v bodě 0:

$$r(t) = 1 - \frac{\omega_0^2}{2}t^2 + o\,(t^2) \quad \text{pro} \quad t \to 0.$$



Obrázek 5: Trajektorie „kosinového procesu“ s $\omega_0 = \pi$.



Obrázek 6: Trajektorie „kosinového procesu“ s $\omega_0 = 10\pi$.

Na obrázku 5 vidíme trajektorie „kosinového procesu“ s $\omega_0 = \pi$ a na obrázku 6 s $\omega_0 = 10\pi$.

2. Gaussovský proces se spektrální hustotou $h(\omega)$ s $\int_{-\infty}^{\infty} \omega^2\, h(\omega)\, d\omega < \infty$. Připomeňme, že korelační funkce je Fourierovou transformací spektrální hustoty, a tedy platí

$$r(t) = \int_{-\infty}^{\infty} e^{i\omega t} h(\omega)\, d\omega \quad \text{a} \quad h(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \cos(\omega t) r(t)\, dt.$$

Inženýři si intuitivně představují takový proces jako „nekonečnou směs kosinových procesů“ s náhodnými amplitudami a posunutími:

$$X(t) = \sum_{\omega} A_\omega \cos(\omega t) + B_\omega \sin(\omega t).$$

My zde budeme uvažovat pouze případ, kde spektrální hustota $h(\omega)$ má jednoduchý tvar:

$$\begin{aligned} h(\omega) &= \frac{1}{4a} \quad \text{pro} \quad \omega \in [-\omega_0 - a, -\omega_0 + a] \cup [\omega_0 - a, \omega_0 + a], \\ &= 0 \qquad \text{jinde,} \end{aligned}$$

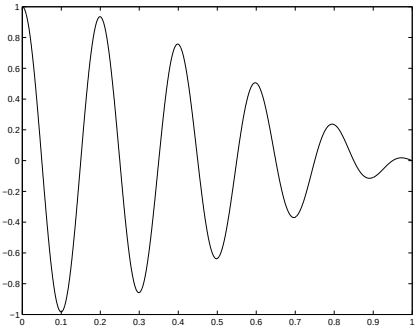
kde ω_0, a jsou nějaké známé konstanty. Korelační funkce procesu $\{X(t)\}$

$$r(t) = (\cos(\omega_0 t))\, \frac{\sin(at)}{at} \tag{9}$$

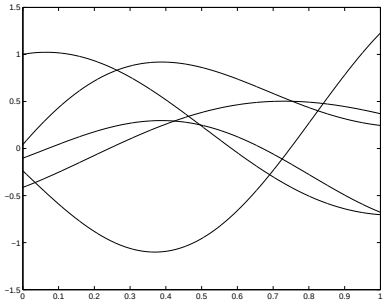
má rozvoj v bodě 0:

$$r(t) = 1 - \frac{1}{2} \left(\omega_0^2 + \frac{a^2}{3} \right) t^2 + o\left(t^2\right) \quad \text{pro} \quad t \to 0.$$

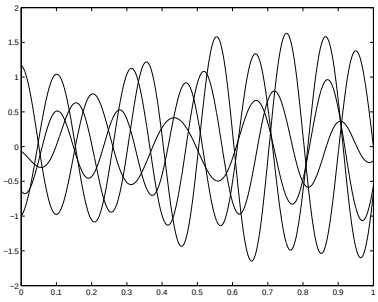
Obrázek 7 ukazuje příklad korelační funkce (9) pro $\omega_0 = 10\,\pi$ a $a = \pi$.



Obrázek 7: Autokorelační funkce (9).



Obrázek 8: Trajektorie procesu s korel. fcí (9), kde $\omega_0 = \pi$ a $a = \pi/3$.



Obrázek 9: Trajektorie procesu s korel. fcí (9), kde $\omega_0 = 10\,\pi$ a $a = \pi$.

Obrázek 8 ukazuje trajektorie procesu s korelační funkcí (9), kde $\omega_0 = \pi$ a $a = \pi/3$, a obrázek 9 trajektorie procesu, kde $\omega_0 = 10\,\pi$ a $a = \pi$.

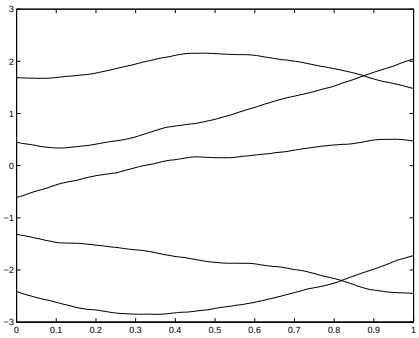
3. Gaussovský proces s korelační funkcí

$$r(t) = \frac{3}{2}\exp\left(-\sqrt{\frac{2}{3}}\,t\right) - \frac{1}{2}\exp\left(-3\sqrt{\frac{2}{3}}\,t\right). \tag{10}$$

Korelační funkce (10) má rozvoj v bodě 0:

$$r(t) = 1 - t^2 + o\,(t^2) \quad \text{pro} \quad t \to 0.$$

Rozptyl procesu derivace $\mathsf{Var}\left(\dot{X}(t)\right) = 2$. Proces nemá druhou derivaci, protože $r'''(0) \neq 0$. Obrázek 10 ukazuje jeho trajektorie. Proces vzniká při studiu testové statistiky v problémech vzniku spojitého lineárního trendu v posloupnosti dat, viz sekce 4.2.



Obrázek 10: Trajektorie procesu s korelační funkcí (10).

3.4 Maximum derivovatelného procesu na pevném intervalu

Nechť $\{X(t)\}$ je standardizovaný gaussovský proces s derivací $\{\dot{X}(t)\}$. Označme $\mathsf{Var}\left(\dot{X}(t)\right) = \dot{\sigma}_x^2$. Pro derivovatelný proces je velmi důležitý pojem překročení úrovně trajektorií. (Funkce f překročí v bodě t_0 úroveň u směrem vzhůru, jestliže existuje $\epsilon > 0$ takové, že pro všechna $t \in [t_0 - \epsilon, t_0]$ platí $f(t) \le u$ a pro všechna $t \in [t_0, t_0 + \epsilon]$ platí $f(t) \ge u$.) Pro střední počet překročení úrovně u směrem vzhůru $\mathsf{E}\left(N_u[0,T]\right)$ na intervalu $[0,T]$ totiž existuje analytické vyjádření, viz věta 3.2. Ze stacionarity procesu plyne

$$\mathsf{E}\left(N_u[0,T]\right) = T\,\mathsf{E}\left(N_u[0,1]\right).$$

Pro nás je však důležité, že pomocí středního počtu překročení lze omezit pravděpodobnost překročení úrovně.

Věta 3.1. *Pro standardizovaný derivovatelný gaussovský proces* $\{X(t)\}$ *platí:*

$$\begin{aligned} P\left(\max_{0\le t\le T} X(t) > u\right) &\le (1-\Phi(u)) + \mathsf{E}\left(N_u[0,T]\right) \\ &= (1-\Phi(u)) + T\,\mathsf{E}\left(N_u[0,1]\right). \end{aligned}$$

Jestliže $r(t) < 1$ *pro všechna* $t \neq 0$*, pak*

$$P\left(\max_{0\le t\le T} X(t) > u\right) \sim (1-\Phi(u)) + T\,\mathsf{E}\left(N_u[0,1]\right) \quad \textit{pro} \quad u \to \infty. \tag{11}$$

Vzhledem k tomu, že pravděpodobnost $(1-\Phi(u))$ *je asymptoticky zanedbatelná vzhledem ke střednímu počtu překročení úrovně, platí také*

$$P\left(\max_{0\le t\le T} X(t) > u\right) \sim T\,\mathsf{E}\left(N_u[0,1]\right) \quad \textit{pro} \quad u \to \infty. \tag{12}$$

Důkaz.

$$\begin{aligned} P\left(\max_{0\le t\le T} X(t) \ge u\right) &= P\left(X(0) \ge u\right) + P\left(X(0) < u \cap N_u[0,T] \ge 1\right) \\ &\le P\left(X(0) \ge u\right) + P\left(N_u[0,T] \ge 1\right) \\ &\le P\left(X(0) \ge u\right) + \mathsf{E}\left(N_u[0,T]\right). \end{aligned}$$

Poslední nerovnost plyne z toho, že

$$P\left(N_u[0,T] \ge 1\right) = \sum_{i=1}^{\infty} P\left(N_u[0,T] = i\right),$$

$$\mathsf{E}\left(N_u[0,T]\right) = \sum_{i=1}^{\infty} i\, P\left(N_u[0,T] = i\right).$$

Důkaz limitního chování (11) je poměrně složitý, zájemci ho mohou najít např. v [6]. □

Všimněme si toho, že horní odhad v (11) je tím přesnější, čím menší je pravděpodobnost, že proces překročí úroveň směrem vzhůru více než jednou. Pro „pomalu se pohybující procesy“ bude horní odhad těsnější než pro „rychlejší procesy“. Horní odhad je také tím těsnější, čím vyšší hladinu překročení uvažujeme. Dříve než budeme zkoumat těsnost horní meze v (11), ukažme, čemu se rovná střední počet překročení úrovně.

Věta 3.2. *(Riceova věta)*
Pro standardizovaný stacionární derivovatelný gaussovský proces $\{X(t)\}$ *platí*

$$\mathsf{E}\left(N_u[0,1]\right) = \dot{\sigma}_x \frac{1}{2\pi} \exp\left(-\frac{u^2}{2}\right), \tag{13}$$

kde $\dot{\sigma}_x$ *je směrodatná odchylka procesu derivace* $\dot{X}(t)\}$.

Pokud proces $\{X(t)\}$ *má střední hodnotu 0 a směrodatnou odchylku* σ_x, *pak*

$$\mathsf{E}\left(N_u[0,1]\right) = \frac{\dot{\sigma}_x}{\sigma_x} \frac{1}{2\pi} \exp\left(-\frac{u^2}{2\sigma_x^2}\right). \tag{14}$$

Důkaz provedeme ve dvou krocích.

Věta 3.3. *Pro stacionární derivovatelný proces platí*

$$\mathsf{E}\left(N_u[0,1]\right) = \lim_{q\to 0} \frac{P\left(X(0) < u \cap X(q) > u\right)}{q}.$$

Důkaz. Pro $n = 1, \ldots$ zaveďme procesy $\{X_n(t),\, t \in [0,1]\}$, které mají po částech lineární spojité trajektorií, přičemž

$$X_n\left(i/2^n\right) = X\left(i/2^n\right), \quad i = 0, \ldots, 2^n.$$

Označme $N_n(u)$ počet překročení úrovně u směrem vzhůru procesem $\{X_n(t)\}$. Zřejmě $N_n(u)$ konverguje monotónně k $N_u[0,1]$ pro $n \to \infty$. Z Leviho věty plyne, že $\mathsf{E}N_n(u) \to \mathsf{E}N_u[0,1]$. Vzhledem k tomu, že proces $\{X_n(t)\}$ může překročit úroveň u v každém z intervalů $[(i-1)/2^n, i/2^n]$ nejvýše jednou, plyne:

$$\begin{aligned}\mathsf{E}N_n(u) &= \sum_{i=1}^{2^n} P\left(X((i-1)/2^n < u \cap X(i/2^n) > u\right) \\ &= 2^n P\left(X(0) < u \cap X(1/2^n) > u\right).\end{aligned}$$

Existuje-li limita $P\left(X(0) < u \cap X(q) > u\right)/q$, pak k ní poslední výraz konverguje. □

Věta 3.4. *Pro stacionární derivovatelný proces platí*

$$\begin{aligned}\lim_{q\to 0} \frac{P\left(X(0) < u \cap X(q) > u\right)}{q} &= \int_0^\infty w\, f_{(X(0),\dot{X}(0))}(u,w)\, dw \\ &= \int_0^\infty w\, f_{\dot{X}(0)}(w\,|X(0)=u)\, f_{X(0)}(u)\, dw.\end{aligned}$$

Důkaz.

$$\begin{aligned}&\frac{1}{q} P\left(X(0) < u \cap X(q) > u\right) \\ &= \frac{1}{q}\int_{-\infty}^{u} P\left(\frac{X(q)-X(0)}{q} > \frac{u-v}{q} \,\Big|\, X(0)=v\right) f_{X(0)}(v)\, dv \\ &= \frac{1}{q}\int_{-\infty}^{u}\int_{(u-v)/q}^{\infty} f_{(X(0),(X(q)-X(0))/q)}(v,w)\, dw\, dv \\ &= \frac{1}{q}\int_{0}^{\infty}\int_{u-wq}^{u} f_{(X(0),(X(q)-X(0))/q)}(v,w)\, dv\, dw.\end{aligned}$$

Obrázek 11 ilustruje platnost poslední rovnosti. Dále pokud $q \to 0$, pak poslední dvojný integrál konverguje k integrálu

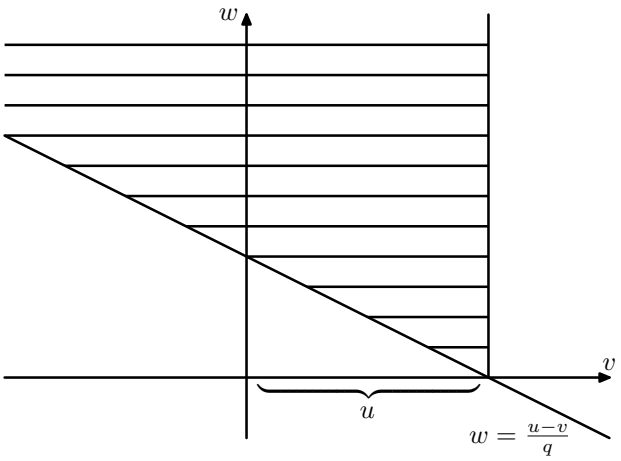
$$\int_0^\infty w\, f_{(X(0),\dot{X}(0))}(u,w)\, dw = \int_0^\infty w\, f_{\dot{X}(0)}(w\,|X(0)=u)\, f_{X(0)}(u)\, dw.$$

□

Důkaz věty 3.2. Veličiny $X(0)$ a $\dot{X}(0)$ jsou nezávislé normálně rozdělené, proto

$$\int_0^\infty w\, f_{\dot{X}(0)}(w\,|X(0)=u)\, dw = \int_0^\infty w\, f_{\dot{X}(0)}(w)\, dw = \frac{\dot{\sigma}_x}{\sqrt{2\pi}}.$$

□



Obrázek 11: Obor hodnot dvojného integrálu ve větě 3.4.

Poznamenejme, že v předchozích tvrzeních jsou třeba jisté předpoklady na sdruženou hustotu vektoru $(X(0), (X(q) - X(0))/q)$, které zabezpečují existenci potřebných limit a záměnu limity a integrálu. Pro normálně rozdělené veličiny jsou tyto předpoklady splněné.

Důsledek věty 3.1 a věty 3.2.
Pro standardizovaný stacionární derivovatelný gaussovský proces $\{X(t)\}$ platí:

$$P\left(\max_{0\le t\le T} X(t) > u\right) \sim T\,\frac{\dot{\sigma}_x}{\sqrt{2\pi}}\,\phi(u) \quad \text{pro} \quad u \to \infty. \tag{15}$$

Položme si otázku, jak lze Riceovu větu použít ve statistické analýze pro odhad středního počtu přechodu úrovní směrem vzhůru, a odtud popřípadě pro odhad pravděpodobnosti přechodu vysokých úrovní. Rozptyl procesu derivace totiž obvykle neznáme a jeho odhad je často nespolehlivý. Můžeme však použít vztah (14). Po zlogaritmování získáme lineární vztah mezi kvadrátem úrovně překročení u^2 (nezávislou proměnnou) a logaritmem středního počtu překročení úrovně (závisle proměnnou) $\log\left(\mathsf{E}(N_u[0,1])\right)$:

$$\log\left(\mathsf{E}(N_u[0,1])\right) = c + d\,u^2.$$

Z údajů o počtu překročení nižších úrovní můžeme odhadnout regresní technikou parametry c a d a získaný vztah použít pro odhad počtu překročení vysoké úrovně u, která nás zajímá.

Nyní si ukažme na našich příkladech, jak těsně odhaduje horní hranice (11) pravděpodobnosti překročení meze u směrem vzhůru na intervalu $[0,1]$. Nejprve se zabývejme případem 1, tj. „kosinovým procesem“:

$$X(t) = A\cos(\omega_0 t) + B\sin(\omega_0 t) = C\cos(\omega_0 t + \phi).$$

Uvažujme případ, kdy $\omega_0 \leq \pi$. V takové situaci je trajektorie procesu v bodě 0 buď větší než u, což je ekvivalentní s tím, že trajektorie už úroveň u během časového intervalu $[0,1]$ nepřekročí, nebo je v bodě 0 menší než u, což je ekvivalentní s tím, že překročí úroveň u právě jednou. Odtud plyne, že se nerovnost (11) změní v rovnost:

$$P\left(\max_{0\leq t\leq 1} X(t) > u\right) = (1-\Phi(u)) + \mathsf{E}\left(N_u[0,1]\right).$$

Naopak, jestliže $\omega_0 \geq 2\pi$, pak

$$P\left(\max_{0\leq t\leq 1} X(t) > u\right) = P(C > u) = e^{-u^2/2}.$$

Proces $\{X(t)\}$ v případu 2 s korelační funkcí $r(t) = (\cos(\omega_0 t))\,\frac{\sin(at)}{at}$ je do jisté míry „podobný“ procesu z případu 1, a to tím více, čím menší je hodnota a. Není tedy divu, že pro $\omega_0 \leq \pi$ a malé a bude horní hranice v (11) těsná, zatímco pro $\omega_0 > 2\pi$ bude velmi volná, s výjimkou pro vysoké hranice překročení, viz tabulky 9 a 10.

Nakonec se ještě podívejme na kvalitu horního odhadu (11) pro proces v případu 3 s korelační funkcí (10). Tabulka 11 ukazuje, že i zde je horní mez poměrně těsná.

Poznamenejme, že odhady pravděpodobností překročení byly provedeny ze simulací.

	$1-\Phi(u)$	$\mathsf{E}N_u$	$1-\Phi(u)+\mathsf{E}N_u$	$P(\max X(t) > u)$
$u=1$	0.1587	0.3197	0.4784	0.4675
$u=2$	0.0228	0.0713	0.0941	0.0871
$u=3$	0.0013	0.0059	0.0072	0.0077

Tabulka 9: Porovnání pravděpodobnosti překročení s horní mezí (11) pro proces s korelační funkcí (9), kde $\omega_0 = \pi$ a $a = \pi/3$.

	$1-\Phi(u)$	$\mathsf{E}N_u$	$1-\Phi(u)+\mathsf{E}N_u$	$P(\max X(t) > u)$
$u=1$	0.1587	3.0730	3.2317	0.8731
$u=2$	0.0228	0.6773	0.7001	0.2920
$u=3$	0.0013	0.0556	0.0569	0.0301

Tabulka 10: Porovnání pravděpodobnosti překročení s horní mezí (11) pro proces s korelační funkcí (9), kde $\omega_0 = 10\pi$ a $a = \pi$.

	$1-\Phi(u)$	$\mathsf{E}N_u$	$1-\Phi(u)+\mathsf{E}N_u$	$P(\max X(t) > u)$
$u=0$	0.5000	0.2254	0.7254	0.6734
$u=1$	0.1587	0.1365	0.2952	0.2737
$u=2$	0.0228	0.0305	0.0533	0.0514

Tabulka 11: Porovnání pravděpodobnosti překročení s horní mezí (11) pro proces s korelační funkcí (10).

3.5 Maximum procesu Ornsteinova – Uhlenbeckova typu na pevném intervalu

Věta 3.5. *Pro standardizované stacionární gaussovské procesy Ornsteinova – Uhlenbeckova typu, tj. pro procesy, jejichž korelační funkce má v bodě 0 rozvoj* (6) *platí*

$$P\left(\max_{0\le t\le T} X(t) > u\right) \sim T\,C\,u\,\phi(u) \quad \textit{pro} \quad u \to \infty, \tag{16}$$

kde konstanta C je konstanta vyskytující se v rozvoji (6).

Důkaz. Důkaz je poměrně obtížný. Zájemci ho mohou opět najít v knize [6]. □

Zajímavé je srovnání (15) a (16). V obou případech lze pravděpodobnost překročení vysoké úrovně (chvost rozdělení maxima) aproximovat výrazem $T\,\lambda(u)$, kde $\lambda(u)$ je funkcí hranice překročení u. Pro derivovatelné procesy platí, že $\lambda(u) = C_d\,\phi(u)$, zatímco pro procesy Ornsteinova – Uhlenbeckova typu $\lambda(u) = C_{nd}\,u\,\phi(u)$ (konstanty $C_d = \sqrt{C_2/(2\pi)}$ a $C_{nd} = C_1$ z rozvojů (6) a (7)).

3.6 Maximum stacionárních procesů na rozšiřujícím se intervalu

V této kapitole se budeme studovat chování

$$P\left(\max_{0\le t\le T} X(t) > u_T\right)$$

pro případ, kdy $T \to \infty$ a úroveň překročení $u_T \to \infty$, ovšem v závislosti na T. Opět se budeme zabývám pouze procesy, jejichž korelační funkce splňuje buď (6) nebo (7). Navíc budeme předpokládat, že $r(t)\,\log(t) \to 0$ pro $t \to \infty$. Je zřejmé, že limitní chování bude odlišné pro procesy Ornsteinova – Uhlenbeckova typu a pro derivovatelné procesy.

Věta 3.6. *Pro standardizovaný stacionární derivovatelný proces $\{X(t)\}$ platí*

$$P\left(\max_{0\le t\le T} X(t) > b_T + a_T\,x\right) \to 1 - e^{-e^{-x}} \quad \textit{pro} \quad T \to \infty, \tag{17}$$

kde

$$b_T = \sqrt{2\log T} + \frac{1}{\sqrt{2\log T}}\log\frac{\dot\sigma_x}{2\pi} \quad \text{a} \quad a_T = \frac{1}{\sqrt{2\log T}}. \tag{18}$$

Věta 3.7. *Pro standardizovaný stacionární proces* $\{X(t)\}$ *Ornsteinova-Uhlenbeckova typu platí*

$$P\left(\max_{0\le t\le T} X(t) > b_T + a_T\, x\right) \to 1 - e^{-e^{-x}} \quad \textit{pro} \quad T \to \infty, \tag{19}$$

kde

$$b_T = \sqrt{2\log T} + \frac{1}{\sqrt{2\log T}}\left(\frac{1}{2}\log\log T + \log\left(\frac{C}{\sqrt{\pi}}\right)\right) \quad \text{a} \quad a_T = \frac{1}{\sqrt{2\log T}}. \tag{20}$$

Náznak důkazu:

Uvažujme $\lambda(u)$ ze vztahů (12) a (16) a volme u v závislosti na T tak, aby $\lambda(u)\,T \to \tau$, jestliže $T \to \infty$ a současně $u \to \infty$. Pro vysoké úrovně u zřejmě platí

$$P\left(\max_{0\le t\le 1} X(t) < u\right) \sim (1 - \lambda(u))\,.$$

Kdyby maxima na po sobě jdoucích intervalech o délce 1 byla nezávislá, pak by platilo

$$P\left(\max_{0\le t\le T} X(t) < u\right) \sim (1 - \lambda(u))^T \sim e^{-T\,\lambda(u)} \sim e^{-\,\tau}.$$

Ve skutečnosti bohužel maxima na po sobě jdoucích intervalech nejsou nezávislá, a proto se musí postupovat při důkazu opatrněji, a to tak, že se mezi po sobě jdoucími intervaly vynechávají mezery, které na jedné straně zeslabují závislost mezi maximy a na druhé straně neovlivňují výslednou limitní pravděpodobnost.

Naším přáním je nyní najít úroveň překročení ve tvaru $u_T = b_T + a_T x$ tak, aby $T\,\lambda(u_T) \sim e^{-x}$. Postup při odvození konstant a_T a b_T a limitního rozdělení je stejný jako ve větě 2.1.

Z vět 3.6 a 3.7 vyplývá, že v obou případech má maximum gaussovského procesu na velmi dlouhých intervalech opět přibližně Gumbelovo rozdělení. Rozdíl mezi oběma rozděleními je však v parametru měřítka a parametru posunutí.

4 Aplikace

Poté, co jsem se seznámila s teorií extrémů gaussovských procesů jsem byla na pochybách, jak teorii použít ve statistice. Představme si, že měříme ve velmi krátkých časových okamžicích po velmi dlouhou dobu hodnoty gaussovského procesu. Na naměřenou řadu se můžeme dívat jako na realizaci gaussovské posloupnosti. Jsou-li však intervaly mezi po sobě jdoucími měřeními velmi krátké, bude korelace mezi odpovídajícími údaji blízká jedné. Jak bylo již řečeno dříve, aproximace rozdělení maxima pomocí Gumbelova rozdělení s parametry (1) nebude asi příliš dobrá. Na druhé straně se také můžeme dívat na

data jako na hodnoty gaussovského procesu měřeného v časech $t = 1, \ldots, T$. Jak ale v tom případě zjistit chování korelační funkce na okolí bodu 0? Navíc může existovat mnoho trajektorií, pro které $[X(1) < u] \cap \cdots \cap [X(T) < u]$, ale $\max_{0 \le t \le T} X(t) > u$. Odtud plyne například i neshoda mezi konstantami a_n a b_n pro maximum $AR(1)$ posloupnosti a konstantami a_T a b_T (při $T = n$) pro maximum Ornsteinova – Uhlenbeckova procesu. Pozorovaný proces také nemusí být dokonale normální. Zdá se, že nejrozumněji lze použít teorii extrémů tak, že se díky obecnému principu modeluje rozdělení maxim Gumbelovým rozdělením s parametry odhadnutými z naměřených dat. Aproximace Gumbelovým rozdělením je tím lepší, čím vyšší úrovně překročení nás zajímají.

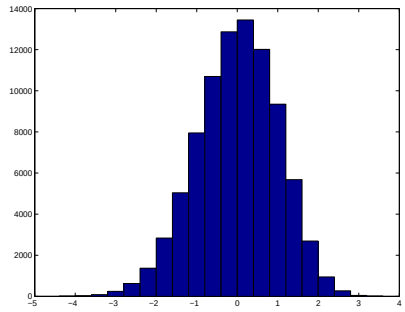
4.1 Odchylky milánské teplotní řady

Řada denních průměrných teplot měřených v Miláně je příkladem velmi dlouhé řady vzniklé přirozeným způsobem, nikoliv simulací. Bohužel díky změnám teploty během roku se nejedná o řadu stacionární. Řadu jsme stacionarizovali tak, že jsme odečetli průměr a vydělili směrodatnou odchylkou spočtených z hodnot v daném kalendářním dni. To znamená, že například od údajů pro 1. leden jsme odečetli průměr spočtený ze všech teplot v prvních lednech a rozdíly vydělili směrodatnou odchylkou prvních lednů, podobně od údajů z 2. ledna jsme odečetli průměry z druhých lednů a podělili směrodatnou odchylkou druhých lednů atd. Tímto způsobem jsme získali řadu standardizovaných odchylek řady od průměrné teploty pro příslušný kalendářní den. Velké hodnoty tedy znamenají neobvykle vysoké teploty pro daný kalendářní den v roce.

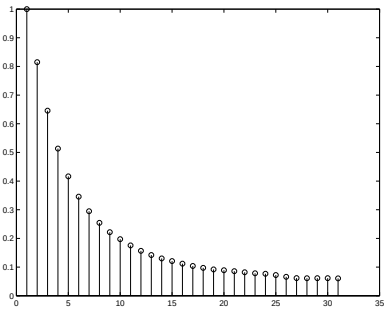
Studovaná řada obsahuje 86 000 údajů. Na obrázku 12 vidíme její histogram. Je zřejmé, že rozdělení je mírně záporně sešikmené (šikmost=-0.2195), takže není dokonale normální, jak bychom si přáli. Na obrázku 13 vidíme autokorelační funkci. Zdá se, že vhodným modelem by mohla být autoregresní posloupnost 1. řádu s autoregresním koeficientem 0.81. Shoda empirické autokorelační funkce s teoretickou autokorelační funkcí je až zarážejícím způsobem dobrá pro posunutí (lag) menší než 10. Pro větší hodnoty posunutí klesá empirická autokorelační funkce pomaleji než by odpovídalo autoregresní posloupnosti 1. řádu. To může být způsobeno pomalým stochastickým kolísáním řady, tedy jejím „slabě nestacionárním chováním“. Je známo, že se delší (někdy i několikaletá) období s vyšší teplotou střídají s obdobími s nižší teplotou. Také kladný trend v průměrné roční teplotě pozorovaný během posledních let se může v autokorelační funkci projevit podobným způsobem.

Řadu o délce 86 000 údajů jsme rozdělili na 86 úseků po 1000 datech a z každého úseku jsme spočítali maximum. Q-Q plot pro Gumbelovo rozdělení je na obrázku 14.

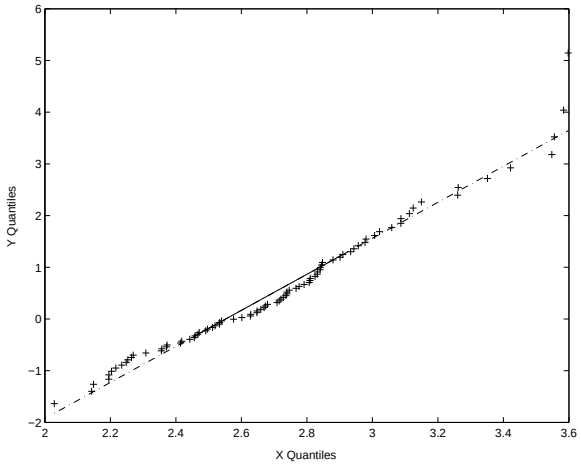
Metodou maximální věrohodnosti jsme odhadli parametr měřítka a a parametr posunutí b Gumbelova rozdělení $\hat{a} = 0.2893$ a $\hat{b} = 2.5377$, které jsou výrazně odlišné od a_n a b_n pro $n = 1000$ z tabulky 1. Tabulka 12 ukazuje



Obrázek 12: Histogram milánských teplotních odchylek.



Obrázek 13: Autokorelační fce milánských teplotních odchylek.



Obrázek 14: Q-Q plot pro maxima milánských teplotních odchylek proti Gumbelovu rozdělení.

	empiricky	Gumbel
$u = 3$	0.1860	0.1832
$u = 3.2$	0.0930	0.0964
$u = 3.4$	0.0581	0.0495

Tabulka 12: Pravděpodobnosti překročení úrovně u odhadované empiricky a pomocí Gumbelova rozdělení.

pravděpodobnosti překročení úrovně $P(\max X_i > u)$ odhadované jednak neparametricky, z empirické distribuční funkce, a pak pomocí Gumbelova rozdělení s parametry $\hat{a}$ a $\hat{b}$.

Příklad ilustruje, že navzdory tomu, že konvergence rozdělení extrémů stacionární posloupnosti ke Gumbelovu rozdělení je pomalá (zvláště pro řady se silnou kladnou korelací mezi blízkými členy), je pro dostatečně dlouhé řady a vysoké úrovně překročení shoda mezi relativní četností překročení a odpovídající pravděpodobností spočtenou z Gumbelova rozdělení vcelku dobrá.

4.2 Aplikace teorie extrémů na problémy detekce bodu změny

Nakonec uveďme ještě dvě aplikace teorie extrémů pro *detekci bodu změny.* Problémy bodu změny se zabývaly již dva příspěvky na minulých Robustech, viz [5] a [1], budeme proto struční.

Náhlá změna střední hodnoty

Uvažujme situaci, která se objevuje ve statistické kontrole jakosti. Pokud je výrobní proces „pod kontrolou“, kolísá měřená charakteristika kvality kolem určité známé hodnoty se známým rozptylem. Statistik by řekl, že v časových okamžicích $t = 1, 2, \ldots, n$ pozoruje nezávislé náhodné veličiny se známou střední hodnotou a známým rozptylem. Je zřejmé, že místo původních veličin můžeme uvažovat jejich standardizované verze $\{X_i\}$, kde $\mathsf{E}X_i = 0$ a $\mathsf{Var}X_i = 1$ pro $i = 1, \ldots, n$. Navíc budeme předpokládat, že všechna $\{X_i\}$ mají normální rozdělení.

Díky poruše ve výrobním procesu se však může stochastické chování pozorovaných veličin změnit. Jedním z typů změn, které mohou nastat, je posun střední hodnoty při zachování stejného rozptylu. To znamená, že počínaje nějakým neznámým časovým okamžikem není již střední hodnota pozorovaných veličin nulová, ale je rovna nějakému $\mu \neq 0$. Předpokádejme, že víme předem, že $\mu > 0$. V matematické statistice se rozhodnutí, zda ke změně došlo, obvykle provádí na základě výsledku testování hypotéz. Testujeme nulovou hypotézu H proti alternativě A_1:

$$\begin{aligned} H &: X_i = e_i, && i = 1, \ldots, n, \\ A_1 &: \exists\, k \in \{0, \ldots, n-1\} \quad \text{takové, že} \\ &\quad X_i = e_i, && i = 1, \ldots, k, \\ &\quad X_i = \mu + e_i, && i = k+1, \ldots, n, \end{aligned} \tag{21}$$

kde $\{e_i\}$ jsou nezávislé $N(0,1)$ rozdělené náhodné veličiny a $\mu > 0$ je neznámá konstanta. Protože v čase n známe všechny hodnoty pozorovaných veličin můžeme (kvůli zjednodušení matematických formulí) změnit jejich pořadí na pořadí uvažované od konce $X_i^{nov} = X_{n-i+1}^{star}$ (dále budeme zjednodušeně psát místo X_i^{nov} pouze X_i) a místo alternativy A_1 testovat alternativu A_1':

$$\begin{aligned} A_1' &: \exists\, k \in \{1, \ldots, n\} \quad \text{takové, že} \\ &\quad X_i = \mu + e_i, && i = 1, \ldots, k, \\ &\quad X_i = e_i, && i = k+1, \ldots, n. \end{aligned}$$

Kdybychom věděli, že pokud ke změně dojde, pak může nastat jedině v čase k, použili bychom k testování odhad μ metodou nejmenších čtverců založený pouze na prvních k pozorování, tj $\hat{\mu}_k = \left(\sum_{i=1}^k X_i\right)/k$, nebo ještě lépe jeho standardizovanou verzi $\left(\sum_{i=1}^k X_i\right)/\sqrt{k}$. Nulovou hypotézu bychom zamítli, jestliže by tato testová statistika překročila kritickou hodnotu. Pokud bod změny k neznáme, počítáme statistiku pro všechna možná k a nulovou hypotézu zamítneme, jestliže alespoň jedna ze statistik $\{\left(\sum_{i=1}^k X_i\right)/\sqrt{k},\ k = 1,\dots,n\}$ nabude velké hodnoty, nebo jinak řečeno, jestliže

$$\max_{1\le k\le n}\left\{\frac{1}{\sqrt{k}}\sum_{i=1}^{k} X_i\right\} \tag{22}$$

nabude velké hodnoty. Kromě statistiky (22) také používáme její useknutou verzi:

$$\max_{[\beta n]\le k\le n}\left\{\frac{1}{\sqrt{k}}\sum_{i=1}^{k} X_i\right\}, \tag{23}$$

kde β je nějaké malé kladné číslo.

Nulovou hypotézu zamítáme, jestliže statistika (22), resp. statistika (23), nabude tak velké hodnoty, že překročí kritickou hodnotu odpovídající požadované hladině významnosti. Je tedy zapotřebí znát rozdělení statistik (22), resp. (23), za platnosti nulové hypotézy. Obě dvě statistiky jsou maximem posloupnosti standardizovaných normálních veličin s korelací:

$$\mathsf{corr}\left(\frac{\sum_{i=1}^k X_i}{\sqrt{k}}, \frac{\sum_{i=1}^m X_i}{\sqrt{m}}\right) = \sqrt{k/m} \quad \text{pro} \quad k \le m. \tag{24}$$

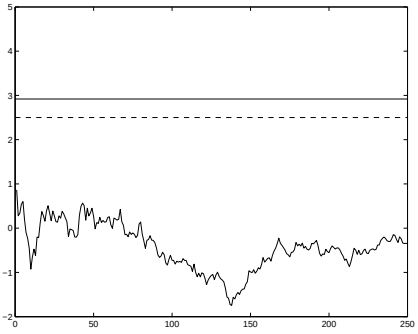
Použít přesné rozdělení statistik (22) a (23) za platnosti H_0 k nalezení kritických hodnot je velmi obtížné. Pro velký počet dat n lze použít přibližné kritické hodnoty založené na limitním chováním (22) a (23). Definujme náhodné procesy $\{U_n(t),\ t \in [1/n, 1]\}$ se spojitými po částech lineárními trajektoriemi, pro které

$$U_n(k/n) = \frac{\sum_{i=1}^k X_i}{\sqrt{k}} \quad \text{pro} \quad k = 1,\dots,n,$$

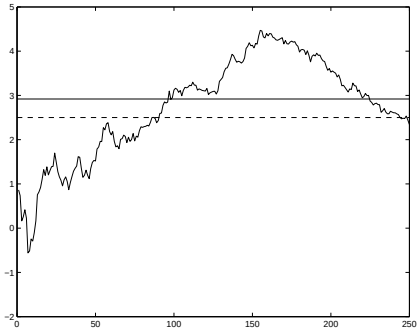
a tedy, viz (24),

$$\mathsf{corr}(U_n(t)\, U_n(s)) = \sqrt{\frac{[nt]/n}{[ns]/n}} + O(1/n) \quad \text{pro} \quad t \le s., \tag{25}$$

a zajímejme se o rozdělení maxima procesu $\{U_n(t)\}$ na intervalu $[1/n, 1]$, resp. na intervalu $[\beta, 1]$, neboť



Obrázek 15: Proces $\{U_n(t)\}$ pro $n = 250$ za platnosti H_0 a 5% kritické hodnoty spočtené podle (27) s $\beta = 0.1$ a (28).



Obrázek 16: Proces $\{U_n(t)\}$ pro $n = 250$ za platnosti A_1', kde $k = 140$ a $\mu = 0.25$ a 5% kritické hodnoty spočtené podle (27) s $\beta = 0.1$ a (28).

$$\max_{1/n\le t\le 1} U_n(t) = \max_{1\le k\le n}\left\{\frac{1}{\sqrt{k}}\sum_{i=1}^{k} X_i\right\},$$

$$\max_{\beta\le t\le 1} U_n(t) = \max_{[\beta\, n]\le k\le n}\left\{\frac{1}{\sqrt{k}}\sum_{i=1}^{k} X_i\right\}.$$

Na obrázku 15 vidíme trajektorii procesu $\{U_n(t)\}$ za platnosti nulové hypotézy H a na obrázku 16 trajektorii téhož procesu za platnosti alternativy A_1' s $k = 140$, $\mu = 0.25$.

Maximum procesu $\{U_n(t)\}$ je možno aproximovat maximem limitního procesu $\{U(t)\}$, který bude mít vzhledem k (25) korelační funkci

$$\mathsf{corr}\,(U(t), U(s)) = \sqrt{t/s} \quad \text{pro} \quad t \le s.$$

Použijeme-li transformaci času, pak

$$\max_{1/n\le t\le 1} U(t) = \max_{0\le\tau\le\log n} U\left(\exp(-s)\right), \ \max_{\beta\le t\le 1} U(t) = \max_{0\le\tau\le\log 1/\beta} U\left(\exp(-s)\right). \tag{26}$$

Proces $\{U\left(\exp(-\tau)\right), \tau \ge 0\}$ je Ornsteinův – Uhlenbeckův proces s korelační funkcí $r(\tau) = \exp(-\tau/2)$. Maxima (26) jsou pak maxima tohoto procesu na pevném a na rozšiřujícím se intervalu, proto z (16) a (19) plyne

$$P\left(\max_{[\beta\, n]\le k\le n} X_i > x\right) \approx (1/2)\, x\, \phi(x) \log(1/\beta), \tag{27}$$

$$P\left(\max_{1\le k\le n} X_i > x\right) \approx 1 - \exp\left(-e^{-\frac{x-b_n}{a_n}}\right), \tag{28}$$

kde

$$b_n = \sqrt{2\log\log n} + \frac{1}{\sqrt{2\log\log n}}\left(\frac{1}{2}\log\log\log n + \log\left(\frac{1}{2\sqrt{\pi}}\right)\right),$$
$$a_n = \frac{1}{\sqrt{2\log\log n}}.$$

Postupná lineární změna střední hodnoty

Kromě náhlého skoku ve střední hodnotě se můžeme též snažit odhalit pomalu vznikající kladný lineární trend. Pomocí matematické statistiky můžeme naše rozhodnutí, zda k takovéto změně v neznámém čase došlo, založit opět na testování hypotéz. Jestliže stejně jako v minulém příkladu pracujeme s veličinami $X_i^{nov} = X_{n-i+1}^{star}$ (opět pro zjednodušení budeme značit X_i^{nov} pouze X_i), budeme testovat nulovou hypotézu H proti alternativě A_2:

$$\begin{aligned} &H : X_i = e_i, && i = 1,\dots,n, \\ &A_2 : \exists\, k \in \{1,\dots,n\} \quad \text{takové, že} \\ &\qquad X_i = b\cdot(k-i+1) + e_i, && i = 1,\dots,k, \\ &\qquad X_i = e_i, && i = k+1,\dots,n, \end{aligned} \tag{29}$$

kde $b > 0$ je neznámá konstanta a $\{e_i\}$ jsou nezávislé $N(0,1)$ rozdělené náhodné veličiny. Nechť $\hat{b}_k$ označuje odhad regresního koeficientu b v čase k V_k jeho standardizovaná verze

$$\hat{b}_k = \frac{\sum_{i=1}^{k}(k-i+1)\,X_i}{\sum_{i=1}^{k}(k-i+1)^2}, \quad V_k = \frac{\sum_{i=1}^{k}(k-i+1)\,X_i}{\sqrt{\sum_{i=1}^{k}(k-i+1)^2}}.$$

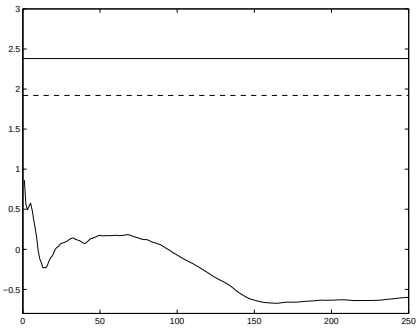
Testová statistika pro testování problému (29) má tvar $\max_{1\le k\le n} V_k$ nebo $\max_{[\beta\, n]\le k\le n} V_k$, přičemž mezi členy posloupnosti $\{V_k,\, k = 1,\dots,n\}$ je korelace:

$$\mathsf{corr}(V_k, V_m) = \frac{\sum_{i=1}^{\min(k,m)}(k-i+1)(m-i+1)}{\sqrt{\sum_{i=1}^{k}(k-i+1)^2}\sqrt{\sum_{i=1}^{m}(m-i+1)^2}}.$$

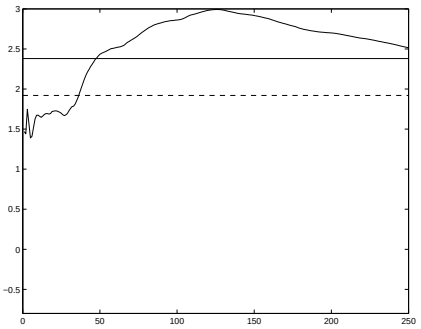
Definujeme-li podobným způsobem jako v předchozím případě spojitý po částech lineární proces $\{V_n(t),\, t \in (0,1]\}$, pak

$$\max_{1\le k\le n} V_k = \max_{1/n\le t\le 1} V_n(t)\,, \quad \max_{[\beta\, n]\le k\le n} V_k = \max_{\beta\le t\le 1} V_n(t).$$

Na obrázku 17 vidíme trajektorii procesu $\{V_n(t)\}$ za platnosti nulové hypotézy H a na obrázku 18 trajektorii téhož procesu za platnosti alternativy A_2 s $k = 140$, $b = 0.25$.
Korelační funkce procesu $\{V_n(t)\}$ pro $t < s$:



Obrázek 17: Proces $\{V_n(t)\}$ pro $n = 250$ za platnosti H_0 a 5% kritické hodnoty spočtené podle (30) s $\beta = 0.1$ a (31).



Obrázek 18: Proces $\{V_n(t)\}$ pro $n = 250$ za platnosti A_2, kde $k = 140$ a $b = 0.25$ a 5% kritické hodnoty spočtené podle (30) s $\beta = 0.1$ a (31).

$$\mathsf{corr}(V_n(t), V_n(s)) = \frac{\frac{1}{n}\sum_{i=1}^{[nt]}(\frac{[nt]-i+1}{n})(\frac{[ns]-i+1}{n})}{\sqrt{\frac{1}{n}\sum_{i=1}^{[nt]}(\frac{[nt]-i+1}{n})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{[ns]}(\frac{[ns]-i+1}{n})^2}} + O\left(\frac{1}{n}\right)$$

konverguje pro $n \to \infty$ k

$$\frac{\int_0^t (t-x)(s-x)\,dx}{\sqrt{\int_0^t (t-x)^2\,dx}\sqrt{\int_0^s (s-x)^2\,dx}} = \frac{3}{2}\sqrt{\frac{t}{s}} - \frac{1}{2}\sqrt{\left(\frac{t}{s}\right)^3}.$$

Pro velká n lze tedy maxima procesu $\{V_n(t)\}$ aproximovat odpovídajícími maximy procesu $\{V(t)\}$ s korelační funkcí $r(t,s) = (3/2)\sqrt{t/s} - (1/2)\sqrt{(t/s)^3}$ pro $t \le s$. Po transformaci času dostáváme

$$\max_{\beta \le t \le 1} V(t) = \max_{0 \le \tau \le \log 1/\beta} V(e^{-\tau}), \quad \max_{1/n \le t \le 1} V(t) = \max_{0 \le \tau \le \log n} V(e^{-\tau}).$$

Proces $\{V(e^{-\tau})\}$ je stacionární derivovatelný proces s korelační funkcí

$$r(\tau) = \frac{3}{2}\exp\left(-\frac{1}{2}\tau\right) - \frac{1}{2}\exp\left(-\frac{3}{2}\tau\right).$$

a je zřejmě až na lineární transformaci času shodný s třetím procesem v kapitole 3.3. Jeho korelační funkce má v bodě 0 rozvoj:

$$r(\tau) = 1 - \frac{3}{8}\tau^2 + o(\tau^2) \quad \text{pro} \quad \tau \to 0.$$

To znamená, že $\dot{\sigma}_x = \sqrt{3/4}$ (srovnej s (8)) , takže podle (15) a (17)

$$P\left(\max_{[\beta\, n] \le k \le n} X_i > x\right) \approx \frac{\sqrt{3}}{2}\,\frac{1}{\sqrt{2\,\pi}}\,\phi(x)\log(1/\beta), \tag{30}$$

$$P\left(\max_{1\le k\le n} X_i > x\right) \approx 1 - \exp\left(-e^{-\frac{x-b_n}{a_n}}\right), \tag{31}$$

kde

$$b_n = \sqrt{2\log\log n} + \frac{1}{\sqrt{2\log\log n}}\left(\log\frac{\sqrt{3}}{4\pi}\right),$$

$$a_n = \frac{1}{\sqrt{2\log\log n}}.$$

Reference

[1] Antoch J., Hušková M., Jarušková D. (1998). *Change – point problém po deseti letech*. Robust'98, JČMF.

[2] Camuffo D., Jones P. (2002). *Improved understanding of past climatic variability from early daily european instrumental sources*. Kluwer Ac. Pub., Dordrecht/Boston/London.

[3] Embrechts P., Küppelberg C, Mikosch T. (1997). *Modelling extremal events*. Springer Verlag, Heidelberg.

[4] Falk M., Hüsler J., Reiss R.-D. (2004). *Laws of small numbers: Extremes and rare events*. Birkhäuser, Basel.

[5] Hušková M. (1988). *Detekce změny regrese a detekce změny rozdělení*. Robust'88, JČMF.

[6] Leadbetter M. R., Lindgren G. and Rootzén H. (1983). *Extremes and related properties of random sequences and processes*. Springer Verlag, Heidelberg.

[7] Rencová M. (2004). *Extrémy v teplotních řadách*. Robust'04, JČMF.

Adresa: ČVUT, FSV, katedra matematiky, Thákurova 7, 160 00 Praha 6

E-mail: jarus@mat.fsv.cvut.cz

LOCATING EYES

Jan Kalina, P. Laurie Davies

Keywords: Image analysis, object detection, template matching, robust estimation, eye pattern, algorithm optimization.

Abstract: The aim is to construct and implement an algorithm, which automatically finds eyes in pictures of human faces. This paper stresses the motivation for this work and its possible use by a group of genetics researchers. Preliminary transformations of the data are then described; we compared the performance of several robust estimation methods. A combination of the template matching method and several characteristics or measures makes it possible to distinguish the eyes from other parts of the face. The paper summarizes the results obtained so far.

1 Motivation

The primary goal is to find both eyes in a black-and-white picture of a human face. This problem has been widely investigated for commercial purposes and fast algorithms implemented in commercial software are now available. Why we are working on our own software procedure should follow from this section. Our hope is to write a program useful for the researchers of the Institute of Human Genetics at the university clinic in Essen. The difficulty consists in considering also dysmorphic faces of people, whose appearance is influenced by genetic deformations.

Figure 1 is an example of the input matrix of data. A *gray value* corresponding to each pixel lies in the interval $[0, 1]$, where low values are black and high values white.

There exist several approaches to locating the eyes (or landmarks, to be more general). Typical examples include wavelet transformations, neural networks, support vector machines or template matching. Our aim is to construct the procedure to answer the needs of the geneticists, so let us shortly describe their research. A picture is taken always in a standard position, when the person is sitting straight against the camera looking in it. Inadmissible are pictures from a side or with eyes not well visible. These artificial conditions make the task easier. Still it often happens that the eyes are not in a perfectly horizontal position. We call such face to be *rotated*. We stress that the whole paper considers only this rotation in a plane, such as in Figure 2, where the whole face is well visible from the front.

The output of our work is primarily to detect the rotation of the face. Then it can be rotated to have the eyes horizontally. Other facial features such as the nose, mouth or ears can be then found based on the position and distance of the eyes. An automatic description of the face by biometric measures is then possible. Now the medical research has the following ambitions: to diagnose

Figure 1: Example of a picture.

Figure 2: A picture with extreme rotation.

genetic defects from a picture of a face; to examine the connection between the genetic code and the size and shape of facial features; and also to visualize a face based only on its biometric measures.

The everyday experience of the geneticists with available software is unsatisfactory for its high sensitivity even to a small rotation of the face. That is why they plan to use *first* our procedure to estimate the rotation of the face and only *then* to use the current software to look for other facial features.

2 Robustification

Filtering or robustification is a transformation often used to remove noise from images. For example [6] is using the median transform, while [5] gives theoretical arguments in favour of the trimmed mean and other L-estimators than the median. We have not found a similar justification of the very robust methods mentioned below.

As the pictures are taken at the clinic in Essen, the light of two bulbs is reflected in each of the two eyes. Their position in the eyes is however not always the same, so they represent a nuisance element which we want to get rid of. However it is desirable to change only small details, but not the picture as a whole. In fact Figure 1 is a picture after a suitable LMS-robustification.

For each pixel we take the gray values from its circular neighbourhood and compute the least median of squares (LMS) or some other very robust estimator. The information about the coordinates is lost. In this context is the LMS estimator equivalent to the mean of the shortest half of the data; see for example [1] for this and other connections among robust estimators. The computation of the least trimmed squares (LTS) is slower, so is the least weighted squares (LWS) estimator proposed by [7]. Here the LTS estimator corresponds to the mean of such half of the data (or a group of some $h < n$ observations), which has the smallest variance. A survey of algorithms for computing these robust estimators in (not only) location-model can be found in [4].

The performance of all these very robust estimators is not very different. We did not attempt to compare the results systematically. The only poor results are given by the median, which removes contrast and the resulting picture is rather grayish.

3 The template approach

Template matching is a tailor made method for object detection which uses the information about the ideal shape. [3] perceives template matching as a convolution of the template with the image and studies its theoretical aspects. A template for the whole face is used for example in [2]. [8] criticizes that template matching ignores the individual variability and prefers using priors and Bayesian approach.

We use the template approach only as one of many steps of the algorithm. Our ambition is not to find the eyes immediately, but only to select several areas, which can be considered suspicious; see Figure 3, where suspicious areas are black and the rest of the picture was reduced to a grayish shade. To convict the non-suspicious areas, other measures and criteria have to be used.

We compared the performance of several eye templates. The highest correlation with real eyes is attained for average eyes, where the gray values are taken as an average of eyes from several pictures. Other our templates inclu-

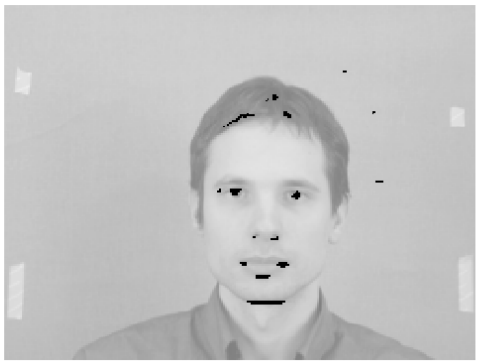


Figure 3: Areas correlated with the template.

ded real or robustified eyes from one of available pictures or simply a black circle with a white neighbourhood.

Each template is projected to various sizes, rotated by various angles and also consider a mirror reflection to detect both right and left eyes. After each of these possible transformations, we look for such size and rotation of the template, which lead to its highest correlation with any area. This size and rotation is then used in the next section.

In the image analysis literature we have found the correlation coefficient to be the only measure of similarity between a template and an area of the picture. It seemed to us natural to examine also other measures, for example the least square estimate of the slope in the regression and also the residual sum of squares in the regression, where the picture is modelled as a response of the template. We studied some nonparametric and robust analogies, including the rank correlation coefficient, robust estimates of the slope or robustified sum of squares, for example the trimmed sum of squares in the LTS context or the weighted sum of squares for the LWS. Aside from a considerable reduction of the speed, the results were not satisfactory. The reason is that an eye consists of both black and white points. Replacing the gray values by ranks removes the constrast between both groups and very robust approaches completely discard one of the groups. A non-robust approach is therefore desired and the correlation coefficient can be recommended.

4 Other measures and criteria

An important task is to describe the eye pattern by means of such characteristics, which would help to distinguish between eyes and other areas. Each of the following paragraphs describes one of such measures.

The total variation compares each pixel with its four neighbours. We add the absolute values of differences between the gray value in the pixel and the gray values of its neighbours. The highest values appear at boundaries between black and white areas, which typically include the eyes.

The eyes usually report a large variability of gray values. Thus we can take a circular neighbourhood of each pixel and compute the sample variance or interquartile range of the gray values.

The distribution of gray values in the eyes has a typical structure, which can be compared with a template. Ignoring the coordinates, we take gray values from a circular area of the picture and arrange them in ascending order. Let us denote them by $y_1, y_2, \ldots, y_n$. Let the ordered values from a template be $x_1, x_2, \ldots, x_n$. We look for such pixels, for which is the minimal sum

$$\sum_{i=1}^{n}(y_i - \alpha - x_i)^2$$

over all possible α very small. This corresponds to measuring the goodness of fit between two curves of ordered gray values, allowing for a vertical shift.

Take a rectangular neighbourhood of a pixel and consider average gray values over each column. Again use a template, based on a typical behaviour of real eyes. If the optimal template has been found rotated, which suggests that the face is rotated, then the rectangle needs to be taken also rotated.

The next measure is based on the idea that the middle part of an eye is more variable than its outer parts. We take a rectangular neighbourhood of each pixel, compute the sample variance in each column (or row) and compare these values with a triangular template with the highest values in the middle. The rectangle should be taken rotated by the same angle as the template.

5 Constructing the algorithm

We have described all our measures. The ambition is to use them to distinguish between eyes and other areas. Unfortunately the eyes do not always have the most extreme values of these measures. Very often there are areas which have a much larger variability than the eyes. This happens typically at boundaries between a black sweater and light background or a black shirt and white collar. Therefore to combine all the measures, we use indicators if a threshold has been achieved rather than the values of the measures themselves.

All the measures are applied to areas rather than to a single pixel. However in the next part we talk about applying these measures to a particular pixel, which means we consider an area centered at the desired pixel.

The measures have significant values somewhere in the eyes, but not necessarily in the center of the iris. For example the total variation is high at the boundary of the eyes, so the centers would be declared not suspicious. We need to distinguish between the global and local concepts of *locating*

and *localizing*, where locating stands for finding the eye-areas in the whole picture, in contrary to localizing, which means detecting the precise, exact position of the center part or most important point of an eye.

We use the following solution several times during the algorithm, which solves the question of localizing the eyes perfectly. The suspicious areas can be made larger by considering with each suspicious pixel also its small circular neighbourhood. An example is Figure 4, obtained from Figure 3 after several steps of the algorithm. This transformation should be supplemented by forgetting all previous results and creating new small suspicious areas centered in the darkest pixels of the previous enlarged areas.

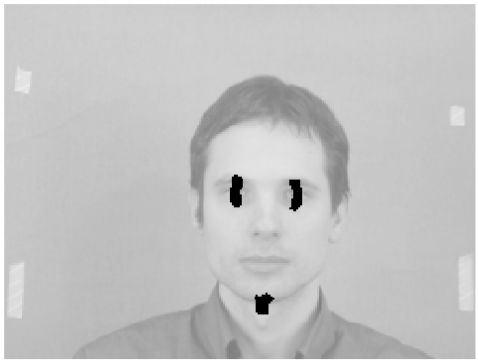


Figure 4: Suspicious areas after combining several measures.

Now we describe how to combine all measures together. We start with areas well correlated with the template of the optimal size and rotation, from the previous section. For the threshold we take the value 0.6 and also 75 % of the maximal correlation attained. These bounds are not very high, which ensures that the eyes are not lost when the size and rotation of the template are reasonable.

For each of the suspicious pixels we compute the values of other measures and compare these with thresholds. The number of suspicious pixels decreases. The thresholds must be chosen so low that the eyes are kept suspicious for as many pictures as possible. The results made us certain that even this liberal approach is sufficient, it is always possible to reject all pixels which are not the eyes. We use all measures described above and repeatedly use the transformation helpful for localizing. Because a good localization of eyes is attained only during the algorithm, some measures can be used repeatedly with stricter thresholds and the eyes are still kept suspicious. So we use 8 measures altogether, some of which are used twice. If the number of suspicious areas after these 8 measures is not exactly two, we repeat the steps for a different rotation of the template.

The order of measures was optimized in the following way. We started by the correlation with the template. Then we repeated the next steps. We found such thresholds for each of the remaining measures, so that the eyes were kept suspicious for a set of 20 pictures. We then added that measure to the algorithm which reduced the number of suspicious areas the most. After examining more pictures we had to lower the thresholds not to lose the eyes.

6 Results, final remarks

We examined a set of 135 pictures of 192×256 pixels taken under similar conditions. The photographer tried on purpose to create them in some way standardized. Thus we could take a fixed size of the templates and rotate them by -10, 0 and 10 degrees. The eyes were found and well located in each case, except for three pictures with halfway closed eyes. These pictures are however not interesting for the medical analysis; the eyes were found in other pictures of the same people.

The program is implemented in C. The results are sent to R, which offers nice methods for graphical output. The advantage of C is a high speed even with such massive data. On the other hand it has neither tools for handling matrices nor image processing functions.

Concerning the speed, the most time is spent on searching for the optimal size and rotation of the template. Right now the process takes one minute for a picture with 250×200 pixels when the template is rotated by $180, 170, \ldots, 0, \ldots, -170$ degrees. The extreme rotation is not likely to be used in practice, allowing only for three possible rotations by $10, 0$ and 10 degrees shortens the time to 10 seconds.

The last two paragraphs of this section discuss what happens with rotated faces. An *extreme* rotation by some 90 or 180 degrees is not likely to appear in practice. Anyway, eye templates with eyebrows are often successful in estimating the rotation of the face. As an example we show Figure 2, where both eyes were found and replaced by templates; these are difficult to recognize at first glance. Unfortunately the algorithm is not always successful, because some of the measures work well only when a correct information about the rotation is available. Modifications are possible, but for practical use not interesting.

A good example of a *small* rotation is Figure 1, where the line through the eyes differs from a horizontal line by an almost invisible angle of 3 degrees. Because our set of pictures included such with a small rotation, we can take this as a document of robustness of the template approach.

In the future, the performance of the algorithm should be examined for pictures with a different size of the head. A possible simplification can be to examine only the top half of a picture, where the eyes typically are. An interesting but complicated open problem is the varying pose, in other words a three-dimensional rotation of the face.

References

[1] Davies P.L., Gather U. (2004). *Robust statistics.* In Gentle J.E., Härdle W., Mori Y. (eds.): Handbook of computational statistics. Concepts and Methods. Springer-Verlag, Heidelberg, 655 – 695.

[2] Graf H.P., Cosatto E., Gibbon D., Kocheisen M., Petajan E. (1996). *Multi-modal system for locating heads and faces.* Technical Report, AT&T Labs. `www.research.att.com/resources/trs`

[3] James M. (1987). *Pattern recognition.* BSP Professional books, Oxford.

[4] Mašíček L. (2004). *Diagnostika a senzitivita robustních modelů.* Dissertation thesis, MFF UK, Praha. (Diagnostics and sensitivity of robust models. In Czech.)

[5] Pitas I., Venetsanopoulos A.N. (1990). *Nonlinear digital filters.* Kluwer, Dordrecht.

[6] Starck J.-L., Murtagh F., Bijaoui A. (1998). *Image processing and data analysis. The multiscale approach.* Cambridge University Press, Cambridge.

[7] Víšek J.Á. (2001). *Regression with high breakdown point.* ROBUST 2000, JČMF and Česká statistická společnost, 324 – 356.

[8] Winkler G. (1995). *Image analysis, random fields and dynamic Monte Carlo methods. A mathematical introduction.* Springer, Berlin.

Acknowledgement: This work was supported by the grant SFB 475, University of Dortmund. Participation of J. K. at Robust was supported by the grant GA ČR 402/03/0084 of the Grant Agency of the Czech Republic. Ing. Michal Dobeš, Ph.D. had several suggestions how to improve the paper. J. K. is thankful to his colleagues for help with hardware and software issues.

Address: J. Kalina, P. Laurie Davies, Universität Duisburg–Essen, Fachbereich Mathematik, 45117 Essen, Germany

E-mail: `kalina@stat-math.uni-essen.de`

KLASIFIKAČNÍ A REGRESNÍ LESY

Jan Klaschka, Emil Kotrč

Klíčová slova: Klasifikační stromy, klasifikační lesy, bagging, boosting, arcing, Random Forests.

Abstrakt: Klasifikační les je klasifikační model vytvořený kombinací určitého počtu klasifikačních stromů. Každý strom přiřazuje hodnotě vektoru prediktorů nějakou třídu a výsledná klasifikační funkce je dána hlasováním. Obdobně regresní les sestává z několika regresních stromů a výsledná regresní funkce je obvykle definována jako vážený průměr regresních funkcí jednotlivých stromů. V práci jsou stručně vysvětleny některé metody vytváření lesů, jmenovitě tzv. bagging, boosting, arcing a Random Forests.

1 Úvod

Klasifikační a regresní stromy se pěstují od 60. let. Silným metodologickým impulsem byla v 80. letech kniha [3], popisující tehdy novou metodu CART (Classification And Regression Trees). Věrozvěstem metody CART a stromů obecně na ROBUSTu se v r. 1988 stal Jaromír Antoch příspěvkem [1].

Mezitím, co od druhé poloviny 90. let přibývaly na ROBUSTu další práce o stromech [8], [11], [12], odstartovala ve světě – nejvíce ale asi v pracovně Leo Breimana, jednoho z „otců" CARTu – nová etapa rozvoje metod analýzy dat založených na stromech. Zdá se, že je na čase probrat ji také na ROBUSTu.

Tématem tohoto článku jsou klasifikační a regresní *lesy*. *Klasifikační les* je klasifikační model, jehož klasifikační funkce je dána kombinací (podle vhodně zvoleného pravidla) klasifikačních funkcí určitého počtu (typicky několika desítek) klasifikačních stromů. Obdobně lze charakterizovat *regresní les* – stačí v předcházející větě všude nahradit slovo „klasifikační" slovem „regresní".

Dlužno poznamenat, že článek nebude v rámci české literatury takovým pionýrským počinem jako již citovaná práce [1]: Přinejmenším se o některých technikách konstrukce lesů zmiňuje (byť v obecnější poloze) Berka v knize [2].

2 Klasifikační a regresní stromy

Les, jak každé malé dítě ví, se skládá ze stromů. Zopakujme některá základní fakta o stromech (podrobněji viz [3],[1]):

- *Klasifikační strom* představuje model pro data, kde každé pozorování patří do některé z tříd $C_1, \ldots, C_k$, $k \geq 2$. Současně je pozorování charakterizováno vektorem $\mathbf{x} = (x_1, \ldots, x_m)$ hodnot vysvětlujících proměnných (prediktorů) $X_1, \ldots, X_m$. V jedné a téže úloze se mohou vyskytovat prediktory kvantitativní i kvalitativní.
- Model lze popsat stromovým grafem sestávajícím z uzlů a orientovaných hran (orientace se nevyznačuje, hrana vede shora dolů).

- V každém neterminálním uzlu se strom větví – z uzlu vedou hrany do dvou nebo (v některých metodách) více dceřinných uzlů. Větvení je založeno na hodnotě jediného prediktoru. Nejběžnější je binární větvení podle odpovědi na otázku tvaru „$x_i < c$?“ pro kvantitativní prediktory X_i a „$x_i \in B$?“ (kde B je neprázdná vlastní podmnožina množiny všech hodnot veličiny X_i) pro prediktory X_i kvalitativní. Jedna hrana je pak přiřazena kladné a druhá záporné odpovědi. (Některé metody umožňují i větvení založená na lineární kombinaci kvantitativních prediktorů.)
- Pozorování podle hodnot prediktorů „postupuje“ od kořenového uzlu přes větvení v neterminálních uzlech k některému terminálnímu uzlu (listu). Množina všech listů určuje disjunktní rozklad prostoru hodnot prediktorů $\mathcal{X}$. Terminálnímu uzlu a zároveň pozorováním, která do něj patří, je přiřazena některá z tříd $C_1, \ldots, C_k$. Strom T tak určuje klasifikační funkci d_T definovanou na $\mathcal{X}$ s hodnotami v množině $\{C_1, \ldots, C_k\}$.

Regresní strom se od klasifikačního stromu liší tím, že každému terminálnímu uzlu je přiřazena reálná konstanta – odhad kvantitativní závisle proměnné Y. Regresní strom T definuje reálnou regresní funkci d_T, která je uvnitř množin odpovídajících terminálním uzlům konstantní.

K vytváření (pěstování) stromů prakticky všechny běžné metody využívají tzv. rekurzivní dělení (recursive partitioning). Konstrukce začíná stromem o jediném uzlu, do kterého patří všechna trénovací data (kořen je zároveň listem). Probere se množina všech možných větvení a pro každé z nich se vypočte kriteriální statistika (splitting criterion), která – budiž řečeno bez podrobností – hodnotí, nakolik jsou potenciální dceřinné uzly co do hodnot závisle proměnné vnitřně homogenní a navzájem odlišné. Větvení s maximální hodnotou kritéria se vybere jako nejlepší a použije se v modelu, k němuž tak přibude dvojice (popř. v některých metodách větší počet) uzlů, jež jsou prozatím terminální. Data, která patří do kořenového uzlu, se rozdělí podle hodnot prediktorů mezi nové dceřinné uzly. Pro každý z těchto provizorních listů se procedura opakuje, jako by se jednalo o kořen – hledá se nejlepší větvení, atd.

Při konstrukci klasifikačního stromu je žádoucí dosáhnout co nejmenší skutečné (generalizační) klasifikační chyby $R_P(T) = P\big(d_T(\mathbf{X}) \neq Y\big)$, kde P je sdružené rozdělení vektoru prediktorů $\mathbf{X}$ a závisle proměnné Y s hodnotami v $\{C_1, \ldots, C_k\}$. V regresních úlohách obdobnou roli nejčastěji hraje (skutečná) střední kvadratická chyba $R_P(T) = E_P\big(Y - d_T(\mathbf{X})\big)^2$. S rostoucí velikostí stromu sice stále klesá (nebo alespoň neroste) chyba na trénovacích datech, ale skutečná chyba v mnoha typických situacích klesá jen do určité velikosti, pak s dalším zvětšováním stromu opět roste.

Podle přístupu k problému stanovení „správné velikosti“ se metody pěstování stromů dělí do dvou skupin. Metody z prvé skupiny přidávají nová větvení, jen dokud to přináší dostatečně velký okamžitý efekt. Když se takové větvení nenajde, proces končí a strom je hotov. Metody druhého typu (např. CART [3]) nejdříve vypěstují tzv. *velký strom* $T_{\max}$, který se následně proře-

zává – některé uzly se opět odstraňují. Při kostrukci $T_{\max}$ se uzel stane terminálním, až když obsahuje méně pozorování než zvolená mez nebo všechna pozorování v uzlu patří do téže třídy, popř. mají stejné hodnoty všech prediktorů. O tom, jak velká část stromu $T_{\max}$ se má odstranit, rozhodují empirické odhady skutečné chyby. Ty se získávají různými způsoby, například pomocí testovacích (validačních) dat, která byla k tomu účelu při konstrukci stromu „ponechána stranou“, nebo složitějším „trikem“, křížovou validací (cross-validation), při níž se v opakovaných analýzách všechna pozorování střídají v rolích prvků trénovací a testovací množiny. (Detaily viz [3], [1]).

Mezi metodami pěstování stromů požívají asi největší prestiže Breimanův, Friedmanův, Olshenův a Stoneův CART (**C**lassification **A**nd **R**egression **T**rees) ([3]) a Quinlanova metoda C4.5 ([9]). Jediná současná implementace CARTu „posvěcená“ autory metody je šířena komerčně firmou Salford Systems[1] (San Diego, CA, USA). Volné programy tree a rpart v rámci projektu R[2] však představují také dosti věrné implementace metodologie CART. Program C4.5 je k dispozici bezplatně[3]. Vylepšenou verzi pod názvy C5.0 (Unix) a See5 (Windows) komerčně šíří Quinlanova firma Rulequest Research[4].

Informace o řadě dalších programů konstruujících stromy lze nalézt např. na webových stránkách Yu-Shan Shiha z Tchai-wanu[5] (spoluautora metod QUEST a CRUISE).

Výhodou klasifikačních a regresních stromů je, že pružně postihují vztahy mezi různými typy veličin, nelineární závislosti, interakce proměnných a závislosti, které mají rozdílnou podobu v různých částech prostoru $\mathcal{X}$. Ve srovnání s klasickými parametrickými metodami dosahují často srovnatelné přesnosti, ale poskytují přitom daleko přehlednější a názornější modely.

Nevýhodou stromů je, že jsou obvykle značně nestabilní: Zhusta pro jedna a tatáž data existuje mnoho různých stromů s přibližně stejnou chybou a při malé změně dat nebo vstupních parametrů se může výsledný strom (a příslušná klasifikační nebo regresní funkce) výrazně změnit.

3 Lesy

Myšlenka klasifikačních a regresních *lesů* je vcelku prostá: Co místo jednoho stromu T vypěstovat L stromů $T_1, \ldots, T_L$ a vytvořit z nich „komisi“, která se bude o zařazení pozorování do tříd, (popř., půjde-li o regresní úlohu, o predikované hodnotě závisle proměnné) „usnášet“[6]? Jinými slovy, „agregovaná“

[1] `http://www.salford-systems.com`

[2] `http://cran.r-project.org`

[3] `http://www.rulequest.com/Personal`

[4] `http://www.rulequest.com`

[5] Užší výběr `http://www.math.ccu.edu.tw/~yshih/trees.html`, širší, ale ne zcela aktuální přehled `http://www.math.ccu.edu.tw/~yshih/tree.html`.

[6] Již jednou jsme se odvolávali na znalosti malých dětí, a i toto si dnes kdekteré dítě dokáže představit, pokud vidělo ve filmu Dvě věže, druhém dílu trilogie Pán prstenů, jak sněm Entů (chodících stromů) rozhoduje o tom, zda jít do války, tj. jak řeší klasifikační úlohu, zda vstupní informace, které jsou k dispozici, patří do třídy „válka“, nebo „mír“.

klasifikační (popř. regresní) funkce $d_A(\mathbf{x})$ vznikne vhodným zkombinováním klasifikačních (popř. regresních) funkcí $d_1(\mathbf{x}), \ldots, d_L(\mathbf{x})$ jednotlivých stromů. Přirozeným a jednoduchým způsobem kombinace je u regrese aritmetický průměr a u klasifikace většinové hlasování, tj.

$$d_A(\mathbf{x}) = C_{i^*}, \text{ pokud } \#\{j; d_j(\mathbf{x}) = C_{i^*}\} = \max_{i=1,\ldots,k} \#\{j; d_j(\mathbf{x}) = C_i\},$$

kde symbol # značí počet prvků. (Nejednoznačnost vyplývající ze shody počtu hlasů se řeší např. znáhodněním.)

Kombinování klasifikačních funkcí (v regresi to je analogické, rozdíly si lze domyslet) může být také o něco složitější. Při hlasování může váha hlasu každé z L klasifikačních funkcí záviset na chybě stromu na trénovacích datech (pochopitelně přesnější strom má vyšší váhu). Váha hlasu jednotlivého stromu případně nemusí být stejná pro všechny hodnoty vektoru prediktorů $\mathbf{x}$ – může záviset např. také na tom, jak je list příslušného stromu, do kterého $\mathbf{x}$ patří, velký (kolik pozorování z trénovacího souboru do něj patří) a „čistý“ (jak výrazná je převaha nejfrekventovanější třídy). Složitější vážení hlasů je dílem už standardní součástí některých metod konstrukce lesů, ale dílem také ještě předmětem výzkumu (viz např. [15], [13]).

3.1 Čtvero způsobů, jak na to

Dobrá, vypěstovat více stromů, ale kde je vzít? Použijeme-li na jedna a tatáž data opakovaně např. program CART se stejnými vstupními parametry, dostaneme pokaždé tentýž strom. Kombinováním totožných stromů pak nic nového nezískáme. Rozdělit data na L disjunktních částí a každou použít ke konstrukci jednoho stromu se také nezdá být dobrý nápad. Problém přesto má řešení, resp. více řešení. Podívejme se na některá z nich.

3.1.1 Bagging je akronym, zkratka „**b**ootstrap **agg**regat**ing**“. Základní citace je Breimanův článek [4].

Z trénovacího datového souboru $\mathcal{L} = \{(\mathbf{x_1}, y_1), \ldots, (\mathbf{x_n}, y_n)\}$ se vytvoří náhodným výběrem *s vracením* L souborů $\mathcal{L}_1, \ldots, \mathcal{L}_L$ velikosti n (bootstrapových výběrů), každý z nich se použije k sestrojení jednoho klasifikačního (popř. regresního) stromu a výsledný klasifikační (nebo regresní) les je pak dán většinovým hlasováním se stejnými vahami (resp. aritmetickým průměrem dílčích regresních funkcí).

Do bootstrapového výběru jsou některá pozorování z $\mathcal{L}$ vybrána opakovaně a některá naopak vůbec. Počet opakování má pro jednotlivé pozorování z $\mathcal{L}$ asymptoticky (pro $n \to \infty$) Poissonovo rozdělení se střední hodnotou 1. Pravděpodobnost, že pozorování nebude vůbec vybráno, je tedy přibližně $e^{-1} \approx 0.37$. Bootstrapový výběr je tudíž tvořen asi 63% pozorování z $\mathcal{L}$, 37% zůstává mimo.

Breiman v [4] uvádí u lesů velikosti $L = 50$ (tvořených stromy vypěstovanými metodou CART) v několika reálných i umělých klasifikačních úlohách snížení generalizační chyby oproti jednomu stromu o 20-47% a velmi podobná čísla – 22-46% – udává také pro úlohy regresní.

3.1.2 Boosting a arcing. *Boosting* (to boost – zesilovat) je původně pojem z teorie strojového učení ([14], [5]) a v analýze dat se tak obvykle označuje algoritmus AdaBoost (**ada**ptive **boost**ing), navržený v článku [7].

Mějme klasifikační metodu (nemusí se jednat jen o stromy), která vytváří klasifikační model T na základě trénovacích dat $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ a vektoru $\mathbf{w} = (w_1, \ldots, w_n)$ vah přiřazených jednotlivým pozorováním. Algoritmus AdaBoost konstruuje posloupnost rozdílných modelů $T_1, \ldots, T_L$ s klasifikačními funkcemi $d_1(\mathbf{x}), \ldots, d_L(\mathbf{x})$ tak, že se podle předcházejících výsledků postupně upravují váhy případů. V prvním kroku se použije váhový vektor $\mathbf{w}_1$ zadaný uživatelem (např. rovnoměrné váhy) a vytvoří se model T_1. V dalších krocích se pak vždy ke konstrukci modelu T_i $(i = 2, \ldots, L)$ použije váhový vektor $\mathbf{w}_i$ získaný takovou úpravou vektoru $\mathbf{w}_{i-1}$, že se váhy pozorování chybně klasifikovaných modelem T_{i-1} zvýší a správně klasifikovaných sníží. Klasifikační metoda tak stále více „soustředí pozornost“ na „obtížná“ pozorování, která vzdorují zařazení do správné třídy. Váha modelu při hlasování závisí na chybě modelu na trénovacích datech. (Konkrétněji viz [7], [5], [2].)

Algoritmus nazvaný v článku [7] AdaBoost je vlastně určen jen pro klasifikační úlohu se dvěma třídami, ale v článku jsou popsány také modifikace AdaBoost.M1 a AdaBoost.M2 pro úlohy s více třídami a AdaBoost.R pro regresní úlohy s hodnotami závisle proměnné v intervalu $[0, 1]$.

Arcing je další Breimanův akronym (**a**daptive **r**esampling and **c**ombining). Základní prací je článek [5]. Arcing představuje spojení myšlenky baggingu a boostingu. Váhy případů se postupně upravují stejným způsobem jako v AdaBoostu, ale používají se jinak: Místo toho, aby s těmito vahami vstupovala všechna pozorování do analýzy, jsou jako v baggingu vytvářeny výběry s vracením, přičemž váhy (náležitě normované) slouží jako pravděpodobnosti „vytažení“.

Ukazuje se (viz např. [10], [5]), že boosting i arcing u klasifikačních stromů většinou snižují generalizační chybu ještě více než bagging. V některých případech však mohou výsledky být naopak katastrofální – zejména tehdy, když data jsou „zašuměná“ a u části trénovacích dat je hodnota závisle proměnné y_i jiná, než „by správně měla být“. Boosting nebo arcing pak vede k tomu, že se učíme opakovat chyby v datech.

3.1.3 Metoda Random Forests je opět „dítkem“ Leo Breimana, jehož článek [6] představuje základní citaci. Je založena na několika „tricích“:

- Trénovací soubory pro jednotlivé stromy jsou (jako v baggingu) bootstrapové výběry z datového souboru $\mathcal{L}$.
- Při volbě větvení pro daný uzel se z m prediktorů $X_1, \ldots, X_m$, které jsou k dispozici, nejdříve náhodně vybere některých m_0, načež se nejlepší větvení hledá již klasicky, ale jen mezi těmi větveními, která jsou založena na vybraných m_0 veličinách.
- Pěstují se velké stromy (viz sekci 2), které se neprořezávají.

Nejen že náhodný výběr prediktorů výrazně urychluje výpočty, ale experimenty v [6] také dávají zejména v klasifikačních úlohách velmi dobré výsledky – mj. srovnání s metodou AdaBoost vyznívá pro Random Forests příznivě. (V regresi je efekt méně výrazný a zdá se, že jsou o něco výhodnější některé modifikace uvedeného postupu.) Zvláště přínosná se metoda ukazuje u problémů, kde existuje velký počet prediktorů, z nichž každý sám o sobě obsahuje jen málo informace o závisle proměnné.

Jak volit parametr m_0? Nejvhodnější hodnota závisí na úloze a uživatel může experimentovat. Jednoduché rozumné doporučení je použít m_0 blízké $\log_2 m$.

U některých problémů se jako dobré, nebo dokonce nejlepší ukazuje $m_0 = 1$. To znamená, že se prediktor, který se má v uzlu použít pro větvení, vybírá zcela náhodně! Jinými slovy, na tom, jaké větvení se vybere, v podstatě nezáleží, hlavně že se prostor hodnot prediktorů vůbec nějak rozparceluje – hlasování (nebo v případě regrese průměrování) to dá do pořádku. To je mimochodem zcela protichůdné optimalizaci stromů, o níž byly v minulých letech na ROBUSTu dva příspěvky ([11], [12]) – připomeňme, že větvení se optimalizuje za cenu drastických výpočetních nákladů a s výsledky (co do generalizační chyby) mnohdy ne právě uspokojivými.

Vidíme zde jeden podstatný rozdíl mezi baggingem, boostingem a arcingem na jedné straně a metodou Random Forests na straně druhé. Bagging, boosting a arcing jsou nadstavbou nad klasickými metodami (jako CART nebo C4.5) zaměřenými na pěstování co nejpřesnějších stromů. U metody Random Forests na kvalitě jednotlivých stromů tolik nezáleží; cílem není minimalizovat chybu stromů, ale jen celého lesa.

3.2 Proč to funguje

Zatím jsme se omezili na popisný přístup – říkáme, jak která metoda postupuje, ale ne proč. Stejně tak informujeme, jaké jsou empirické výsledky, ale nevysvětlujeme, jak je možné, že to tak je. Může se tak zdát, že nové způsoby konstrukce lesů se navrhují podle vzoru „Myslím, myslím, nevím dál, co bych ještě udělal."

Ve skutečnosti se při studiu vlastností lesů neuplatňují jen numerické experimenty, ale také teoretické úvahy, byť částečně heuristického charakteru. Čtenáře v tomto ohledu odkazujeme na dříve citovanou literaturu, ale alespoň něco málo naznačíme.

Z metod uvedených v paragrafu 3.1 se jen Random Forests týká specificky stromů a lesů. Bagging, boosting a arcing lze aplikovat nejen na stromy, ale na vcelku libovolnou klasifikační či regresní metodu. (Všimněme si, že v názvech článků [4] a [5] se mluví obecně o prediktorech a klasifikátorech, nikoli o stromech.)

Platí, vágně řečeno, že agregovaný klasifikační model, jehož klasifikační funkce d_A je dána kombinací dílčích klasifikačních funkcí $d_1, \ldots, d_L$ pomocí hlasování, je tím přesnější (tj. má tím menší skutečnou chybu), čím přes-

nější jsou dílčí klasifikátory a čím jsou si funkce $d_1, \ldots, d_L$ vzájemně méně podobné. U regrese s agregací průměrováním je to obdobné.

„Alchymie“ přidávání náhody do konstrukce modelu sleduje cíl najít co nejlepší kompromis: co možná zvýšit vzájemnou nepodobnost funkcí d_i, ale tak, aby přesnost dílčích modelů neutrpěla přespříliš.

Skrze lesy se ve výhodu obrací nevýhoda stromů – jejich nestabilita, tj. fakt, že malá změna dat zpravidla vede k velké změně stromu T a jeho klasifikační (nebo regresní) funkce d_T. K dalším nestabilním metodám patří třeba neuronové sítě nebo kroková lineární regrese. Příklady stabilních metod jsou lineární diskriminační analýza nebo metoda k nejbližších sousedů. Nestabilní metody lze pomocí baggingu a obdobných technik výrazně zpřesnit, zatímco u stabilních metod je efekt minimální, popřípadě dokonce záporný.

3.3 Software

Repertoár softwaru pro pěstování lesů je zatím skromnější než nabídka programů pro konstrukci stromů. Uveďme několik současných možností.

- Komerční verze programu CART zahrnuje bagging a arcing.
- Quinlanovy programy C5.0 a See5 (viz sekci 2) obsahují boosting.
- Softwarový systém Weka[7] zaměřený na strojové učení, vyvinutý na University of Waikato na Novém Zélandu, obsahuje pod názvem j48 implementaci metody C4.5, a dále procedury Bagging a AdaBoostM1, které se dají v kombinaci s j48 použít k pěstování lesů. Weka je freeware.
- Breimanův vlastní software Random Forests je volně k dispozici[8] a zároveň existuje komerční verze šířená firmou Salford Systems. Implementace randomForest v projektu R je freeware.

Pokud nějaký program pro konstrukci stromů umožňuje snadné zadávání vstupních údajů a dává výstupy ve vhodném tvaru, „dotvořit“ jej na software pro pěstování lesů je programátorsky nenáročné. Do původního programu není třeba zasahovat, takže nevadí, není-li dostupný ve zdrojovém tvaru.

4 Závěr

Klasifikační a regresní lesy představují významné zdokonalení metodologií založených na stromech. Za cenu únosného zvýšení výpočetní náročnosti se dosahuje výrazného zpřesnění modelů.

Patrně jedinou nevýhodou lesů oproti stromům je to, že se ztrácí pro stromy tak charakteristická přehlednost. Na les tvořený desítkami nebo stovkami stromů nejsme schopni, stejně jako je tomu např. u neuronových sítí, se dívat jinak než jako na „černou skříňku“. Jinými slovy, do lesa není vidět.

[7] `http://www.cs.waikato.ac.nz/~ml/index.html`
[8] `http://www.stat.berkeley.edu/users/breiman/RandomForests`

Reference

[1] Antoch J. (1988). *Klasifikace a regresní stromy.* In: ROBUST'1988. Sborník prací 5. letní školy JČSMF, J. Antoch, T. Havránek & J. Jurečková (eds.), Praha, JČSMF, 0 – 6.

[2] Berka P. (2003). *Dobývání znalostí z databází.* Praha: Academia.

[3] Breiman L., Friedman J.H., Olshen R.A., Stone C.J. (1984). *Classification and regression trees.* Belmont CA: Wadsworth.

[4] Breiman L. (1996). *Bagging predictors.* Machine Learning **24**, 123 – 140.

[5] Breiman L. (1998). *Arcing classifiers.* Annals of Statistics **26**, 801 – 849.

[6] Breiman L. (2001). *Random forests.* Machine Learning **45**, 5 – 32.

[7] Freund Y., Schapire R.E. (1997). *A decision-theoretic generalization of on-line learning and an application to boosting.* Journal of Computer and System Sciences **55**, 119 – 139.

[8] Klaschka J., Antoch J. (1997). *Jak rychle pěstovat stromy.* In: ROBUST'1996. Sborník prací 9. letní školy JČMF, J. Antoch & G. Dohnal (eds.), Praha, JČMF, 91 – 106.

[9] Quinlan J.R. (1992). *C4.5: Programs for machine learning.* New York: Morgan Kaufmann.

[10] Quinlan J.R. (1996). *Bagging, boosting, and C4.5.* In: Proceedings of the Thirteenth National Conference on Artificial Intelligence, 725 – 730.

[11] Savický P., Klaschka J., Antoch J. (2000). *Optimální klasifikační stromy.* In: ROBUST'2000. Sborník prací 11. letní školy JČMF, J. Antoch & G. Dohnal (eds.), Praha, JČMF, 267 – 283.

[12] Savický P., Klaschka J. (2002). *Lesk a bída optimálních stromů.* In: ROBUST'2002. Sborník prací 12. zimní školy JČMF, J. Antoch, G. Dohnal & J. Klaschka (eds.), Praha, JČMF, 256 – 267.

[13] Savický P., Kotrč E. (2004). *Experimental study of leaf confidences for random forest.* In: COMPSTAT 2004. Proceedings in Computational Statistics, J. Antoch (ed.), Heidelberg, Physica Verlag, 1767 – 1774.

[14] Schapire R.E. (1990). *The strength of weak learnability.* Machine Learning **5**, 197 – 227.

[15] Schapire R.E., Singer Y. (1999). *Improved boosting algorithms using confidence-rated predictions.* Machine Learning **37**, 297 – 336.

Poděkování: Práce byla podporována granty ME 701 MŠMT ČR a GA ČR 201/02/1456.

Adresa: J. Klaschka, E. Kotrč, Ústav informatiky AV ČR, Pod Vodárenskou věží 2, 182 07 Praha 8

E-mail: {klaschka,kotrc}@cs.cas.cz

MOSUM-TYPE TESTS FOR A CHANGE-POINT PROBLEM WITH CENSORED DATA

Lenka Komárková

Keywords: Change-point problem, censored data, rank tests, permutation principle, limit behavior.

Abstract: The contribution concerns about MOSUM-type test statistics for detection of a change in the distribution of variables that are independent but possibly censored. The MOSUM-type test statistic is suitable if we expect more than one change and it is useful as a diagnostic tool. The test statistics are derived using the same principle as for uncensored data. The limit behavior for such a class of test statistics is investigated under the hypothesis of "no-change" in the distribution of censored variables and further, the consistency of the test procedure is shown.

1 Introduction

We introduce a general model of random censorship, see e.g. Kalbfleisch and Prentice [12]. Typically, $X_1^0, X_2^0, \ldots, X_n^0$ is a sequence of independent nonnegative random variables (*the lifetimes* or *the survival times*), where the index i of X_i^0 corresponds to the chronological order the subject of interest (e.g. patient) has entered the study. The patient can be withdrawn from the study due to many reasons, e.g. an accidental death, a migration of human population or limited time of the study. More precisely, the lifetimes can be censored from the right by independent random variables $C_1, C_2, \ldots, C_n$, the so-called *censoring times*. In other words, instead of the survival times we observe the pairs $(X_1, \Delta_1), (X_2, \Delta_2), \ldots, (X_n, \Delta_n)$ only, where

$$X_j = \min(X_j^0, C_j), \quad \Delta_j = I(X_j^0 \le C_j) = \begin{cases} 1, & \text{if } X_j \text{ is uncensored,} \\ 0, & \text{if } X_j \text{ is censored.} \end{cases}$$

for $j = 1, 2, \ldots, n$. We assume that the lifetimes and the censoring times are independent variables.

We consider the following modification called as *the random censorship model with multiple changes*. We suppose that there exist $0 = \gamma_0 < \gamma_1 \le \cdots \le \gamma_q \le \gamma_{q+1} = 1$ with some finite $q \in \mathbb{N}$ such that the survival times $X^0_{\lfloor n\gamma_i \rfloor+1}, X^0_{\lfloor n\gamma_i \rfloor+2}, \ldots, X^0_{\lfloor n\gamma_{i+1} \rfloor}$ have an absolutely continuous distribution function F_{i+1} and the censoring times $C_{\lfloor n\gamma_i \rfloor+1}, C_{\lfloor n\gamma_i \rfloor+2}, \ldots, C_{\lfloor n\gamma_{i+1} \rfloor}$ have an absolutely continuous distribution function G_{i+1} for $i = 0, 1, \ldots, q$. Further suppose that $F_{i+1} \ne F_i$ and $G_{i+1} \ne G_i$, $i = 1, 2, \ldots, q$. We would

like to test the no-change null hypothesis against the alternative that at least one change has occurred:

$$H_0 : \gamma_1 = \gamma_2 = \cdots = \gamma_q = \gamma_{q+1} = 1,$$
$$H_1 : \gamma_1 \leq \gamma_2 \leq \cdots \leq \gamma_q \leq \gamma_{q+1}, \text{ where at least one inequality is strict}$$

and where the integer $q \geq 1$ can be known or unknown. In our testing problem is the basic task to decide if there is a change (or changes) in the model e.g. due to medical care in our case. The null hypothesis says that the lifetimes are i.i.d. random variables with common distribution function F_1 and moreover, the censoring variables are also i.i.d. and they have the common distribution function G_1.

Remark 1.1. *The considered model introduced above includes the so-called* Koziol–Green model with multiple changes*. In this model, the survival function of the censoring times is supposed to be a power of the survival function of the lifetimes. Supposing at least one change, i.e. $q \geq 1$, we get*

$$\forall\, t \geq 0 \qquad 1 - G_i(t) = (1 - F_i(t))^{\beta_i} \quad \textit{with } \beta_i > 0,\ i = 1, 2, \ldots, q+1.$$

There are only a few papers dealing with detection of a change in the distribution when only censored data are available. Stute [16] suggested estimators based on U-statistics for the change point and he studied their properties. His results were extended by Ferger [4] and Horváth [8]. Aly [1] treated the uncensored and the censored observations separately and finally, he mixed both the proposed test statistics based on the quantile processes together. Gombay and Liu [6] based their test on a generalization of the Wilcoxon rank statistic which can be expressed as a U-statistic. Extensive studies for such procedures were conducted in the doctoral thesis of Liu [15]. In all the papers listed above the censoring times $C_1, C_2, \ldots, C_n$ are supposed to be i.i.d. variables. Hušková and Neuhaus [10] developed their test as a generalization of two-sample rank tests under the random censoring. In contrast to the other mentioned authors, they considered not only the change in the distribution of the survival variables but also the change in the distribution of the censoring variables.

2 Test statistic

Motivated by the work of Hušková and Neuhaus [10] and the type of statistic used to test multiple changes in uncensored case (Csörgő, Horváth [3] or Antoch, Hušková, Jarušková [2]) we propose for testing (H_0, H_1) the MOSUM-type rank statistic as follows

$$T^{\sigma}_{n,D}(\tau_0) = \max_{D<k<n-D} \frac{|\sum_{j=k+1}^{k+D} a_n(j) - \sum_{j=k-D+1}^{k} a_n(j)|}{\sqrt{2D}\,\sigma_n(\boldsymbol{a})} \tag{1}$$

with $\sigma_n^2(\boldsymbol{a}) = \frac{1}{n}\sum_{j=1}^n a_n^2(j)$ and the scores which are weighted and have the following form

$$a_n(j) = \int_0^{\tau_0} w_n(t)\,\mathrm{d}N_j(t) - \int_0^{\tau_0} w_n(t)\,\frac{Y_j(t)}{\sum_{j=1}^n Y_j(t)}\,\mathrm{d}\Big(\sum_{j=1}^n N_j(t)\Big), \quad (2)$$

where $Y_j(t) = I(X_j \geq t)$ and $N_j(t) = \Delta_j\, I(X_j \leq t)$. The value τ_0 denoting the end of medical study is such a positive number for which

$$0 < \tau_0 < \tau := \sup\{x;\, F_i(x)\, G_i(x) < 1, i = 1, \ldots, q+1\}.$$

The test statistic also depends on D. We assume that $D = D(n)$ satisfies, as $n \to \infty$,

$$\frac{D}{n} \to 0, \qquad \frac{n^{2/(2+u)}\,\log n}{D} \to 0, \quad (3)$$

where u is a positive constant such that $\frac{1}{n}\sum_{j=1}^n |a_n(j) - \bar{a}_n|^{u+2} = O_{\mathrm{P}}(1)$, as $n \to \infty$. The assumption (3) means that D tends to infinity together with n but not too fast.

Next we describe the form of the considered weight function $w_n(X_j, \Delta_j;\, 1 \leq j \leq n) \geq 0$. An important class of weight functions is

$$w_n(t) = (\hat{S}_n(t-))^\rho \left(\frac{Y(t)}{n}\right)^\kappa I(Y(t) > 0), \quad (4)$$

where $\rho, \kappa \geq 0$ and $\hat{S}_n(t-) = \prod_{i:X_i<t}\left(1 - \frac{\Delta_i}{Y(X_i)}\right)$ is the left-continuous Kaplan–Meier estimate of the survival function. Such a class of weighted test statistics includes commonly used test statistics in practice like *the log-rank statistic* ($\rho = 0$, $\kappa = 0$), *the Prentice–Wilcoxon statistic* ($\rho = 1$, $\kappa = 0$) and *the Gehan–Wilcoxon statistic* ($\rho = 0$, $\kappa = 1$) which are generalizations of the Savage and the Wilcoxon statistic for uncensored data, for more information see e.g. Hájek et al [7].

3 Test procedure

We describe the exact (permutation) test based on $T_{n,D}^\sigma(\tau_0)$ and we construct the asymptotic decision rule for (H_0, H_1) based on the limit distribution of $T_{n,D}^\sigma(\tau_0)$ under H_0.

Since under H_0 the observations $(\boldsymbol{X}, \boldsymbol{\Delta}) = ((X_1, \Delta_1), \ldots, (X_n, \Delta_n))$ are i.i.d., we can apply *the permutation principle* (see e.g. Lehmann [14] or Good [5]). It means that the permutation distribution $F_{n,D}(\cdot, (\boldsymbol{X}, \boldsymbol{\Delta}))$ of the test statistic $T_{n,D}^\sigma(\tau_0)$ can be described as the conditional distribution given $(\boldsymbol{X}, \boldsymbol{\Delta})$ of

$$T_{n,D}^\sigma(\tau_0, \boldsymbol{Q}) = \max_{D<k<n-D} \frac{|\sum_{j=k+1}^{k+D} a_n(Q_j) - \sum_{j=k-D+1}^{k} a_n(Q_j)|}{\sqrt{2D}\,\sigma_n(\boldsymbol{a})},$$

where $\boldsymbol{Q} = (Q_1, Q_2, \ldots, Q_n)$ is a random permutation of $(1, 2, \ldots, n)$, precisely it can be expressed for $x \in \mathbb{R}$ as follows

$$F_{n,D}(x, (\boldsymbol{X}, \boldsymbol{\Delta})) = \mathrm{P}\big(T^{\sigma}_{n,D}(\tau_0) \le x \,|\, (\boldsymbol{X}, \boldsymbol{\Delta})\big) = \frac{\#\{\boldsymbol{q} \in \mathcal{Q}_n; T^{\sigma}_{n,D}(\tau_0, \boldsymbol{q}) \le x\}}{n!},$$

where $\mathcal{Q}_n$ is the set of all permutations of integers $(1, 2, \ldots, n)$. Denoting by $c_{n,D}(\alpha, (\boldsymbol{X}, \boldsymbol{\Delta}))$ the corresponding $100(1-\alpha)\%$-quantile the critical region of *the exact (permutation) test* based on $T^{\sigma}_{n,D}(\tau_0)$ with the level α has the form

$$T^{\sigma}_{n,D}(\tau_0, \boldsymbol{Q}) \ge c_{n,D}(\alpha, (\boldsymbol{X}, \boldsymbol{\Delta})).$$

Under H_0 the distributions of $T^{\sigma}_{n,D}(\tau_0)$ and $T^{\sigma}_{n,D}(\tau_0, \boldsymbol{Q})$ are the same and the permutation distribution provides the exact critical values for our testing problem.

Practically, for large n it is not possible to calculate the value of the statistic $T^{\sigma}_{n,D}(\tau_0, \boldsymbol{q})$ for all $n!$ permutations $\boldsymbol{q}$. So instead, we generate a random sample from all possible permutations of size B large enough and determine the empirical critical value $x_{n,D}(\alpha, (\boldsymbol{X}, \boldsymbol{\Delta}))$ from this sample. Such calculated critical value $x_{n,D}(\alpha, (\boldsymbol{X}, \boldsymbol{\Delta}))$ provides a good estimate for the actual value $c_{n,D}(\alpha, (\boldsymbol{X}, \boldsymbol{\Delta}))$. We do the so-called *bootstrap without replacement*, for concrete example see Section Simulations.

The other way how to obtain appropriate approximation of critical values is through the limit behavior of the test statistic. Here we state the limit distribution of the MOSUM-type test statistic $T_{n,D}(\tau_0, \boldsymbol{Q})$.

Theorem 3.1. *Suppose the random censorship model with multiple changes. Let*

$$\sup_{0 \le t \le \tau_0} |w_n(t) - w(t)| = o_{\mathrm{P}}(1), \tag{5}$$

where w is a continuous nonrandom function on $[0, \tau_0]$ and

$$\int_0^{\tau_0} w^2(t)\,(1 - G_i(t))\,\mathrm{d}F_i(t) > 0, \quad i = 1, 2, \ldots, q+1, \tag{6}$$

be satisfied. Then for all $y \in \mathbb{R}$ we have, as $n \to \infty$,

$$\mathrm{P}\left(d_1\Big(\frac{n}{D}\Big)\; T^{\sigma}_{n,D}(\tau_0, \boldsymbol{Q}) \le y + d_2\Big(\frac{n}{D}\Big) \,|\, (\boldsymbol{X}, \boldsymbol{\Delta})\right) \xrightarrow{\mathrm{P}} \exp\big\{-2\mathrm{e}^{-y}\big\},$$

where $\boldsymbol{Q} = (Q_1, Q_2, \ldots, Q_n)$ denotes a random permutation of $(1, 2, \ldots, n)$. Moreover under $H_0 : \gamma_i = 1$ for $i = 1, 2, \ldots, q$

$$\mathrm{P}\left(d_1\Big(\frac{n}{D}\Big)\; T^{\sigma}_{n,D}(\tau_0) \le y + d_2\Big(\frac{n}{D}\Big)\right) \to \exp\big\{-2\mathrm{e}^{-y}\big\}$$

with d_1 and d_2 defined as follows

$$d_1(t) = \sqrt{2\log t}, \qquad d_2(t) = 2\log t + \tfrac{1}{2}\log\log t - \tfrac{1}{2}\log\pi + \log\Big(\frac{3}{2}\Big). \tag{7}$$

Důkaz. We mention the basic idea of the proof. Realize that the random variables $\sum_{j=1}^{k} a_n(Q_j)$, $k = 1, 2, \ldots, n$, given $(\boldsymbol{X}, \boldsymbol{\Delta})$, can be viewed as simple linear rank statistics, where the role of ranks is played by the random permutation $\boldsymbol{Q} = (Q_1, Q_2, \ldots, Q_n)$. Consequently, the statistic $T^{\sigma}_{n,D}(\tau_0, \boldsymbol{Q})$ given $(\boldsymbol{X}, \boldsymbol{\Delta})$ can be viewed as a function of a simple linear rank statistic and the theorem (Theorem 2 in Hušková [9]) on rank statistics for change point problem can be used. The assumptions (5) and (6) ensure that the assumptions of the applied theorem are fulfilled for our scores.
The whole proof can be found in Komárková [13], Corollary 2.4. □

We get the same limit distribution as for completely observable data, compare with the results given by Hušková, Slabý [11]. It can be seen that the rejection criterion of *the asymptotic test* based on $T^{\sigma}_{n,D}(\tau_0)$ has the following form

$$T^{\sigma}_{n,D}(\tau_0) > \frac{-\log(-\log(\sqrt{1-\alpha})) + d_2(\frac{n}{D})}{d_1(\frac{n}{D})}. \tag{8}$$

4 Consistency

Here we present the result on limit behavior of $T^{\sigma}_{n,D}(\tau_0)$ under alternatives. Particularly, we concentrate on *the consistency* of the test. Notice that for the considered class of test statistics under H_0 we have, as $n \to \infty$,

$$T^{\sigma}_{n,D}(\tau_0)\,(\log n)^{-1/2} = O_{\mathrm{P}}(1).$$

Therefore to show the consistency it suffices to show that under alternatives

$$T^{\sigma}_{n,D}(\tau_0)\,(\log n)^{-1/2} \xrightarrow{\mathrm{P}} \infty, \quad \text{as } n \to \infty.$$

Now, we formulate the assertion on the consistency of the test procedure based on $T^{\sigma}_{n,D}(\tau_0)$.

Theorem 4.1. *Suppose the random censorship model with multiple changes. Let* (5) *and* (6) *be satisfied. If*

$$\max_{i=1,2,\ldots,q} \left| \sum_{j=1}^{q+1} \left\{ \int_0^{\tau_0} w(t)\, \frac{(\gamma_j - \gamma_{j-1})\,(1-F_j(t))(1-G_j(t))}{\sum_{j=1}^{q+1}(\gamma_j - \gamma_{j-1})\,(1-F_j(t))(1-G_j(t))} \right.\right.$$
$$((1-F_{i+1}(t))(1-G_{i+1}(t))(\lambda_{i+1}(t) - \lambda_j(t))$$
$$\left.\left. - (1-F_i(t))(1-G_i(t))(\lambda_i(t) - \lambda_j(t)))\,\mathrm{d}t \right\} \right| > 0 \tag{9}$$

holds, where $\lambda_j(t) = -\frac{\mathrm{d}\,F_j(t)}{1-F_j(t)}$ *is the hazard functions corresponding to* $F_j(t)$, $j = 1, 2, \ldots, q+1$. *Then we have for any* $u > 0$, *as* $n \to \infty$,

$$D^{u-\frac{1}{2}}\, T^{\sigma}_{n,D}(\tau_0) \xrightarrow{\mathrm{P}} \infty$$

and hence the test based on $T^{\sigma}_{n,D}(\tau_0)$ *is consistent.*

Důkaz. We describe only the main steps of the proof. It can be assumed w.l.g. that

$$\lfloor n\gamma_{i-1}\rfloor \le \lfloor n\gamma_i\rfloor - D < \lfloor n\gamma_i\rfloor + D \le \lfloor n\gamma_{i+1}\rfloor, \quad i = 1, 2, \ldots, q.$$

By the definition of the test statistic $T^{\sigma}_{n,D}(\tau_0)$ we have

$$T^{\sigma}_{n,D}(\tau_0) \ge \max_{i=1,2,\ldots,q} \frac{\left|\sum_{j=\lfloor n\gamma_i\rfloor+1}^{\lfloor n\gamma_i\rfloor+D} a_n(j) - \sum_{j=\lfloor n\gamma_i\rfloor-D+1}^{\lfloor n\gamma_i\rfloor} a_n(j)\right|}{\sqrt{2D}\,\sigma_n(\boldsymbol{a})}.$$

By the appropriate asymptotic approximation we receive, as $n \to \infty$,

$$\sigma_n^2(\boldsymbol{a}) = \sum_{j=1}^{q+1}\Big\{(\gamma_j - \gamma_{j-1})\int_0^{\tau_0} w^2(t)\,(1 - G_j(t))\,\mathrm{d}F_j(t)\Big\} + o_{\mathrm{P}}(1).$$

By the assumption (6) of our theorem we get that $\sigma_n^2(\boldsymbol{a})$ is asymptotically bounded away from 0. The asymptotic representation (convergence in probability) for

$$\frac{\sum_{j=\lfloor n\gamma_i\rfloor+1}^{\lfloor n\gamma_i\rfloor+D} a_n(j) - \sum_{j=\lfloor n\gamma_i\rfloor-D+1}^{\lfloor n\gamma_i\rfloor} a_n(j)}{D}$$

has the form

$$\sum_{j=1}^{q+1}\Big\{\int_0^{\tau_0} w(t)\,\frac{(\gamma_j - \gamma_{j-1})\,(1 - F_j(t))(1 - G_j(t))}{\sum_{j=1}^{q+1}(\gamma_j - \gamma_{j-1})\,(1 - F_j(t))(1 - G_j(t))} \\ ((1 - F_{i+1}(t))(1 - G_{i+1}(t))(\lambda_{i+1}(t) - \lambda_j(t)) \\ - (1 - F_i(t))(1 - G_i(t))(\lambda_i(t) - \lambda_j(t)))\,\mathrm{d}t\Big\}.$$

This and the assumption (9) imply, as $n \to \infty$,

$$\max_{i=1,2,\ldots,q} \frac{\left|\sum_{j=\lfloor n\gamma_i\rfloor+1}^{\lfloor n\gamma_i\rfloor+D} a_n(j) - \sum_{j=\lfloor n\gamma_i\rfloor-D+1}^{\lfloor n\gamma_i\rfloor} a_n(j)\right|}{D} \xrightarrow{\mathrm{P}} const > 0.$$

The whole proof can be again found in Komárková [13], Theorem 2.10. □

5 Discussion

We finalize the paper with few remarks. Firstly, we note that the theorems can easily be modified for the situation when during the observation period the lifetimes change their distribution only and the censoring variables are i.i.d. Further, the approach of this paper could generally be extended to the situation when the distribution of the censoring variables can change

at different times than the distribution of the survival variables. However, the test procedure becomes then rather complicated since some estimators for the time of changes in the censoring distribution are needed. We refer the reader to Komárková [13] where some suggestions can be found. The same work by Komárková [13] contains additionally simulated critical values for the test presented in this paper.

Finally, it is worth to emphasize that we have studied performance of the test under the null hypothesis and proved consistency. However, it would be worthwhile to study properties of the test under the local alternatives.

References

[1] Aly E.-E.A.A. (1998). *Change point tests for randomly censored data.* In: B. Szyskowicz, (Ed.), Asymptotic Methods in Probability and Statistics, North Holland, 503 – 513.

[2] Antoch J., Hušková M., Jarušková D. (1998). *Change point problem po deseti letech (in Czech language).* In: J. Antoch, Ed., ROBUST'98, JČMF, Prague, 1 – 42.

[3] Csörgő M., Horváth L. (1997). *Limit theorems in change-point analysis.* John Wiley & Sons, Inc., New York.

[4] Ferger D. (1998). *Change-point analysis of censored data.* Economic Quality Control **13**, 167 – 177.

[5] Good P. (2000). *Permutation tests (Second edition).* Springer Verlag, New York.

[6] Gombay E., Liu S. (2000). *A nonparametric test for change in randomly censored data.* The Canadian Journal of Statistics **21**, 113 – 121.

[7] Hájek J., Šidák Z., Sen P.K. (1999). *Theory of rank tests (Second edition).* Academic Press, San Diego.

[8] Horváth L. (1998). *Tests for changes under random censorship.* Journal of Statistical Planning and Inference **69**, 229 – 243.

[9] Hušková M. (1997). *Limit theorem for rank statistics.* Statistics & Probability Letters **32**, 45 – 55.

[10] Hušková M., Neuhaus G. (2004). *Change point analysis for censored data.* Journal of Statistical Planning and Inference, accepted for publication.

[11] Hušková M., Slabý A. (2001). *Permutation tests for multiple changes.* Kybernetika **37**, 605 – 622.

[12] Kalbfleisch J.D., Prentice R.L. (2002). *The statistical analysis of failure time data (Second edition).* John Wiley & Sons, Inc., Hoboken, New Jersey.

[13] Komárková L. (2004). *Change point problem for censored data.* Doctoral Thesis, Charles University in Prague, Department of Statistics.

[14] Lehmann E.L. (1991). *Theory of point estimation.* Wadsworth & Brooks/Cole, California.

[15] Liu S. (1998). *Nonparametric tests for change-point problems with random censorship.* Doctoral thesis, University of Alberta, Edmonton.

[16] Stute W. (1996). *Changepoint problems under random censorship.* Statistics **27**, 255 – 266.

Acknowledgement: The work was supported by the grant GAČR 201/03/0945 and by the Research plan MSM 113200008.

Address: L. Komárková, University of Economics, Department of Information Management, Jarošovská 1117/II, 377 01 Jindřichův Hradec

E-mail: koblizle@fm.vse.cz

INFORMACE A DEZINFORMACE – STATISTICKÝ POHLED

Michala Kotlíková, Hana Mašková, Arnoštka Netrvalová, Pavel Nový, Dagmar Spíralová, František Vávra, David Zmrhal

Klíčová slova: Informace, entropie, divergence, dezinformace.

Abstrakt: Při práci se statistickými odhady sdílené informace

$$I(X:T) = \sum_{t\in T}\sum_{x\in X} p(x,t)\lg\frac{p(x,t)}{p(x)p(t)}$$

nastávají některé zdánlivě neočekávatelné situace. Protože pravděpodobnosti (hustoty) $p(x,t), p(x), p(t)$ pro nás nejsou dostupné, pracujeme s jejich odhady $e(x,t) \approx p(x,t), e(x) \approx p(x), e(t) \approx p(t)$ nebo s jejich některou parametrickou reprezentací (ale ta je také odhadem). Potom odhad $\hat{I}_n(X:T) = \frac{1}{n}\sum_{i=1}^{n}\lg\frac{e(x_i,t_i)}{e(x_i)e(t_i)}$ „konverguje k" $\sum_{t\in T}\sum_{x\in X} p(x,t)\lg\frac{e(x,t)}{e(x)e(t)}$. Tento výraz budeme značit $I(X:T;e)$ a bude dále nazýván sdílenou informací mezi náhodnými proměnnými X a T při využití modelů $e(x,t), e(x), e(t)$. Obdobná situace nastává i pro další pojmy z teorie informace jako je entropie a divergence. V tomto sdělení jsou prezentovány některé vybrané možnosti využití uvedeného jevu a vlastnosti předložených „výběrových měr" (např. $I(X:T;e)$ může být i záporná). Rozdíl $Di(X:T;e) = I(X:T;e) - I(X:T)$ pak může být užitečnou charakteristikou „přijatelnosti" (i např. vychýlenosti) zvolených modelů $e(x,t), e(x), e(t)$. Pro tento rozdíl jsme zvolili pracovní název „dezinformace". Název je motivován tím, že „vhodnou" volbou modelů lze dosáhnout poměrně širokého spektra hodnot „výběrové" informace $I(X:T;e)$ a tím i případně možných rozhodnutí (manipulace s daty a modelem).

1 Úvod

Výše uvedený koncept výběrové informace je využitelný jak pro spojitá, tak i diskrétní rozdělení pravděpodobnosti. Autoři se domnívají, že otevírá prostor zkoumání a popisování jevů v úlohách klasifikace, rozhodování a i testování hypotéz. V tomto příspěvku budou prezentovány některé výsledky pro diskrétní rozdělení, zaměřené na využití při odhadování. Prezentace je zaměřena do dvou oblastí. První oblastí budou náznaky toho, jak lze získat klasické formule odhadů s využitím navrhovaného aparátu. Druhou oblastí bude náznak některých neotřelejších postupů. Příspěvek však bude z důvodů přehlednosti členěn podle klasických prostředků teorie informace, obě oblasti se budou prolínat.

2 Výběrová entropie

Mějme abecedu $X = \{x_a, x_b, \ldots, x_m\}$ - soustavu elementárních jevů a n nezávislých pozorování $x_1, x_2, \ldots, x_n$ - „produkce“ zdroje s touto abecedou. Nechť je takový zdroj stacionární s pravděpodobnostmi výskytu jednotlivých písmen $p(x)$. Toto pravděpodobnostní rozdělení je nám nedostupné a je modelováno námi předpokládaným rozdělením $e(x)$. Potom budeme za výběrovou entropii $H_n(X; e)$ považovat výraz:

$$H_n(X; e) = -\frac{1}{n} \sum_{i=1}^{n} \lg \Big(e(x_i) \Big).$$

Ten lze jinak psát:

$$H_n(X; e) = - \sum_{x \in X} \frac{n(x)}{n} \lg \Big(e(x) \Big),$$

kde $n(x)$ je počet pozorování písmene (jevu) x. S rostoucím n výběrová entropie $H_n(X; e)$ konverguje k výrazu $-\sum_{x \in X} p(x) \lg(e(x))$. Konvergence je ve smyslu pravděpodobnosti [4] (zákon velkých čísel). „Limitní výběrovou entropií“ budeme pak rozumět:

$$H(X; e) = - \sum_{x \in X} p(x) \lg \Big(e(x) \Big).$$

Tento pojem není originální. Je znám pod názvem „Cross Entropy“, např. [7]. Nejprve vlastnosti:
1. $H(X; e)$ je konvexním funkcionálem $e()$.
2. $H(X; e)$ je spojitým funkcionálem $e(), \forall e()$, pro které je $e(x) > 0, \forall x \in X$.
3. $H(X; e)$ nabývá svého minima pro $e() = p()$, pokud $p(x) > 0, \forall x \in X$.

Vlastnost 1. je zřejmá, je důsledkem konvexity funkce $-\lg()$. Vlastnost 2. dokážeme pomocí konceptu sub-derivace ve směru.

Rozdělení $e_{\lambda,g}(x) = (1 - \lambda)e(x) + \lambda g(x)$ budeme pro $0 \leq \lambda \leq 1$ nazývat (λ) variací rozdělení $e()$ ve „směru“ $g()$.

Sub-derivací $H(X; e)$ ve směru g budeme rozumět výraz [2], [3]:

$$H'_g(X; e) = \lim_{\lambda \to 0_+} \frac{H(X; e_{\lambda,g}) - H(X; e)}{\lambda}.$$

Potom:

$$\begin{aligned} H'_g(X; e) &= \lim_{\lambda \to 0_+} \frac{d}{d\lambda} H(X; e_\lambda) = \lim_{\lambda \to 0_+} \left\{ - \sum_{x \in X} p(x) \frac{g(x) - e(x)}{e_{\lambda,g}(x)} \right\} = \\ &= - \sum_{x \in X} p(x) \frac{g(x) - e(x)}{e(x)}. \end{aligned}$$

Derivace je definována pro všechny modely $e()$, pro které je $e(x) > 0$, $\forall x \in X$. Tedy je i funkcionál $H(X;e)$ pro takové modely spojitý v $e()$ [2], [3]. Vlastnost 3. je důsledkem obou předchozích a faktu, že: $H'_g(X;p) = 0$ pro taková $p()$, pro které je $p(x) > 0, \forall x \in X$. Požadavek nenulovosti je, u diskrétní verze, přirozený. Mohou s ním však být problémy u odhadů při nulovém počtu pozorování jevu (písmene), u kterého je předpoklad nenulové pravděpodobnosti (zero frequency problem). Minimalita $H(X;e)$ pro $e() = p()$ vede na následující heuristiku při odhadování:

$$\text{Hledáme takové } \mathrm{e}^*(), \text{ pro které je } \mathrm{H_n(X;e^*)} = -\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} \lg\Big(\mathrm{e}^*(\mathrm{x_i})\Big) \text{ minimální.}$$

To však není nic jiného než metoda maximální věrohodnosti, protože, pokud takové $e^*()$ existuje, pak:

$$-\tfrac{1}{n}\sum_{i=1}^{n} \lg\Big(e^*(x_i)\Big) \le -\tfrac{1}{n}\sum_{i=1}^{n} \lg\Big(e(x_i)\Big) \Leftrightarrow$$

$$\Leftrightarrow \sum_{i=1}^{n} \lg\Big(e^*(x_i)\Big) \ge \sum_{i=1}^{n} \lg\Big(e(x_i)\Big) \Leftrightarrow \prod_{i=1}^{n} e^*(x_i) \ge \prod_{i=1}^{n} e(x_i).$$

A to je velice užitečný prostředek, zvláště pro parametrické odhadování. Dokonce lze i pomocí klasického diferenciálního počtu nalézt (v případě konečného souboru elementárních jevů) odhad $\hat{e}(x)$ minimalizující $H_n(X;e)$. Tedy hledáme čísla $\hat{e}(x_a), \hat{e}(x_b), \ldots, \hat{e}(x_m)$, pro která platí: $\sum_{x\in X} \hat{e}(x) = 1$ a $H_n(X;\hat{e}) \to \min$. Potom s použitím metody Lagrangeových multiplikátorů se jedná o nalezení minima:

$$Q\{\hat{e}(x_a), \hat{e}(x_b), \ldots, \hat{e}(x_m), \psi\} = -\sum_{x\in X} \frac{n(x)}{n} \lg\Big(\hat{e}(x)\Big) + \psi\Big(1 - \sum_{x\in X} \hat{e}(x)\Big),$$

pak

$$\frac{\partial Q}{\partial \hat{e}(x)} = -\frac{n(x)}{n}\frac{1}{\hat{e}(x)} - \psi \quad \text{a} \quad \frac{\partial^2 Q}{\partial \hat{e}^2(x)} = \frac{n(x)}{n}\frac{1}{\hat{e}^2(x)} \ge 0, \text{ pokud } \hat{e}(x) > 0.$$

Řešením rovnic:

$$\frac{\partial Q}{\partial \hat{e}(x)} = 0 \text{ a } \sum_{x\in X} \hat{e}(x) = 1 \text{ dostaneme } \hat{e}(x) = \frac{n(x)}{n}, \text{ pokud } n(x) > 0.$$

A to není nic jiného než klasický frekvenční odhad pravděpodobnosti za předpokladu, že pro všechny možné jevy (písmena) máme alespoň jedno pozorování.

3 Výběrová divergence a informace ve speciálních případech

Divergence pravděpodobnostních modelů je definována [1], [2]:

$$D(p\|s) = \sum_{x\in X} p(x) \cdot \lg \frac{p(x)}{s(x)}.$$

Z obdobných důvodů jako výše, můžeme zavést:

$$D_n(p\|s;e) = \sum_{x\in X} \frac{n(x)}{n} \cdot \lg \frac{e(x)}{s(x)}$$

výběrovou divergenci modelu $e()$ proti modelu $s()$ a její limitní formu:

$$D(p\|s;e) = \sum_{x\in X} p(x) \cdot \lg \frac{e(x)}{s(x)}.$$

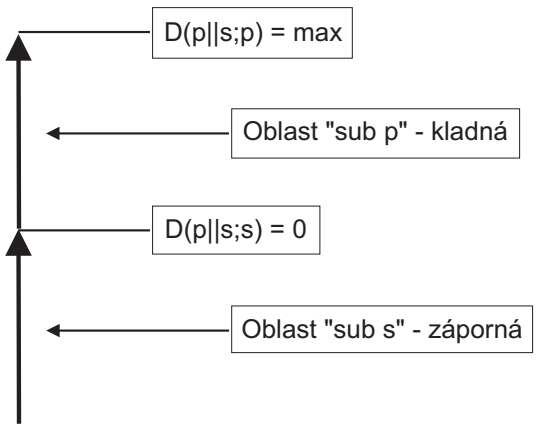
Za použití shodných prostředků jako u entropie lze dokázat obdobné vlastnosti:
1. $D(p\|s;e)$ je konkávním funkcionálem $e()$.
2. $D(p\|s;e)$ je spojitým funkcionálem $e(), \forall e()$, pro které je $e(x) > 0, \forall x \in X$.
3. $D(p\|s;e)$ nabývá svého maxima pro $e() = p()$, pokud $p(x) > 0, \forall x \in X$.
4. $D(p\|s;s) = 0$.
Vlastnosti 1. a 4. jsou triviální. Vlastnosti 2. a 3. jsou dokazatelné opět za použití sub-derivace ve směru:

$$D'_g(p\|s;e) = \lim_{\lambda\to 0_+} \frac{D(p\|s;e_{\lambda,g}) - D(p\|s;e)}{\lambda} = \sum_{x\in X} p(x)\frac{g(x)-e(x)}{e(x)}.$$

Ta je až na znaménko shodná se sub-derivací u limitní výběrové entropie. To je přirozené, neboť:

$$\begin{aligned} D(p\|s;e) &= \sum_{x\in X} p(x)\lg\frac{e(x)}{s(x)} = \sum_{x\in X} p(x)\lg\Big(e(x)\Big) - \sum_{x\in X} p(x)\lg\Big(s(x)\Big) = \\ &= -H(X;e) - \sum_{x\in X} p(x)\lg\Big(s(x)\Big). \end{aligned}$$

Jak bylo uvedeno v práci [8], uvedené vlastnosti umožňují klasifikovat možné modely $e()$ ve vztahu ke srovnávacímu $s()$ podle následující stupnice:



Samozřejmě se též nabízí problém nalezení „nejvíc odlišného“ modelu od $s()$. Je však namístě poznamenat, že divergence není metrikou.

To můžeme naznačit následující úlohou: Hledáme sdružené rozdělení pravděpodobnosti $p(x, y)$ za předpokladu, že jsou dány a známy jeho marginály $p(x)$ a $p(y)$, a to takové, že „maximalizuje“ divergenci „od“ sdruženého rozdělení za předpokladu nezávislosti $p(x)p(y)$. Tedy:

$$\sum_{x\in X}\sum_{y\in Y}\frac{n(x,y)}{n}\lg\frac{p(x,y)}{p(x)p(y)}\xrightarrow{p(x,y)}\max$$

při vazbách:

$$p(x)=\sum_{y\in Y}p(x,y);\forall x\in X\qquad\text{a}\qquad \mathrm{p(y)}=\sum_{\mathrm{x\in X}}\mathrm{p(x,y)};\forall \mathrm{y}\in \mathrm{Y}.$$

Řešení tohoto problému klasickou cestou vede opět na využití Lagrangeových multiplikátorů:

$$Q(p(,),\alpha,\psi)=\sum_{x\in X}\sum_{y\in Y}\frac{n(x,y)}{n}\lg\frac{p(x,y)}{p(x)p(y)}+$$

$$+\sum_{x\in X}\alpha_x\Big(p(x)-\sum_{y\in Y}p(x,y)\Big)+\sum_{y\in Y}\psi_y\Big(p(y)-\sum_{x\in X}p(x,y)\Big).$$

Z toho dostaneme rovnici „sedlových bodů“:

$$0=\frac{\partial Q}{\partial p(x,y)}=\frac{n(x,y)}{np(x,y)}-\alpha_x-\psi_y$$

a pro matici druhých parciálních derivací (přesněji její diagonálu):

$$\frac{\partial^2 Q}{\partial p^2(x,y)}=-\frac{n(x,y)}{np^2(x,y)}\,,$$

tedy v případných sedlových bodech bude maximum. Řešením rovnice sedlových bodů dostaneme:

$$p(x,y)=\frac{n(x,y)}{n(\alpha_x+\psi_y)}\,.$$

Hodnoty α_x, ψ_y zjistíme numerickým řešením vazebních rovnic:

$$p(x)=\frac{1}{n}\sum_{y\in Y}\frac{n(x,y)}{(\alpha_x+\psi_y)},\forall x\in X;\quad p(y)=\frac{1}{n}\sum_{x\in X}\frac{n(x,y)}{(\alpha_x+\psi_y)},\forall y\in Y,$$

pokud takové řešení existuje.

Uvedenou úlohu lze modifikovat, např.:

$$\sum_{x\in X}\sum_{y\in Y}\frac{n(x,y)}{n}\lg\frac{p(x,y)}{p(x)p(y)}\xrightarrow{p(x,y)}\max,$$

$$\text{při vazbě}: p(x)=\sum_{y\in Y}p(x,y), \forall x\in X.$$

Tj. příliš nám nezáleží na splnění podmínek:

$$p(y)=\sum_{x\in X}p(x,y), \forall y\in Y.$$

Pak dostaneme řešení pro odhad $p(x,y)$ ve tvaru $p(x,y)=\dfrac{n(x,y)}{n(x)}p(x)$

a odpovídající marginální pravděpodobnost $\bar{p}(y)=\sum_{x\in X}\dfrac{n(x,y)}{n(x)}p(x)$.

Tato řešení jsou vlastně klasickým (frekvenčním) odhadem podmíněné pravděpodobnosti $p(y/x)$.

Ještě volnější modifikace je:

$$\sum_{x\in X}\sum_{y\in Y}\frac{n(x,y)}{n}\lg\frac{p(x,y)}{p(x)p(y)}\xrightarrow{p(x,y)}\max.$$

Tj. vůbec nám nezáleží na dodržení marginálů. Pak dostaneme řešení pro odhad $p(x,y)$ ve tvaru:

$$p(x,y)=\frac{n(x,y)}{n}$$

a tomu odpovídající marginální pravděpodobnosti:

$$\bar{p}(x)=\frac{n(x)}{n};\qquad \bar{p}(y)=\frac{n(y)}{n}.$$

A to je vlastně „nejklasičtější“ frekvenční tvar odhadu.

4 Závěr

Předložená práce má ambici prokázat, že nastíněné pojetí „výběrových“ entropií, divergencí a informací spolu s jejich limitními tvary je konzistentní s některými klasickými technikami odhadování. Nejen to, ukazuje se, že je v něm možná skryt další, vlastní, potenciál pro rozvoj. Nakolik je tento dojem přesvědčivý i mimo skupinu autorů, ponecháváme na čtenáři.

Reference

[1] Vajda I. (1982). *Teória informácie a štatistického rozohodovania*. Alfa, Bratislava.

[2] Cover T.M., Thomas J.A. (1991). *Elements of information theory*. Wiley.

[3] Hiriart-Urruty J.B., Lemaréchal C. (2001). *Fundamentals of convex analysis*. Springer.

[4] Rényi A. (1972). *Teorie pravděpodobnosti*. Academia, Praha.

[5] Cziszár I. (1998). *Information theoretic methods in probability and statistics*. Conference Fifty Years of Shannon Theory, 1998 and Shannon Lecture ISIT'97.

[6] Cziszár I., Körner J. (1986). *Information theory, coding theorems for discrete memoryless systems*. KIADÓ, BUDAPEST, 1986.

[7] Zhai C.X. (2004). *Essential probability and statistics*. Department of Computer Science, University of Illinois, Urbana-Champaign, Lecture for CS397-CXZ Algorithms in Bioinformatics.

[8] Vávra F., Nový P. (2004). *Informace a dezinformace. Seminář z aplikované matematiky*. Katedra aplikované matematiky, Přírodovědecká fakulta Masarykovy Univerzity, Brno, 2004.

Poděkování: This work was supported by the grant of the Ministry of Education of the Czech Republic No: MSM-235200005 - Information Systems and Technologies.

Adresa: M. Kotlíková, H. Mašková, A. Netrvalová, P. Nový, D. Spíralová, F. Vávra, D. Zmrhal, Západočeská univerzita v Plzni, Fakulta aplikovaných věd, Katedra informatiky a výpočetní techniky, Univerzitní 8, 306 14 Plzeň

E-mail: vavra@kiv.zcu.cz

PRAVDĚPODOBNOST A MATEMATICKÁ STATISTIKA V INFORMATICKÝCH OBORECH

Alena Koubková, Jaroslav Král

Klíčová slova: Teorie pravděpodobnosti, matematická statistika, informatika.

Abstrakt: Cílem tohoto příspěvku je podnítit diskusi o použití pravděpodobnostních a statistických metod v informatice, o spolupráci informatiků a statistiků při řešení teoretických i praktických úloh z nejrůznějších oblastí informatiky a hlavně o výuce pravděpodobnosti a statistiky pro studenty informatických oborů.

1 Kde mohou informatici uplatnit pravděpodobnost a statistiku?

Pokud si někdo pod označením informatik představuje jen rutinního programátora známých algoritmů, správce serveru nebo tvůrce webových stránek, pak bude celkem oprávněně tvrdit, že informatika se bez pravděpodobnosti a statistiky obejde. Musíme bohužel konstatovat, že tuto představu má nejen většina laické veřejnosti, ale i nezanedbatelná část studentů, kteří se na naší fakultě na tento obor hlásí. Práce informatika ovšem zahrnuje i analýzy problémů, které se mají řešit, návrhy nových algoritmů a zkoumání jejich vlastností, testování programových systémů v různých podmínkách provozu atd. A zde už se prostor pro pravděpodobnostní a statistické metody najde.

Následující přehled není v žádném případě vyčerpávající a ani dostatečně reprezentativní. Vycházíme pouze z našich osobních zkušeností a víme, že řada našich kolegů by ho mohla doplnit o zajímavé příklady ze své praxe.

Příklad 1: Teoretická analýza algoritmů
Analýza algoritmů se kromě důkazů správnosti zabývá také odhadováním rychlosti a paměťové náročnosti výpočtu (tzv. *časové a prostorové složitosti*) v závislosti na rozsahu vstupních dat. Jde tedy vlastně o nalezení nějaké funkce f takové, že doba výpočtu (resp. velikost použité paměti) pro data velikosti n je $f(n)$. Protože hledaná funkce by měla mít obecnou platnost bez ohledu na to, v jakém jazyce bude daný algoritmus naprogramován a na jakém konkrétním počítači bude výpočet realizován, měří se časová složitost ne v jednotkách času, ale počtem provedených operací. Přitom aditivní i multiplikativní konstanty se ve výpočtech zanedbávají, protože asymptoticky pro velká n nehrají podstatnou roli. Je zřejmé, že jeden a tentýž algoritmus může pracovat různě rychle pro dvě stejně velké, ale jinak uspořádané množiny vstupních dat. Z toho důvodu se rozlišuje *složitost v nejhorším případě* a *očekávaná složitost*. Zatímco v prvním případě je hodnota $f(n)$ dobou výpočtu pro nejhorší možná data, tj. dobou, která pro ostatní data dané velikosti

rozhodně nebude překročena, ve druhém případě je to střední (očekávaná) hodnota doby výpočtu, která závisí na pravděpodobnostním rozložení vstupních dat. Ta by měla být doplněna ještě alespoň o rozptyl nebo odhady pravděpodobností velkých odchylek. Jenže tyto výpočty mohou být i pro velmi jednoduché algoritmy komplikované, a to i tehdy, když předpokládáme, že vstupní data mají rovnoměrné rozdělení. Často se ani samotnou střední hodnotu nepodaří spočítat přesně a musíme se spokojit s nějakým jejím horním či dolním odhadem. Rozptyl nebo další momenty se většinou z důvodu obtížnosti ani nepočítají. Řadu ukázkových analýz tohoto typu lze najít např. v monografiích [2], [3], [17]. Je z nich vidět, že člověk, který má podobné výpočty provádět, musí být schopný matematik a rozhodně nevystačí s pouhou definicí střední hodnoty nebo vytvořující funkce. Musí to tedy být buď informatik s rozsáhlými matematickými znalostmi, nebo matematik se zájmem o informatické problémy. Teoretická informatika, jejíž částí je i analýza algoritmů, má vůbec k matematice velmi blízko a pro člověka zabývajícího se pravděpodobností by mohla být docela zajímavou aplikací.

Příklad 2: Randomizované algoritmy

Takzvané *randomizované* (nebo také *pravděpodobnostní)* algoritmy se od klasických *deterministických* liší v tom, že v některých krocích provádějí náhodnou volbu, která určuje další pokračování výpočtu. Typickým příkladem jsou třeba náhodné procházky na grafech. Tato náhodná volba může v některých případech výpočet urychlit, v jiných zpomalit. Podstatné je ovšem to, že spustíme-li takový algoritmus několikrát na naprosto stejných datech, dostaneme pokaždé jiný průběh a tím i jinou dobu výpočtu. U randomizovaných algoritmů má tedy smysl pouze očekávaná složitost, která závisí na pravděpodobnostním rozdělení náhodně generovaných hodnot. Některé pravděpodobnostní algoritmy mají navíc tu vlastnost, že v některých případech můžeme dostat naprosto chybný výsledek. Je jasné, že aby takový algoritmus byl vůbec použitelný, musíme umět odhadnout pravděpodobnost takové chyby a tato pravděpodobnost musí být zanedbatelná. Pravděpodobnostní algoritmy se často používají k řešení velmi obtížných úloh, kdy přesný výpočet deterministickým algoritmem by trval neúnosně dlouho. Z pravděpodobnostních metod se při analýze randomizovaných algoritmů uplatní kromě různých technik výpočtu střední hodnoty a dalších momentů také markovské procesy. Více se lze dočíst např. v monografii [14].

Příklad 3: Experimentální algoritmika

Experimentální algoritmika je empirickým protějškem teoretické analýzy algoritmů. Má několik cílů. Předně chce navržené algoritmy uvést do praxe, to znamená vytvořit efektivní implementace a ve formě programových knihoven je nabídnout uživatelům. Dále prakticky prověřit chování algoritmů nebo pomocí experimentů odhadnout očekávanou složitost v případech, kdy ji teoreticky spočítat neumíme. Výsledky experimntů pak mohou zpětně pomáhat teoretické analýze. V neposlední řadě je úkolem experimentální algoritmiky vytvořit knihovny testovacích dat pro experimentování s algoritmy a posloužit tak tvůrcům algoritmů i jejich uživatelům.

Experimentální algoritmika nabízí možnost prověřit chování algoritmů pro nejrůznější typy rozdělení vstupních dat, pokud tato data umíme nasimulovat. Při vyhodnocování výsledků experimentů jsou použitelné např. statistické testy o střední hodnotě a dalších charakteristikách a zejména regrese při snaze experimentálně určit funkční závislost doby výpočtu algoritmu na velikosti vstupních dat. Aktuální jsou otázky o potřebném počtu provedených měření, aby získané výsledky byly statisticky průkazné, uplatnilo by se i plánování experimentů. Vše bývá náležitě komplikováno tím, že data získaná při experimentech většinou nesplňují ideální předpoklady běžně známých statistických metod, zejména všudypřítomnou nezávislost a normalitu rozdělení, a je zde tedy opět zapotřebí zkušenějšího statistika. Statistické povědomí je nutné i při závěrečné interpretaci výsledků.

Experimentální algoritmika je poměrně velmi mladou disciplínou, její rozvoj byl umožněn zejména možností publikovat v elektronické formě a získávat programové produkty a datové soubory po internetu. Zájemce o tuto problematiku odkazujeme např. na časopis Journal of Experimental Algorithmics vydávaný Association for Computing Machinery (ACM) nebo na publikace [11] a [13].

Příklad 4: Datové struktury a databáze
Teoreticky patří tento obor do analýzy algoritmů, zabývá se ale jen speciálními algoritmy, které umožňují vyhledávání v datech, vkládání nových záznamů, mazání starých, třídění apod. Při různém způsobu uložení dat v počítači dosahují tyto algoritmy různé efektivity. Úkolem tedy je najít takový způsob reprezentace dat, který je optimální pro nejčastěji prováděné operace. Mírou časové složitosti algoritmů pracujících nad datovými strukturami uloženými v interní paměti počítače je opět počet provedených operací (především porovnávání klíčů), u algoritmů pracujících nad databázemi uloženými na externích médiích je to počet přenosů datových bloků mezi externí pamětí (diskem) a interní pamětí, protože tato operace je časově nejnáročnější. Problémy spojené s výpočtem střední hodnoty této míry jsou tytéž jako v Příkladu 1. Datovými strukturami a jejich analýzou se zabývá např. monografie [10].

Příklad 5: Dokumentografické informační systémy
V těchto systémech se pracuje s databázemi speciálního typu. Záznamy v nich nejsou přísně strukturované, ale jsou to *texty*. Proto se v nich nevyhledává podle hodnot nějakého klíče, ale podle jakéhosi *dotazu* uživatele. Příkladem jsou třeba rešeršní systémy, kde můžeme vyhledávat nejen podle jména autora nebo názvu publikace, ale i podle tématu nebo klíčových slov. Účelem ovšem není vyhledat všechny články, ve kterých se zadané heslo vyskytne alespoň jednou, protože většina z nich by se patrně ukázala jako irelevantní. Proto se hledají modely, v nichž se termínům v dokumentech i v dotazech přiřazují *váhy* podle důležitosti, a stanovují se *míry podobnosti* mezi dokumentem a dotazem. Na základě této podobnosti by měl systém s velkou pravděpodobností vyhledat *relevantní* dokumenty. Pravděpodobnost se zde používá jednak v procesu stanovení vah, ale i jako charakteristika úspěšnosti

práce systému. Konkrétně se používají např. metody bayesovské analýzy, bayesovské sítě a shluková analýza. Nejjednodušší pravěpodobnostní modely dokumentografického informačního systému lze najít např. v [16].

Příklad 6: Data mining

Je opravdu těžké si představit disciplínu, která by více vybízela informatiky a statistiky ke vzájemné spolupráci. Místo ní jsme ale leckdy svědky řevnivosti až nesnášenlivosti, pohrdání partnerem a snahy přivlastnit si tento obor a veškeré zásluhy o jeho rozvoj. Jako lidé, kteří se ve svém životě zabývali jak statistikou tak informatikou, máme možnost vidět společné i rozdílné znaky obou disciplín. Společná je především snaha najít vhodný model pro analyzovaná data, případně detekovat odchylky od tohoto modelu, hledat závislosti, trendy, předpovídat budoucí vývoj. Rozdílný je způsob získávání dat a hlavně jejich objem. Zatímco klasická statistika pracuje obvykle s výběrem nevelkého rozsahu, na jehož základě dělá závěry platné (s nějakou přípustnou statistickou chybou) pro celý základní soubor, data mining chce využít všechna dostupná data. Neuplatní se zde tedy příliš metody výběrových šetření nebo plánování experimentů. Před vlastní analýzou je nutno potřebné záznamy v databázích efektivně vyhledat, odfiltrovat z nich nadbytečné údaje, vyřadit záznamy neúplné nebo nějak poškozené a podobně, což je informatická záležitost. Kvůli obrovskému objemu vstupních dat se vhodnost statistické metody musí posuzovat i z hlediska časové náročnosti výpočtu, je třeba na ni pohlížet jako na kterýkoli jiný algoritmus, který má být analyzován a efektivně implementován, což je opět informatický problém. Naopak výsostně statistická je volba správné metody, ověření jejích předpokladů a závěrečná interpretace výsledků. Statistický a informatický přístup by se zde tedy měl vhodně doplňovat.

Příklad 7: Operační systémy

Z pohledu statistika není operační systém nic jiného než systém hromadné obsluhy, kde obslužným zařízením je *procesor* (nebo server, sdílená tiskárna apod.) a zákazníci jsou *procesy* (klientské terminály, požadavky na tisk). Úkolem je najít strategii režimu fronty a stanovování priorit, která by byla optimální jak z hlediska provozu systému, tak z hlediska čekajících uživatelů, což bývá často v přímém rozporu. O možném použití teorie front v operačních systémech pojednávají některé kapitoly [1] nebo [9].

Příklad 8: Počítačová grafika

V počítačové grafice se pravděpodobnostní metody dají použít např. při analýze a rekonstrukci digitálního obrazu. Na obraz se pohlíží jako na pole bodů (*pixelů*) různých barev oddělených *mikrohranami* a případně doplněných dalšími charakteristikami (např. příslušností k určité *textuře* apod). K obrazu bývá často přimíchán *šum*, který se tam může dostat třeba nekvalitním nasnímáním nebo při přenosu dat. Úkolem je tento šum odstranit a co nejlépe zrekonstruovat originál. Jinou úlohou může být popis a následné syntetické generování textur a mnoho dalších. Dají se použít různé osvědčené heuristiky, ale také např. bayesovská analýza nebo rovinná markovská pole. Zájemce o tuto problematiku odkazujeme na publikace [7], [8], [18].

Příklad 9: Tvorba softwarových systémů
Statistika se zde může uplatnit z několika důvodů. Předně i balíky statistických programů vytvářejí informatici, takže by měli mít alespoň nějakou povědomost o tom, co vlastně programují. V programech pro ekonomickou sféru se uplatní metody analýzy dat, v systémech pro podporu managementu třeba teorie rizik atd. Kromě toho, ať už je softwarový systém určen pro použití v jakékoli oblasti, bude vždy posuzován z hlediska spolehlivosti, jeho tvůrci tedy musí řešit takové otázky, jako kdy se ještě vyplatí programy dále upravovat a ladit a odkdy už je lepší je kompletně přepsat. V tom jim mohou pomoci znalosti z teorie spolehlivosti.

Příklad 10: Kryptografie
Kryptografie již dávno není pouze předmětem zájmu tajných služeb, ale stává se záležitostí doslova každého z nás. Problematika elektronických podpisů, internetové bankovnictví či ochrana dat všeho druhu si zasluhuje nejvyšší pozornost. Kryptografie sice stojí hlavně na teorii čísel, ale i statistické metody se zde uplatní, a to hlavně při *náhodném generování klíčů*. V kryptografii se nedají použít jednoduché kongruenční generátory, protože jsou málo náhodné a snadno dešifrovatelné. Navíc klíčem by obvykle mělo být hodně velké prvočíslo, řádově o stovkách cifer. Protože není možné tak velké prvočíslo efektivně spočítat, generuje se posloupnost cifer náhodně a výsledek se podrobuje testům, zda je to skutečně prvočíslo či ne. Řada těchto testů jsou randomizované algoritmy. Dalším problémem je důkladná pravděpodobnostní analýza algoritmů používaných v tzv. *veřejných kryptografických systémech*. Jde o to, aby algoritmy, pomocí nichž by mohla zpráva být i bez znalosti klíče dešifrována, měly exponenciální složitost nejen v nejhorším, ale i v průměrném případě. Více se lze dočíst např. v [15].

2 Jak to vypadá s výukou?

Před zhruba patnácti lety se studenti informatiky na MFF UK neučili pravděpodobnost a statistiku vůbec. Situace se od té doby výrazně změnila k lepšímu. V současné době mají studenti jednosemestrální přednášku z pravděpodobnosti a jednosemestrální přednášku ze statistiky, obojí se cvičeními, výuku zajišťuje KPMS. Tato výuka sice není striktně povinná, ale tyto předměty se až do letošního roku zkoušely u státnic jako součást všeobecného základu, takže je většina studentů navštěvovala. V začínajícím reformovaném studiu dokonce ještě jeden semestr přibude. V bakalářském studiu bude povinná jednosemestrová přednáška ze základů pravděpodobnosti a statistiky, na ni budou navazovat v magisterském studiu (alespoň pro některé obory) semestr pravděpodobnosti a semestr statistiky věnované pokročilejším partiím, které budou alespoň v některých specializacích povinné či povinně volitelné. Kromě toho existuje i několik výběrových přednášek zaměřených pouze na některé vybrané kapitoly potřebné pro daný obor a přednášek snažících se ukázat aplikace stochastických metod v různých oborech informatiky, které

si jednotlivé katedry informatické sekce zajišťují samy. Obecně ale požadavky z pravděpodobnosti a statistiky nebudou nadále součástí státních závěrečných zkoušek.

S počtem hodin i s tématy přednášek můžememe být celkem spokojeni, přesto podle nás výuka zcela nesplňuje svůj účel. Proč? Jeden důvod je, že je zařazena až do vyšších ročníků. V současnosti si ji studenti sice mohou zapsat kdykoli, ale z nejrůznějších důvodů to odkládají, dokud je k tomu okolnosti nedonutí. Těmi okolnostmi bývá nejčastěji to, že na něco ze statistiky narazí při práci na diplomce. K tomu dojde až na samém konci studia, takže to, co se naučí, už nemohou použít v jiných odborných předmětech a často se ani nedozvědí, kde všude by se jim to mohlo hodit. Učitelé řady odborných předmětů se dostávají do situace, že potřebné partie z pravděpodobnosti si buď musí sami odpřednášet (na úkor něčeho jiného a často i v různých předmětetch opakovaně), nebo se těm tématům, kde by je potřebovali, prostě vyhnout. Najde se i pár dobrodruhů, kteří předpokládají, že to studenti znají, a když ne, že mají možnost se to naučit v předmětech k tomu určených, ale ti většinou spláčou nad výsledkem, protože pro studenty se pak jejich přednáška stane nesrozumitelnou (a neoblíbenou). V reformovaném studiu se situace poněkud zlepší u magistrů, ale bakaláři na tom budou prakticky stejně, protože pravděpodobnost je zařazena až do třetího, tj. posledního ročníku jejich studia. Druhým důvodem k nespokojenosti je, že studenti většinou považují statistiku za nezajímavou a hlavně pro ně nepotřebnou.

Protože současné studium už tak jako tak dobíhá, bavme se dále pouze o reformovaném studiu. Přeřadit výuku pravděpodobnosti a statistiky do druhého ročníku se zdá být momentálně neprůchodné. Počet vyučovacích hodin v nižších ročnících je pevně daný a je třeba do nich doslova nacpat množství odborných předmětů, bez kterých studenti nemohou začít pracovat na závěrečném projektu. Taková je alespoň momentální představa a teprve dalších několik let ukáže, zda je správná a jediná možná. Teorie (a to nejen pravděpodobnost) se zkrátka z tohoto hlediska jeví jako postradatelný luxus. Přednáška z pravděpodobnosti pro bakaláře tak bude mít význam hlavně pro ty z nich, kteří budou ve studiu pokračovat, pokud s ní ovšem něco neuděláme.

Nelze-li posunout výuku pravděpodobnosti a statistiky do nižších ročníků, tak, aby bylo možno studentům v odborných předmětech ukázat její využití, nezbývá zřejmě nic jiného, než zařadit příklady aplikací v informatice přímo do této přednášky. Namísto příkladů "ze života", kde se statisticky vyhodnocuje třeba vývoj kojenecké úmrtnosti, znečištění životního prostředí nebo závislost barvy očí a barvy vlasů, je třeba volit příklady z oborů, které naši studenti studují. To bude vyžadovat úzkou spolupráci učitelů z obou stran, vybrat vhodné příklady totiž vůbec nebude lehké. Bylo by vhodné, aby zvolené příklady pokrývaly co nejvíce informatických disciplín, bude tedy zapotřebí spolupráce většího počtu lidí. Navíc bude nutná synchronizace s odbornými předměty, aby nedocházelo k opačnému extrému, že by se v prav-

děpodobnosti počítal příklad z oboru, který studenti ještě neměli, a naopak statistik by jim musel vysvětlovat jeho základy. A kromě toho, opravdu jednoduchých učebnicových příkladů nebude nijak mnoho. Ideální by bylo mít časem skripta nebo učebnici pravděpodobnosti a statistiky přímo pro informatiky, podobnou, jako je ve svém oboru monografie [12]. V současné době bude ale zřejmě kamenem úrazu nedostatek času na obou stranách, protože rozbíhající se bakalářské studium s sebou přináší např. větší počet zkoušek, vedení a oponování většího počtu diplomových prací a v začátcích i podstatné změny ve vyučovaných předmětech.

3 Víme, co chceme?

K tomu, co zde o významu pravděpodobnosti a statistiky pro informatiku již bylo řečeno, dodejme ještě, že v poslední době zaznamenáváme úvahy o tom, jaký by vlastně měl být profil absolventa oboru informatika na matematicko-fyzikální fakultě (viz např. [4], [5], [6]). Jestli by to měl být v první řadě zručný programátor nebo spíš všestranněji vybavený člověk, který by se uplatnil při analýze problémů a návrzích jejich řešení, dovedl by se orientovat v množství již existujících programů a programových komponent a posoudit jejich výkonnost, spolehlivost i cenu, který by dovedl jednat s uživatelem a měl i základní manažerské schopnosti. Je jasné, že statistika může k všestrannosti informatika podstatně přispět. Schopnost odhadnout a porovnat parametry systémů nebo programů, určit spolehlivost, odhadnout rizika volby apod. se může jedině hodit.

Na druhé straně doufáme, že i statistici si uvědomují, že informatika pro ně neznamená jen to, že mohou mít na stole počítač s editorem pro psaní vědeckých publikací a s přístupem na internet, ale že ji chápou jako jednu z aplikací svého oboru. Že se o ni budou zajímat stejně jako o ekonomii, biologii nebo fyziku, že si někteří z nich osvojí terminologii a základy některého z informatických oborů a budou řešit problémy, které mohou být i ze statistického hlediska zajímavé. V některých situacích by zcela jistě bylo jednodušší vysvětlit statistikovi několikařádkový algoritmus, na jehož pravděpodobnostní analýzu by mohl vynaložit všechny své zkušenosti získané léty studií a vlastní odborné práce, než učit informatika pět let statistiku, aby si tu analýzu mohl udělat sám.

Co říci na závěr? Snad jen to, že doufáme, že diskuse na toto téma bude pokračovat a spolupráce statistiků a informatiků se bude prohlubovat k užitku všech.

Reference

[1] Hansen P.B. (1979). *Principy operačních systémů.* STNL Praha.

[2] Hofri M. (1987). *Probabilistic analysis of algorithms.* Springer-Verlag, New York, Berlin, Heidelberg.

[3] Knuth D. (1973). *The art of computer programming.* Sorting and Searching **3**, Addison-Wesley, Reading, Massachusetts.

[4] Král J., Töpfer P. (2000). *Education of software experts for changing world.* 2^{nd} Global Congress on Engineering Education, Wismar, Germany, Z. Pudlowski ed. UICEE. 267 – 271

[5] Král J., Žemlička M. (2004). *What literacy for software developers.* (Manuscript).

[6] Král J., Žemlička M. (2004). *Mathematical statistics in SW engineering education.* Information Technology and Organizations: Trends, Issues, Challenges and Solutions, M. Khosrow–Pour ed. Idea Group Publishing, Hershey, PA, U.S.A. 889 – 890.

[7] Statistics and Images: 1. (1993). Advances in Applied Statistics, Carfax Publ. Comp. K. V. Mardia and G. K. Kanji eds.

[8] Statistics and Images: 2. (1994). Advances in Applied Statistics, Carfax Publ. Comp. K. V. Mardia ed.

[9] McDermid J.A. (1991). *Software engineer's reference book.* Butterworth-Heinemann, Oxford, London.

[10] Mehlhorn K. (1984). *Data structures and algorithms 1: Sorting and searching.* Springer-Verlag, Berlin, Heidelberg, New York.

[11] Mehlhorn K., Näher S. (1999). *The LEDA platform of combinatorial and geometric computing.* Cambridge Univeristy Press.

[12] Meloun M., Militký J. (1994). *Statistické zpracování experimentálních dat.* PLUS, Praha.

[13] Moret B.M.E. (2002). *Towards a discipline of experimental algorithmic.* DIMACS Series in Discrete Mathematics and Theoretical Computer Science **59**, 197 – 213.

[14] Motwani R., Raghavan P. (1995). *Ranomized algorithms.* Cambridge University Press.

[15] Neuenschwander D. (2004). *Probabilistic and statistical methods in cryptology: An introduction by selected topics.* Lecture Notes in Comp. Sci. **3028**, Springer-Verlag, Berlin, Heidelberg, New York, Tokyo.

[16] van Rijsbergen C.J. (1979). *Information retrieval* (second ed.). London, Butterworths.

[17] Sedgewick R., Flajolet P. (1996). *An introduction to the analysis of algorithms.* Addison-Wesley, Reading, Massachusetts.

[18] Winkler G. (1995). *Image analysis, random fields and dynamic Monte Carlo methods.* Springer-Verlag, Berlin, Heidelberg.

Adresa: A. Koubková, J. Král, Katedra softwarového inženýrství, Matematicko-fyzikální fakulta UK, Malostranské nám. 25, 118 00 Praha 1

E-mail: `koubkova@ksi.ms.mff.cuni.cz`,
`kral@ksi.ms.mff.cuni.cz`

A STATISTICAL PROPOSAL FOR SEQUENTIAL CLINICAL TRIALS IN DIFFERENT CANCER LOCATIONS

A. Koubková, F.T. Barbosa, G. Molenberghs

Keywords: Phase II clinical trials, decision function, response rate distribution, cancer location.

Abstract: This work concerns on phase II cancer clinical trials design, which create an important step in drug testing, before it can be used in practice. We specialize on anticancer drugs.

Since different cancers are similar, the idea of this wirk is to use to estimate the activity of the drug in some specific cancer based on the results of testing the drug in another cancer. The main aim is then to determine an optimal sequence of the cancer locations, how they should be tested, in order to gain as much as possible.

1 Introduction

This contribution is about drug testing and it is a summary of the prepared paper [3].

When a new drug is developed, it needs to come through a number of tests before it is accepted for use in clinical practice. First are the chemical tests in laboratories, then come tests on animals and the last stage consists of tests on humans, so called clinical trials. There are four stages (phases) of clinical trials. In phase I clinical trials the toxicity of the new agent is screened, in phase II its efficacy is estimated, in phase III the drug is compared with the standard treatments and in phase IV rare side effects are monitored.

This work deals with phase II cancer clinical trials. The main aim is to estimate the efficacy of the drug described by its response rate. It is the number of patients for which the drug shows desired activity divided by the total number of patients treated. There are strong criteria on patients, who can enter the tests. The result of phase II clinical trial is either to recommend the drug for phase III trials or to reject it from any further study.

For most of the cancers, there exist some active drugs. These drugs determine the minimal response rate, based on which the new agent is evaluated as active or not. The new agent needs to be preferable than the standard one and it should show at least the same activity.

The phase II clinical trials are one arm studies, in which only moderate number of patients (about 30) with one specific cancer type is treated. Since different cancers behave similarly, one anticancer drug can be active against more of them. Herson [1] came with an idea of a multi-stage design including patients with different tumor types. He wanted to estimate the predictive

probabilities of response using some prior information on the response rate as well as on the degree of tumor non-specificity in response, which would be adjusted as the trial would go on.

In this work we use a similar idea, i.e. to involve the similarity between the cancers in prediction of the drug efficacy. We propose a series of sequentially conducted clinical trials in different tumor locations. To this purpose we develop a decision function determining order of the tumors in the sequence, so that one can gain as much as possible by conducting a trial in the first tumor type. This function takes into account similarity between the cancer types, similarity between the drugs, aggressivity and incidence of the cancers, penalty for treating a patient successfully or unsuccessfully and the results from the clinical trials finished with the new agent. The functional values are updated each time a trial with the new agent is finished.

2 Correlation matrices between the tumor types

Since no easy measure of similarity between the cancers is known, we estimate it by ourselves. There exist more than 100 different cancers. Since many of them are very rare, we used only 25 selected cancer types instead of the particular diseases. These are adrenal, bladder, brain, breast, cervix, colon, endometrium, esophagus, kidneys, acute lymphocytic leukemia (ALL), acute myelogenous leukemia (AML), chronic lymphocytic leukemia (CLL), chronic myelogenous leukemia (CML), liver, Hodgkin's lymphoma, non-Hodgkin's lymphoma, small cells lung, non-small cells lung, ovaries, pancreas, prostate, skin, stomach, testes, and urethra cancers.

We studied these cancer types from 4 different points of view based on which we calculated 5 correlation structures. Finally, we add one more matrix concerning the healthy body locations corresponding to the above cancers, instead of the diseases themselves.

The matrix assuming the healthy body organs names *Biological correlation matrix* and it is calculated as a sample correlation matrix based on 51 features like similarity in structure, cellular base and functions of the organs. Since about 90% of the cancers develop in the epithelial tissue, special emphasis is laid on it.

Next we create *Two matrices based on drug response*. The first one takes into account 99 different drugs and 105 drug combinations, whereas the second one concerns only the 99 drugs. For the first matrix, the activity of the drugs was described by a weighted average of response rates coming from the recent clinical trials. For the second one we use an indicator describing, whether the drug is used alone in the location, or it is used only in combinations, or it is not used at all.

All the cancers are caused due to mutations in genes involved in cell growth and division. There are two kinds of such mutations, the first one activates the genes for rapid growth (oncogenes), the second one inactivates the genes controlling the growth. The *Genetic correlation matrix* takes into

account properties of the mutations, their locations and expression of some markers.

The last two matrices, each concerns mutation only in one special gene. The *P53 correlation matrix* characterizes mutations in P53 gene, which controls the cell growth and division. The *EGFR correlation matrix* is calculated based on expression of Epidermal Growth Factor Receptor, which is involved in cellular growth.

Each of the proposed matrices shows quite different pattern. Based on *First correlation matrix based on drug response*, the tumors seem almost independent (the coefficients take values around 0.05). On the other hand, in the *EGFR correlation matrix*, there are more than 68 coefficients between different tumor types equal to one. The coefficients of the other matrices take values from quite wide range, but mostly they are around 0.2.

3 Decision function

Denote θ_i the true response rate of the new agent in the i-th tumor location, g_i, h_i the gain and loss for treating a patient with i-th tumor type successfully or unsuccessfully respectively and $G(i)$ the value of the gain function for i-th location. The appropriate gain function consists of three parts, and its idea is partially based on [2].

Part I. The first part of the function describes the gain obtained in the phase II trial. For phase II trials, it is reasonable to assume the same sample size in all the tumor locations and we choose $n_1 = 30$. In the trial performed in the i-th tumor location, there will be $\theta_i n_1$ patients treated successfully and $(1-\theta_i)n_1$ unsuccessfully, so the gain will be

$$G(\text{phase II}, i) = g_i\theta_i n_1 - h_i(1-\theta_i)n_1 = (g_i + h_i)\theta_i n_1 - h_i n_1.$$

Part II. With probability

$$\sum_{x=k_{1,i}+1}^{n_1} \binom{n_1}{x} \theta_i^x (1-\theta_i)^{n_1-x}$$

the new agent will enter the phase III trials. The critical values $k_{1,i}$ are calculated based on the equality $P_{\theta_{0,i}}\{\text{the drug is evaluated as ineffective}\} \leq \beta = 0.1$ describing the type II error.

In phase III trials the sample sizes are different for the particular tumor types and reflect their incidence. We assume the sample sizes from the range $[200, 600]$ and calculate them as $n_{2,i} = 200 + 1.81 \cdot (\text{incidence per year})/1000$. In the phase III trial conducted in the tumor type i, there will be $\theta_i n_{2,i}$ successfully treated patients and $(1-\theta_i)n_{2,i}$ unsuccessfully treated patients. So the gain will be

$$G(\text{phase III}, i) = g_i\theta_i n_{2,i} - h_i(1-\theta_i)n_{2,i} = (g_i + h_i)\theta_i n_{2,i} - h_i n_{2,i}.$$

Part III. To the clinical practice the new agent will come with probability

$$\sum_{x=k_{2,i}+1}^{n_{2,i}} \binom{n_{2,i}}{x} \theta_i^x (1-\theta_i)^{n_{2,i}-x}.$$

For calculation of the critical values $k_{2,i}$ we used the threshold response rates $\theta_{0,i}$ as before, but this time we determine them from the equality for the type I error rate (with $\alpha = 0.05$).

The numbers of patients treated with the new agent in clinical practice, $n_{3,i}$ can highly vary for the different tumor types and they will depend on the incidence of the diseases. We assume $n_{3,i} = 1/2 \cdot$ (incidence per year) yearly treated patients per year. It means that in the i-th tumor location, $\theta_i n_{3,i}$ patients will be treated successfully and $(1-\theta_i)n_{3,i}$ unsuccessfully per year. This leads to the last part of the gain function

$$G(\text{practice}, i) = g_i \theta_i n_{3,i} - h_i(1-\theta_i)n_{3,i} = (g_i + h_i)\theta_i n_{3,i} - h_i n_{3,i}.$$

The three parts together give the complete gain function. Since we don't know the true response rates θ_i, we need to estimate them. If we know the distributions of θ_i, we can use their expectations, or their sample means, or we can integrate the gain function over the parameter θ_i using its density.

Finally, the pure gain is often not of interest. Preferentially, the more aggressive cancers are treated first. So we add one more term to our gain function. This term is kt_i, where k is an appropriate constant used to influence the decision function in the desired way, and t_i is an aggressivity coefficient, for example the death rate. The final gain function is

$$\begin{aligned} G(i) \; &= (g_i + h_i)\theta_i n_1 - h_i n_1 + \left(\sum_{x=k_{1,i}+1}^{n_1} \binom{n_1}{x} \theta_i^x (1-\theta_i)^{n_1-x} \right) \\ &\times \left[(g_i + h_i)\theta_i n_{2,i} - h_i n_{2,i} + \left(\sum_{y=k_{2,i}+1}^{n_{2,i}} \binom{n_{2,i}}{y} \theta_i^y (1-\theta_i)^{n_{2,i}-y} \right) \right. \\ &\left. \times \left((g_i + h_i)\theta_i n_{3,i} - h_i n_{3,i} \right) \right] + k \cdot t_i. \end{aligned}$$

4 Multimodal response rate distribution

To evaluate the above decision function, the distributions of the response rates θ_i need to be estimated. The calculation is divided into two parts. In the first one we assume only the ability of the tumors to respond and the similarity between the drugs, based on which the basic prior distribution, $q(\theta_i)$ is determined. In the second part we adjust the basic prior for the information of the finished phase II trials performed with the new agent and we get the advanced response rate distribution $r(\theta_i|\mathbf{X})$, where $\mathbf{X}$ denotes the

results from the finished trials. The distribution $r(\theta_i|\mathbf{X})$ can be either prior or posterior depending on whether a trial in tumor type i was done or not.

Since the response rates θ_i are in fact proportions of the binomial distribution, their appropriate prior distribution is a beta distribution. The first part of the basic prior distribution of θ_i, $f_1(.)$ is calculated based on the ability of the tumor to respond. If the tumor didn't respond to any drug in the past, there is a high probability that it will not respond to the new agent neither. This suggests a beta distribution with the parameters $a < 1$ and $b > 1$ and since we want to have an uninformative prior, we choose $a = 0.8, b = 1.2$.

If the tumor responded in the past to any drug, denote the probability of responding (i.e. the number of active drugs divided by the number of all drugs) by η_i and the average response rate of the active drugs as $\bar{\theta}_i$. Then we can assume that the new agent will have the true response rate $\bar{\theta}_i$ with the probability η_i and the other values will be less probable. The distribution $f_1(\theta_i)$ is calculated as a solution of the system of two equations:

$$\begin{aligned} f_1'(\bar{\theta}_i) &= \frac{1}{B(a,b)}\bar{\theta}_i^{a-2}(1-\bar{\theta}_i)^{b-2}((a-1)(1-\bar{\theta}_i)-(b-1)\bar{\theta}_i) = 0, \\ f_1(\bar{\theta}_i) &= \frac{1}{B(a,b)}\bar{\theta}_i^{a-1}(1-\bar{\theta}_i)^{b-1} = 1+\eta_i, \end{aligned} \tag{1}$$

which correspond to the conditions that $f_1(\theta_i)$ should have mode at $\bar{\theta}_i$ and its functional value here should be $f_1(\bar{\theta}_i) = 1 + \eta_i$.

The second part of the basic prior distribution, $f_2(.)$, describes the similarity of the new agent to drugs with known response rates. Denote the correlation coefficient between the new agent and the known drug as D (if there are more similar known drugs, it is an average of the corresponding correlation coefficients) and the response rate of the similar known drug in tumor type i as $\theta_{D,i}$ (again, for more similar drugs, it is an averaged response rate). The distribution $f_2(\theta_i)$ should have mode at $\theta_{D,i}$ and its functional value here should be $f_2(\theta_{D,i}) = 1 + D$, which leads to the following system of equations determining $f_2(\theta_i)$

$$\begin{aligned} f_2'(\theta_{D,i}) &= \frac{1}{B(a,b)}\theta_{D,i}^{a-2}(1-\theta_{D,i})^{b-2}((a-1)(1-\theta_{D,i})-(b-1)\theta_{D,i}) = 0, \\ f_2(\theta_{D,i}) &= \frac{1}{B(a,b)}\theta_{D,i}^{a-1}(1-\theta_{D,i})^{b-1} = 1+D. \end{aligned} \tag{2}$$

The basic prior distribution is a standardized sum of $f_1(.)$ and $f_2(.)$, i.e.

$$q(\theta_i) = \frac{f_1(\theta_i)+f_2(\theta_i)}{\int_0^1 (f_1(\theta_i)+f_2(\theta_i))\,d\theta_i} = \frac{f_1(\theta_i)+f_2(\theta_i)}{2}.$$

If a trial is conducted in the tumor location i, then the basic prior $q(\theta_i)$ should be replaced by a basic posterior distribution

$$q(\theta_i|\mathbf{X}_i) = \frac{q(\theta_i)L(\mathbf{X}_i|\theta_i)}{\int_0^1 (q(\theta_i)L(\mathbf{X}_i|\theta_i))\,d\theta_i},$$

where $L(\mathbf{X}_i|\theta_i)$ is the likelihood from the trial.

These basic distributions should be adjusted by information of the trials conducted with the new agent in other tumor locations. We will add a part of the corrected likelihoods to the basic distributions. The amount added depends on the similarity between the tumor types. Assume that the trials were conducted in tumor types $l_1, \ldots, l_m$ and denote the correlation matrix between these tumors as $\mathbf{R}$. Next denote S_i the vector of correlation coefficients between the examined tumor type i and the tumors $l_1, \ldots, l_m$ already tested and calculate the product $\mathbf{R}S_i$. The likelihoods from the conducted trials should be corrected with respect of the signs of the terms in $\mathbf{R}S_i$. Denote the vector of corrected likelihoods as $\mathbf{L}^c = (L_1^c, \ldots, L_m^c)$, where $L_j^c = L(\theta_i|\mathbf{X}_j)$ if $\{\mathbf{R}S_i\}_j > 0$ and $L_j^c = 1 - L(\theta_i|\mathbf{X}_j)$ if $\{\mathbf{R}S_i\}_j < 0$. Then the advanced distribution of the response rate $r(\theta_i|\mathbf{X})$ is calculated as

$$r(\theta_i|\mathbf{X}) = \frac{q(\theta_i) + (\mathbf{L}^c)^T|\mathbf{R}S_i|}{\int_0^1 (q(\theta_i) + (\mathbf{L}^c)^T|\mathbf{R}S_i|)\,d\theta_i},$$

where $|\mathbf{R}S_i|$ denotes the absolute value of $\mathbf{R}S_i$ taken by terms and where $q(\theta_i)$ is replaced by $q(\theta_i|\mathbf{X}_i)$, when a trial in tumor type i was conducted.

5 Unimodal response rate distribution

The final estimated distribution proposed in the previous section is in general multimodal. Sometimes it is more convenient to have a unimodal one. In such a case, it should be a beta distribution with parameters $a_{F,i}$ and $b_{F,i}$, which need to be estimated. The estimation procedure consists of three parts.

Part I. Here is estimated the base of the distribution concerning the ability of the tumor types to respond. If the tumor i didn't respond in the past to any drug, the parameters of the beta distribution are $a_{B,i} = 0.8$ and $b_{B,i} = 1.2$. If the tumor responded to some chemical agent, the parameters $a_{B,i}$ and $b_{B,i}$ are calculated by solving the system of equations (1).

Part II. The parameters of a beta distribution, $a_{D,i}$ and $b_{D,i}$, calculated in this part describe the similarity of the new drug with the known drugs. These parameters are established by solving the system of equations (2).

Part III. This part concerns on the phase II clinical trials possibly conducted with the new agent. Assume the trials were performed in tumor types $l_1, \ldots, l_m$ and as before, denote by $\mathbf{R}$ the correlation matrix between them and by S_i the vector of correlation coefficients between them and the examined tumor type i. Next denote the sample size in the trial conducted in tumor location j as n_j and the number of responses observed there as r_j. Now create two vectors v_i and u_i of size m, such that $v_{i,j} = r_j$ and $u_{i,j} = n_j - r_j$

if the j-th term of $\mathbf{R}S_i$ is positive, and $v_{i,j} = n_j - r_j$ and $u_{i,j} = r_j$, if it is negative. The parameters $a_{T,i}$ and $b_{T,i}$ are then calculated as

$$a_{T,i} = |\mathbf{R}S_i|v_i, \qquad b_{T,i} = |\mathbf{R}S_i|u_i,$$

where $|\mathbf{R}S_i|$ denotes the vector of absolute values.

The parameters of the final response rate distribution are then simply the sums of the three parts:

$$a_{F,i} = a_{B,i} + a_{D,i} + a_{T,i}, \qquad b_{F,i} = b_{B,i} + b_{D,i} + b_{T,i}.$$

The only one exception is the case, where $a_{B,i} = a_{D,i} = 0.8$. Then the final parameters are $a_{F,i} = 0.8 + a_{T,i}$ and $b_{F,i} = 1.2 + b_{T,i}$.

6 Simulations

A small simulation study to show the properties of the above introduced procedure was conducted. A drug with the following response rates was assumed: 0.6 in breast cancer and $2|s_i|(0.5 + (\text{sign}s_i)(0.6 - 0.5))$ in the other cancers, where s_i are the correlation coefficients with breast cancer from the *Biological correlation matrix*. This drug was assumed to be dissimilar with any known drug and the *Biological correlation matrix* was taken for the calculations. The gains and losses $g_i = h_i = 1$, $g_i = 1 < h_i = 2$ and $g_i = 2 > h_i = 1$ for treating a patient were compared. Also the multimodal and unimodal response rate distributions were compared.

The simulation procedure was as follows. At first a tumor type for the first test was selected. Next a trial was simulated here and the decision function was updated to choose a tumor type for second trial. Then again a trial was simulated and the decision function was updated once more.

The procedure using equal gain and loss for treating a patient and the one preferring gain recommended the tumor types with high incidence. The third procedure gives some balance between the incidence and aggressivity of the disease. The main difference between the two types of the response rate distribution is that the unimodal distribution is much more informative than the multimodal one.

7 Conclusion

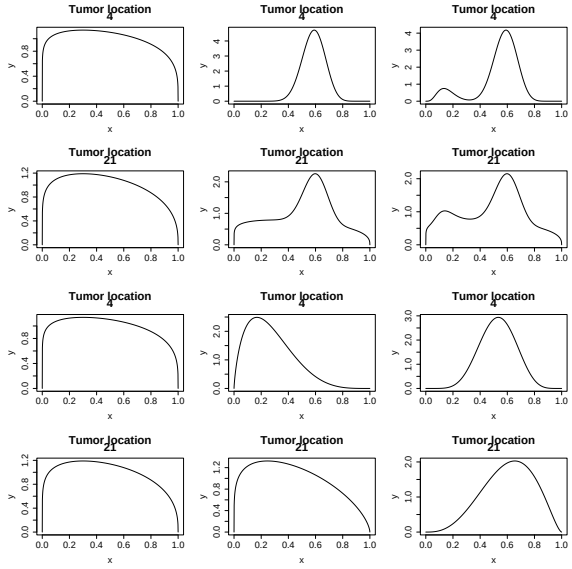
A decision procedure was proposed to determine an optimal order of tumor types for sequential phase II clinical trials with a new anticancer drug. At first a prior response rate distribution for each tumor type was estimated. It can be either unimodal or multimodal distribution. Using this distribution a gain function is evaluated and so the order in the tumor sequence determined.

The gain function takes into account properties of the diseases, such as its incidence, aggressivity, similarity with other tumors and evaluation for treating a patient. Also the characteristics of the new drug, such as its similarity with the known drugs or target, for which it is designed, are assumed. This function is updated each time a clinical trial with the new agent is finished.

The properties of the procedure are shown by a small simulation study, which compares the two types of response rate distribution and also different penalties for treating a patient.

Appendix

Here is a graph depicting how the distribution of the response rate is evolved. The first two lines correspond to the multimodal response rate distribution and the second two lines to the unimodal one. In both cases, the distributions of response rate in breast cancer and in prostate cancer is depicted.



References

[1] Herson, J (1979), *Predictive probability early termination plans for phase II clinical trials*. Biometrics **35**, 775–783.

[2] Sylvester, R. J. (1988), *A bayesian approach to the design of phase II clinical trials*. Biometrics **44** 823–836.

[3] Koubková, A., Barbosa, F.T. and Molenberghs, G. (2004) *A statistical proposal for sequential clinical trials in different cancer locations*, prepared for submission.

Acknowledgement: The work was supported by the Research plan MSM 113200008.

Address: A. Koubková, Department of Statistics, Charles University, Prague, Czech Republic
F.T. Barbosa, G. Molenberghs, Center for Statistics, Limburgs Universitair Centrum, Diepenbeek, Belgium

E-mail: koubkova@ksi.ms.mff.cuni.cz

PROJEVY GLOBÁLNÍCH ZMĚN V BIOSFÉRICKÉ REZERVACI TŘEBOŇSKO

Milena Kovářová

Klíčová slova: Mokré Louky, klimatologie, stanice meteorologická, ekosystém mokřadní, teplota vzduchu, vlhkost vzduchu, úhrny srážkové.

Abstrakt: Teplota vzduchu, především maximální teplota vzduchu, na Mokrých Loukách v období 1977-2003 stoupá. Relativní vlhkost vzduchu v tomto období klesá. Srážky, z hlediska dlouhodobých srážkových úhrnů značně stabilní, vykazují rostoucí tendenci v hodnotách velkých srážek a současně klesá srážková proměnlivost.

1 Úvod

V souvislosti s globální změnou klimatu je třeba se zabývat jejími projevy a jejím vlivem na krajinu a naopak. Termínem globální změna se označují procesy změn planetární geobiosféry způsobené nebo umocněné lidskou činností. Patří sem zejména globální změna klimatu a s ní spojená chemie atmosféry, změny hydrologického cyklu včetně jeho interakcí s půdou a krajinou a výskytů extrémních situací (sucho,povodně), biogeochemické cykly (hlavně cyklus uhlíku, dusíku) a jejich interakcí, geomorfologické procesy, změny využití území a zemského krytu a redukce biodiverzity. Pro předpovídání budoucího vývoje klimatu existuje celá řada modelů. Tyto modely, které je třeba neustále verifikovat pomocí reálných skutečně existujících dat, vycházejí ze znalostí fyzikálních zákonistostí o chování atmosféry a jejich prostřednictvím je odhadován budoucí vývoj klimatu. Jiným přístupem odhadu budoucího vývoje klimatu je odhad na základě skutečně naměřených dat. Existují rozsáhlé databáze klimatologických údajů, jejichž analýzou lze získat informace a znalosti o vzájemném působení jednotlivých klimatologických prvků v závislosti na změnách těchto charakteristik. V tomto příspěvku jsou hodnoceny denní srážkové úhrny, průměrná, maximální a minimální denní teplota vzduchu, průměrná a minimální denní relativní vlhkost naměřené v období 1977-2003 na jedné konkrétní stanici.

2 Meteorologická stanice Mokré Louky u Třeboně

V roce 1970 byl na konferenci UNESCO vyhlášen mezinárodní program ekologické spolupráce MaB Člověk a biosféra, který měl za cíl rozvinout v rámci přírodních a socioekonomických věd základnu pro racionální využívání přírodních zdrojů biosféry a předpovídat důsledky dnešních aktivit na budoucnost. V rámci tohoto projektu byla v roce 1977 zřízena Biosférická rezervace

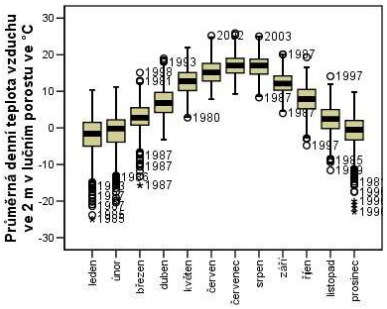
Třeboňsko, oblast velkých a rozmanitých mokřadů. V roce 1976 byla na levém břehu Prostřední stoky ve výtopě rybníka Rožmberk vybudována meteorologická stanice Mokré Louky, obrázek obr. 1. Od začátku roku 1977 jsou zde pravidelně denně měřeny základní meteorologické prvky v lučním porostu vysokých ostřic (např. *Carex gracilis, Carex vesicaria, Glyceria aquatica* a *Calamagrostis canescens*), od dubna 1978 do konce roku 1991 také v blízkém porostu vrby popelavé (*Salix cinerea*). Od roku 1983 jsou teploty a vlhkosti vzduchu měřeny v hodinových intervalech. Projekt sběru dat včetně meteorologické stanice založené Botanickým ústavem AV ČR přešel v roce 2003 pod Ústav ekologie krajiny AV ČR. Tato téměř třicet let dlouhá řada meteorologických pozorování je ojedinělá zejména tím, že stanice je umístěna přímo v přirozeném rozsáhlém mokřadním porostu. Naměřené údaje tedy představují dlouhodobé mikroklimatické podmínky v reálném mokřadním ekosystému.

Obrázek 1: Meteorologická stanice Mokré Louky.

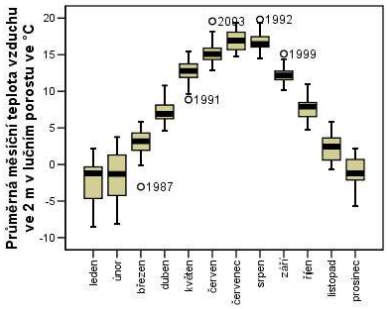
3 Teplota vzduchu

Průměrná roční teplota vzduchu na Mokrých Loukách ve 2m nad povrchem v období 1977-2003 byla 7.4°C. Nejvyšší nárůst průměrné teploty vzduchu byl pozorován v květnu o 3.8°C a v srpnu o 3.5°C. Tyto hodnoty, odhadnuté z okrajových období 1977-1980 a 2001-2003 jsou však nadhodnocené. Pro odhad růstu teploty je vhodnější porovnávat teploty v osmé a deváté dekádě, kde skutečný nárůst činí v červnu 1.2°C a v srpnu 1.3°C. Měsíční hodnoty průměrné teploty vzduchu počítané z denních a měsíčních dat jsou znázorněny na obrázku obr. 2. a obr. 3.

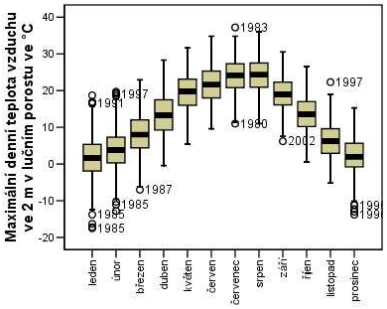
Průměrná roční maximální teplota vzduchu byla ve stejném období 13.1°C Měsíční hodnoty průměrné maximální teploty vzduchu počítané z denních a měsíčních dat jsou znázorněny na obrázku obr. 4. a obr. 5.



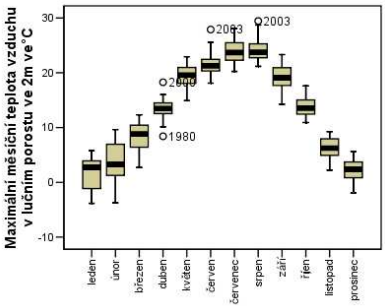
Obrázek 2: Průměrná denní teplota vzduchu ve 2m v lučním porostu ve °C.



Obrázek 3: Průměrná měsíční teplota vzduchu ve 2m v lučním porostu ve °C.



Obrázek 4: Průměrná denní maximální teplota vzduchu ve 2m v lučním porostu ve °C.

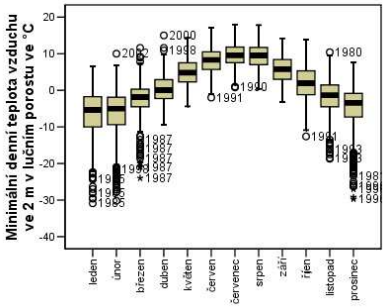


Obrázek 5: Průměrná měsíční maximální teplota vzduchu ve 2m v lučním porostu ve °C.

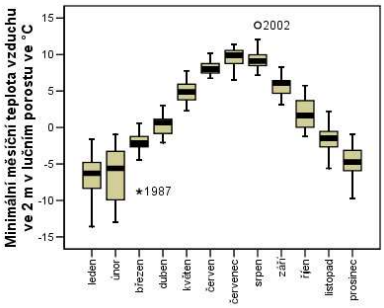
Absolutní teplotní maximum ve sledovaném období 1977-2003 zaznamenané na Mokrých Loukách 27.7.1983 dosáhlo hodnoty 37.2°C. Nejčastější výskyt teplot nad 30°C byl však pozorován v roce 2003, kdy teplota překročila 30°C 33krát. Celkový počet tropických dnů, to znamená dnů s maximálními teplotami nad 30°C včetně se pohybuje od 0 v letech 1977, 1978, 1979 a 1996 do třicetitří dnů v roce 2003. Maximální teploty mezi obdobími rostou pro období 1977-1980 a 2001-2003 v srpnu o 5.2°C a květnu o 4.7°C, za rok o 2°C.

Průměrná roční minimální teplota vzduchu byla ve stejném období 1.5°C. Měsíční hodnoty průměrné minimální teploty vzduchu počítané z denních a měsíčních dat jsou znázorněny na obrázku obr. 6. a obr. 7.

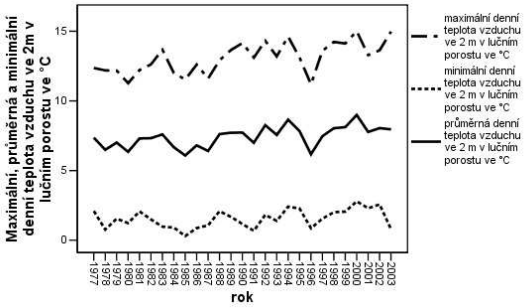
Absolutní teplotní minimum zaznamenané na Mokrých Loukách 7.1.1985 dosáhlo hodnoty −30.9°C. Arktické dny s maximální teplotou pod −10°C



Obrázek 6: Průměrná denní minimální teplota vzduchu ve 2m v lučním porostu ve °C.



Obrázek 7: Průměrná měsíční minimální teplota vzduchu ve 2m v lučním porostu ve °C.



Obrázek 8: Průměrná roční maximální, průměrná a minimální teplota vzduchu ve 2m v lučním porostu ve °C.

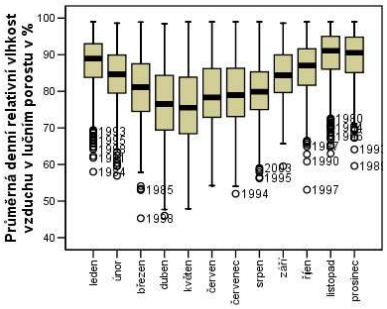
včetně se vyskytly od prosince do února roce 1985 - 8krát, 1996 - 6krát, 1987 - 3krát a jednou v letech 1979, 1980, 1982, 1986 a 1993. Minimální teploty mezi obdobími stoupají méně než průměrné a maximální, v některých měsících klesají. Celkový vzestup minimálních teplot mezi obdobími 1977-1980 a 2001-2003 i mezi osmou a devátou dekádou činí zhruba 0.5°C.

Hodnoty průměrné roční maximální, průměrné a minimální teploty vzduchu v jednotlivých letech jsou znázorněny na obrázku obr. 8.

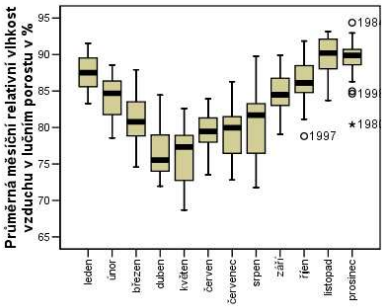
4 Relativní vlhkost vzduchu

Mokré Louky se vyznačují poměrně vysokými hodnotami relativní vzdušné vlhkosti, současně však lze pozorovat pokles vlhkosti v jarních a letních měsících přes dekády. Měsíční průměrná relativní vlhkost vzduchu poklesla mezi osmou a devátou dekádou v srpnu z 82.2% na 78.1%, významnost 0.01,

v červnu z 80.3% na 78.1%, významnost 0.1. Z dlouhodobého hlediska lze nejnižší průměrnou relativní vlhkost vzduchu na Mokrých Loukách pozorovat v květnu, případně v dubnu, nejvyšší v prosinci, případně v lednu či v listopadu. Mezi nejnižší hodnotou pro květen a nejvyšší pro prosinec je rozdíl 13.2%. Měsícem s nejvyšší proměnlivostí vlhkosti je duben, nejnižší proměnlivost vlhkosti bývá v prosinci. Z charakteru relativní vzdušné vlhkosti je zřejmé, že maximální průměrná denní relativní vlhkost 99% se může vyskytnout kdykoliv během roku při déletrvajících srážkách nebo mlze. Nízké hodnoty průměrné denní vlhkosti se mohou vyskytnout téměř v kterémkoliv měsíci, v zimních měsících jsou tyto výskyty spíše ojedinělé. Nejnižší průměrná denní vlhkost se na Mokrých Loukách pohybuje nad 45%. Měsíční hodnoty průměrné relativní vlhkosti vzduchu počítané z denních a měsíčních dat jsou znázorněny na obrázku obr. 9. a obr. 10.



Obrázek 9: Průměrná denní relativmí vlhkost vzduchu v lučním porostu v procentech.

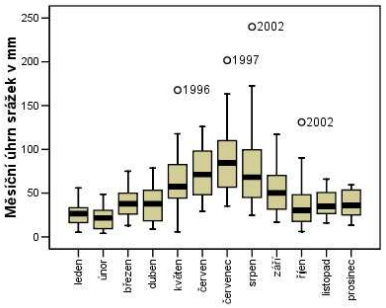


Obrázek 10: Průměrná měsíční relativmí vlhkost vzduchu v lučním porostu v procentech.

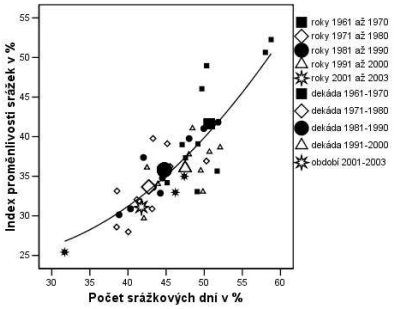
5 Srážky

Hodnocení srážkových poměrů Mokrých Luk se vztahuje k období 1961-2003. Na základě porovnání srážkových úhrnů meteorologické stanice Českého hydrometeorologického ústavu Třeboň a stanice Mokré Louky v období 1977-1986, které vykazují jen nepatrné statisticky nevýznamné odchylky v hodnotách denních srážkových úhrnů, byla pro období 1961-1976 použita data z nedaleké stanice Třeboň.

Průměrný roční srážkový úhrn za celé období 1961-2003 činí 605,3 mm vodního sloupce. Pro oblast Třeboně jsou typické vyšší srážkové úhrny v letních měsících a poměrně nízké úhrny v zimním období. Měsíční hodnoty srážkových úhrnů jsou znázorněny na obrázku obr. 11. Měsíční srážkové úhrny se v období 1961-2003 pohybují od 4 mm v únoru 1982 do 240 mm v srpnu 2002. Maximální srážkové úhrny se nejčastěji vyskytnou od května do srpna, minimální v lednu, únoru a říjnu. Nejnižší průměrné denní srážkové úhrny byly



Obrázek 11: Průměrné měsíční srážkové úhrny v mm.



Obrázek 12: Závislost indexu srážkové proměnlivost na počtu srážkových dní.

zaznamenány v lednu a činí 0.84 mm srážek na den, nejvyšší v červnu 2.85 mm na den. Hodnota nejvyššího pozorovaného denního úhrnu ze dne 7.8.2000 dosahuje 129 mm vodního sloupce.

Vzhledem k velkému meziročnímu rozptylu nelze prokázat změny v hodnotách dlouhodobých měsíčních srážkových úhrnů v průběhu let [4]. Dochází však ke změnám v proměnlivosti srážek, které se projevují poklesem četnosti srážkových dní, snižováním četnosti střídání srážkových a bezesrážkových období a růstem hodnot velkých srážek. Označme n počet dní v měsíci, s počet srážkových dní v měsíci a b počet bezesrážkových dní v měsíci. Pak $ps = 100 * s/n$ udává procento srážkových dní v měsíci. Označme $b1$ počet bezesrážkových dní v měsíci bezprostředně následujících po srážkovém dni. Definujme index proměnlivosti srážek v měsíci: $indexdiv = 100 * b1/b$.

Na obrázku obr. 12. je znázorněna závislost indexu proměnlivosti srážek na procentu srážkových dní. Body ležící nad křivkou označují roky srážkově proměnlivější a body ležící pod křivkou označují roky srážkově méně proměnlivé. Z grafu je patrný velký pokles počtu srážkových dní i indexu proměnlivosti srážek mezi šestou dekádou a následujícími dekádami. V deváté dekádě, i přes mírný nárůst počtu srážkových dní, lze pozorovat pokles proměnlivosti srážek, který dále pokračuje, včetně poklesu počtu srážkových dní i v počínající první dekádě třetího tisíciletí. Pokles procenta srážkových dní mezi šestou a devátou dekádou je významný na hladině 0.05. Pokles indexu proměnlivosti mezi šestou a devátou dekádou je významný na hladině 0.005.

Průměrná četnost dnů se srážkami se dlouhodobě pohybuje kolem 50%, nejnižší je v říjnu s průměrem 13.1 srážkových dnů za měsíc. Nejčastěji pršelo v roce 1966 s 215 srážkovými dny. Nejnižší počet 116 srážkových dní byl pozorován v roce 2003. Srážky nad 1mm jsou pravděpodobnější v letních měsících, srážky nad 10mm se mohou vyskytnout v kterémkoliv měsíci, častější jsou v létě. Srážky 20mm a více se vyskytly nejméně jedenkrát v letech 1969, 1983, 1989 a 1992 a nejvíce desetkrát v roce 2002. Za celé období 1961-2003

se srážky větší nebo rovné 20mm vyskytly 165krát, což představuje zhruba 2% z celkového počtu srážkových dní v uvedeném období. Podíl těchto srážek na celkovém srážkovém úhrnu však dosahuje 20%. Kruskal-Wallisovým pořadovým testem lze ukázat, že mezi dekádami dochází k růstu hodnot velkých srážek. V šesté dekádě se vyskytlo 41 srážek nad 20mm včetně, hodnota pořadového testu je 70.3, v sedmé dekádě se 40 výskyty je hodnota pořadového testu 85.7, v osmé dekádě s 30 výskyty nabývá pořadový test hodnoty 70.2, v deváté dekádě s 37 výskyty 97.2 a v období 2001-2003 se 17 výskyty hodnoty 99, významnost 0.035.

Z provedených analýz se zdá být pravděpodobné, že v posledních letech dochází k narušení malého vodního cyklu [9], a že srážkové úhrny, ačkoliv dlouhodobě velice stabilní, jsou dorovnávány na úroveň obvyklých normálů prostřednictvím zvyšujících se, jednorázových srážek. Pokles proměnlivosti srážek se na srážkových úhrnech ani na počtu srážkových dní neprojeví, může však vést k vysychání krajiny [8] s následným růstem především maximálních teplot, případně ke vzniku povodňových situací. Je zřejmé, že uvedená tvrzení by bylo třeba ověřit na větším počtu co nejdelších srážkových řad. Srážkové řady před rokem 1960 jsou však poměrně často zatíženy chybami, protože spíš než na přesné zaznamenání denních výskytů srážek, byl kladen důraz na zaznamenání celkových srážkových úhrnů, někdy i vícedenních.

6 Závěr

Z hodnot sledovaných charakteristik v poslední dekádě minulého století a v počínajících letech první dekády třetího tisíciletí plynou možné důsledky změny klimatu, jež mohou vyústit v degradaci mokřadů a v omezení jejich regulačního a stabilizačního vlivu na krajinu a ekosystém.

Reference

[1] Houghton J. (1998). *Globální oteplování.* Academia, Praha.

[2] Klimadata Bot. Inst. Ac. Sci. (2003). *Klimatologická data Mokré Louky u Třeboně.* Botanický ústav AVČR, hydrobotanické oddělení, 1977 – 2002.

[3] Kovářová M. (2004). *Databáze klimatologických údajů - Mokré Louky.* III. seminář z ekologie mokřadů a hydrobotaniky, pořádaný na paměť Slavomila Hejného, Třeboň.

[4] (1969) *Podnebí Československé socialistické republiky. Souborná studie*, Hydrometeorologický ústav, Praha.

[5] Pokorný J., Kučerová A. (2000), *Monitoring klimatu a atmosférických depozic v CHKO Třeboňsko*, -In Pokorný J., Šulcová J., Hátle M., Hlásek J. (eds.) Třeboňsko 2000. Ekologie a ekonomika Třeboňska po dvaceti letech, : UNESCO/MaB, ENKI o.p.s., Třeboň, 87- 99.

[6] Přibáň K. a kol. (1992) *Analysis and modeling of wetland microclimate.* The case study Třeboň Biosphere Reserve. - Stud. ČSAV 1992/2, ed. Academia, Praha, 1 – 168.

[7] Přibáň K. (1978). *Ekologické aspekty třeboňského klimatu*. In: Jeník J. et Květ J.(eds.), Ekologie a ekonomika Třeboňska.

[8] Ripl W., Pokorný J., Eiseltová M., Ridgill S. (1994). *A holistic approach to the structure and function of wetlands, and their degradation*. In Eiseltová M. (ed), Restoration of lake ecosystems - a holistic approach. IWRB Publ. **32**, 182 pp.

[9] Ripl W.: *Management of water cycle and energy flow for ecosystem control - the energy transport reaction (ETR) model*. Ecological Modelling **78**, 61 – 76.

Adresa: Milena Kovářová, Botanický ústav AV ČR

TESTING GOODNESS OF FIT IN THE COX–AALEN MODEL

David Kraus

Keywords: Counting process, Cox–Aalen model, goodness-of-fit, martingale, residual, survival analysis.

Abstract: The Cox–Aalen regression model for the intensity of counting processes suggested by Scheike and Zhang [13] extends the Cox proportional hazards model as well as the Aalen additive model. We study goodness-of-fit tests based on the stratified martingale residual process. Asymptotic distribution of this process is complicated (a Gaussian process with a complex covariance structure). Therefore, a direct use of, e.g., the Kolmogorov–Smirnov type test is impossible. We show two ways out of this problem. One possibility is to simulate realisations from the limiting distribution of the residual process, as suggested by Lin, Wei and Ying [7] for the Cox model. Another approach consists of tranforming (compensating) the limiting process to a martingale (following ideas of Khmaladze [4]). Both methods are compared in a simulation study.

1 Introduction

In survival analysis, regression models are used to explain occurence of events (failures) in time by the influence of explanatory variables (covariates). Let $Z_i = \{(Z_{i1}(t), \dots, Z_{ip}(t))^\mathsf{T}, t \in [0, \tau]\}$ be a vector of covariates (possibly time-dependent, i.e. predictable stochastic processes) for the i-th observed individual, Y_i be an indicator process (indicating by its value at time t whether the i-th individual is at risk of the event) and λ_i be the intensity process of the corresponding counting process N_i. Then the most popular model for the intensity process is the Cox proportional hazards model of the form

$$\lambda_i(t) = Y_i(t) \exp\{\beta_0^\mathsf{T} Z_i(t)\} \lambda_0(t),$$

where λ_0 is an unknown baseline hazard function and β_0 is a p-vector of unknown regression parameters. Another frequently used model is the Aalen additive model

$$\lambda_i(t) = Y_i(t) X_i(t)^\mathsf{T} \alpha(t)$$

with a vector of unknown functions $\alpha = \{(\alpha_1(t), \dots, \alpha_q(t))^\mathsf{T}, t \in [0, \tau]\}$ and a vector of covariates X_i.

There are several ways of combining these two models (e.g. [8], [12]). Here we study the Cox–Aalen model, which is due to Scheike and Zhang [13]. Their model follows the form

$$\lambda_i(t) = Y_i(t) \exp\{\beta_0^\mathsf{T} Z_i(t)\} X_i(t)^\mathsf{T} \alpha(t), \qquad 0 \le t \le \tau, \tag{1}$$

where $(X_i^\mathsf{T}, Z_i^\mathsf{T})^\mathsf{T}$ is a $(q+p)$-vector of predictable covariates (usually $X_{i1} \equiv 1$). Some components of $\alpha(t)$ or $X_i(t)$ can even be negative, provided the whole term $X_i(t)^\mathsf{T}\alpha(t)$ is nonnegative (an intensity always has to be so).

In [13] an estimation procedure was suggested and asymptotic properties of the estimators were derived. Here we propose a test of goodness of fit. For the Cox model many goodness-of-fit tests were developed. One approach, which is here adapted for the situation of the Cox–Aalen model, is based on the stratified martingale residual process.

The idea was originated by Arjas [3] who suggested a graphical procedure for assessing goodness of fit of the Cox model. The method is based on comparison of observed and expected number of failures within a given stratum. For each observed individual $i \in \{1, \dots, n\}$ this difference is expressed by the process $\hat{M}_i = N_i - \hat{\Lambda}_i$. Hence, if a stratum $I \subset \{1, \dots, n\}$ is chosen, the process $\Xi_I = \sum_{i \in I} \hat{M}_i$ should fluctuate around zero. Let $0 \le T_{(I,1)} \le T_{(I,2)} \le \cdots \le T_{(I,K)} \le \tau$ be ordered times of the actual failures in I. Then Arjas's plots are plots of the values $\sum_{i \in I} \hat{\Lambda}_i(T_{(I,k)})$ against $k = \sum_{i \in I} N_i(T_{(I,k)})$. The graph should be close to the line with slope 1 when the fit is good, and should differ otherwise.

Asymptotic behaviour of the residual process for the Cox model was studied by Marzec and Marzec [9]. Under certain conditions they showed that the limiting distribution of $n^{-1/2}\Xi_I$ is that of a continuous zero-mean Gaussian process. Later, in [10] they presented some generalisations and used a transformation of the limiting process to a martingale which enables construction of the Kolmogorov–Smirnov type test. The idea of Arjas's plots and stratified residual processes was successfully used by Volf [14] also for the Aalen additive model.

In the situation of the Cox–Aalen model of (1), the stratified martingale residual process has the form

$$\begin{aligned}\Xi_I(t) &= \sum_{i \in I} (N_i(t) - \hat{\Lambda}_i(t)) \\ &= \sum_{i \in I} \int_0^t \left[dN_i(s) - Y_i(s) \exp\{\hat{\beta}^\mathsf{T} Z_i(s)\} X_i(s)^\mathsf{T} d\hat{A}(s) \right] \quad (2)\end{aligned}$$

(where $\hat{\beta}$ and $\hat{A}(t)$ are estimates of β_0 and $A(t) = \int_0^t \alpha(s)ds$, respectively).

The paper is organised as follows. Section 2 establishes asymptotic properties of the residual process. Section 3 is devoted to the study of several testing procedures. In Section 4 a simulation study is presented and in Section 5 we mention some generalisations. Due to the lack of space, some results are only sketched in a symbolic way and many details are omitted; the author refers an interested reader to his paper [5].

2 Asymptotic distribution of the residual process

The estimation procedure for the parameters β_0 and $A(t) = \int_0^t \alpha(s)ds$ is described in detail in [13]. Here we only mention that it consists of two steps: first the parameter β_0 is estimated by solving a score equation and then the increments of $A(t)$ are estimated by a weighted least squares principle. Denote the estimates by $\hat{\beta}$ and $\hat{A}$. These two steps can be repeated to obtain better estimates.

By Taylor's expansion around β_0 and after some rearrangement, Ξ_I can be expressed as

$$\Xi_I(t) = \Gamma_1(t) - Q(\tilde{\beta}, \tilde{\tilde{\beta}}, t)^{\mathsf{T}}\Gamma_2(\tau), \tag{3}$$

where

$$\Gamma_1(t) = \int_0^t K_1(\beta_0, s)^{\mathsf{T}} dM(s) \qquad \text{and} \qquad \Gamma_2(t) = \int_0^t K_2(\beta_0, s)^{\mathsf{T}} dM(s)$$

are martingales. Here $M(t) = (M_1(t), \ldots, M_n(t))^{\mathsf{T}}$, $\tilde{\beta}$ and $\tilde{\tilde{\beta}}$ lie on the line segment between β_0 and $\hat{\beta}$, and K_1, K_2 and Q are some processes of dimensions $n \times 1$, $n \times p$ and $p \times 1$, respectively. The process $Q(\tilde{\beta}, \tilde{\tilde{\beta}}, \cdot)$ converges uniformly in probability to a function, say $q(\beta_0, \cdot)$. The weak convergence in $(D[0,\tau])^{1+p}$ of the martingale $n^{-1/2}(\Gamma_1, \Gamma_2^{\mathsf{T}})^{\mathsf{T}}$ to a zero-mean continuous Gaussian martingale $\gamma = (\gamma_1, \gamma_2^{\mathsf{T}})^{\mathsf{T}}$ follows by Rebolledo's martingale central limit theorem [2, Theorem I.2]. Therefore we have the weak convergence

$$n^{-1/2}\Xi_I(\cdot) \xrightarrow[n\to\infty]{\mathcal{D}} \xi_I(\cdot) = \gamma_1(\cdot) - q(\beta_0, \cdot)^{\mathsf{T}}\gamma_2(\tau) \qquad \text{in } (D[0,\tau]). \tag{4}$$

The limiting process ξ_I is a Gaussian process, which, however, is neither a martingale nor a process with a well-known distribution.

The (co)variance function of γ is the limit of the predictable (co)variation of $n^{-1/2}\Gamma$ and can be estimated uniformly consistently by the quadratic (optional) (co)variation of $n^{-1/2}\Gamma$. Details concerning the covariance structure as well as explicit forms of all of the processes, a further notation, assumptions, exact formulation of the results and proofs can be found in [5].

3 Testing procedures

A graphical technique, an analogue of the residual plots of Arjas [3], has already been explained in Section 1. In this section we describe visual and mainly numerical methods of investigation of the martingale residual process.

Except for a special case of the model with one dichotomous covariate in the Cox part [5, Section 4.1], a direct use of the asymptotic distribution is not possible because of its complexity. For instance, the limiting distribution of the Kolmogorov–Smirnov type statistic is untractable. We describe how some approximations and transformations enables us to cope with this difficulcy.

3.1 The simulation approximation

We shall describe how to approximate the asymptotic distribution of the martingale residual process through simulations. The simulations can be performed to obtain a sample from the limiting distribution, and hence to assess both graphically and numerically how unusual the observed residual process is. Our approach is based on the idea of [7].

In (3) we have found the martingale representation of the residual process of the form

$$\Xi_I(t) = \Gamma_1(t) = \int_0^t K_1(\beta_0, s)^{\mathsf{T}} dM(s) - Q(\tilde{\beta}, \tilde{\tilde{\beta}}, t)^{\mathsf{T}} \int_0^t K_2(s)^{\mathsf{T}} dM(s),$$

The limiting distribution can be approximated by plugging in the consistent estimate $\hat{\beta}$ in place of $\beta_0, \tilde{\beta}, \tilde{\tilde{\beta}}$, and by replacing the martingale increments $dM_i(t)$ by their simulated values. For the martingales M_i, $i = 1, \ldots, n$ it holds that $\mathsf{E}\, M_i(t) = 0$ and $\operatorname{var} M_i(t) = \mathsf{E}\,[M_i(t)^2] = \mathsf{E}\,\Lambda_i(t) = \mathsf{E}\, N_i(t)$. Therefore Lin, Wei and Ying [7] suggested to approximate M_i by $\tilde{M}_i = G_i N_i$ (i.e. $\tilde{M} = \operatorname{diag}[G]N$), where $G = (G_1, \ldots, G_n)$ is a random sample of standard normal variables independent of the data. We obtain the approximation

$$\tilde{\Xi}_I(t) = \int_0^t K_1(\hat{\beta}, s)^{\mathsf{T}} \operatorname{diag}[G] dN(s) \\ - Q(\hat{\beta}, \hat{\beta}, t)^{\mathsf{T}} \int_0^\tau K_2(s)^{\mathsf{T}} \operatorname{diag}[G] dN(s).$$

It can be shown (by means similar to that of [7]) that the asymptotic distribution of $n^{-1/2}\tilde{\Xi}_I$ is equal to that of $n^{-1/2}\Xi_I$, i.e. the distribution of ξ_I. Thus, generating repeatedly a standard normal sample G and computing $\tilde{\Xi}_I$, we obtain a sample from the desired limiting distribution. We can visually assess goodness of fit by plotting Ξ_I together with an appropriate number of simulated $\tilde{\Xi}_I$. To test the hypothesis numerically we generate an adequately large number of realisations of $\tilde{\Xi}_I$ and estimate critical values or the p-value.

Note that the simulation of the asymptotic distribution is conditional on the data and therefore we do not obtain universal critical values of this distribution. In other words, the test is not distribution-free and we have to carry out the simulations for each particular data set separately. This can be computationally demanding.

3.2 The transformation method

Another way of overcoming the problem with complexity of the asymptotic distribution of the residual process is based on the idea of Khmaladze [4]. In the framework of testing whether the distribution of a random variable follows a parametric form, he suggested a transformation of empirical processes with plugged-in estimated parameters in order to obtain a well-known asymptotic distribution which does not depend on the distribution of the data. The idea

was then used in [1, VI.3.3.4] for testing goodness of fit of parametric models for intensities and in [10] for assessment of the Cox model.

Here we will find the compensator of ξ_I, say $\bar{\xi}_I$, and use the empirical counterpart Ψ_I of the Gaussian martingale $\psi_I = \xi_I - \bar{\xi}_I$ as a basis for testing. The process ψ_I should be a martingale with respect to the filtration generated by ξ_I, i.e. with respect to $\mathcal{G}_t = \sigma\{\gamma(s), s \le t; \gamma_2(\tau)\}$, $t \in [0, \tau]$. As the term $q(\beta_0, \cdot)^{\mathsf{T}}\gamma_2(\tau)$ in (4) is measurable w.r.t. $\mathcal{G}_0$, we only need to compensate γ_1. The compensator of γ_1, say $\bar{\gamma}_1$, can be derived rather heuristically as follows (cf. [1, VI.3.3.4, pp. 464–466]). Since the process γ is a Gaussian martingale, it has independent increments and $(d\gamma_1(t), \gamma_2(\tau) - \gamma_2(t))^{\mathsf{T}}$ is jointly normally distributed. Therefore

$$\begin{aligned}\mathsf{E}\,[d\gamma_1(t)|\gamma(s), s \le t; \gamma_2(\tau)] &= \mathsf{E}\,[d\gamma_1(t)|\gamma_2(\tau) - \gamma_2(t)] \\ &= \operatorname{cov}\{d\gamma_1(t), \gamma_2(\tau) - \gamma_2(t)\}[\operatorname{var}\{\gamma_2(\tau) - \gamma_2(t)\}]^{-1}[\gamma_2(\tau) - \gamma_2(t)] \\ &= c(\beta_0, dt)^{\mathsf{T}}\theta(\beta_0, t)^{-1}[\gamma_2(\tau) - \gamma_2(t)],\end{aligned}$$

where the definition of c and θ is obvious. Hence a natural candidate for the compensator of γ_1 is

$$\begin{aligned}\bar{\gamma}_1(t) &= \int_0^t \mathsf{E}\,[d\gamma_1(s)|\gamma(u), u \le s; \gamma_2(\tau)] \\ &= \int_0^t [\gamma_2(\tau) - \gamma_2(s)]^{\mathsf{T}}\theta(\beta_0, s)^{-1}c(\beta_0, ds), \quad t \in [0, \tau].\end{aligned}$$

The compensated residual process is then

$$\psi_I(t) = \gamma_1(t) - \int_0^t [\gamma_2(\tau) - \gamma_2(s)]^{\mathsf{T}}\theta(\beta_0, s)^{-1}c(\beta_0, ds).$$

Formal verification of the fact that ψ_I is actually a Gaussian martingale (with the same variance function as γ_1) is analogous to the proof of Lemma 3.2 in [10]. Finally, the empirical counterpart of ψ_I can be

$$\Psi_I(t) = \Xi_I(t) - \int_0^t \left[\int_s^\tau K_2(s)^{\mathsf{T}} d\hat{M}(s)\right]^{\mathsf{T}} \Theta(\hat{\beta}, s)^{-1} C(\hat{\beta}, ds) \tag{5}$$

with C and Θ being some estimates of c and θ. The issue whether $n^{-1/2}\Psi_I$ truly converges to ψ_I is discussed in [5]. Our computational experience shows that the convergence can be considered valid generally.

Hence the Kolmogorov–Smirnov type statistic $\sup|\Psi_I(t)|/\{\widehat{\operatorname{var}}\,\gamma_1(\tau)\}^{1/2}$ is asymptotically distributed as the variable $\sup|W(t)|$ (where W denotes the Brownian motion).

4 Simulations

We performed a small simulation study in order to investigate performance of the tests. We generated survival data following various models with various censoring patterns and estimated the Cox–Aalen model of the form

$$\lambda_i(t) = \{\alpha_1(t) + \alpha_2(t)X_i\} \exp\{\beta_1 Z_i\}.$$

Under the null hypothesis H_0 the samples came from the distribution with the intensity

$$\lambda_i(t) = \{0.5 + 0.2tX_i\} \exp\{0.5Z_i\},$$

where X_i was uniformly distributed on $[0, 1]$ and Z_i had the standard normal distribution. Then two alternatives with additional covariates were considered: H_1 with

$$\lambda_i(t) = \{0.5 + 0.2tX_i + 0.7tX_i^*\} \exp\{0.5Z_i\},$$

where X_i^* had the alternative distribution on $\{0, 1\}$ with probability 0.5, and H_2 having

$$\lambda_i(t) = \{0.5 + 0.2tX_i\} \exp\{0.5Z_i + 0.7Z_i^*\}$$

with Z_i^* having the same distribution as X_i^* in the previous situation. The covariates were generated independently. Three censoring schemes were considered: no censoring, moderate censoring (with censoring times having the uniform distribution on $[0, 5]$) and heavy censoring (with uniform censoring times on $[0, 2.5]$). The censoring times were mutually independent and independent of the survival times and covariates. The corresponding censoring rates are indicated in the tables. The sample sizes were $n = 100$ and 200. For the null hypothesis, stratification with respect to both covariates was studied, i.e. the stratum was first $I = \{i : X_i > 0.5\}$ and then $I = \{i : Z_i > 0\}$. Under the alternatives, the data were stratified with respect to the missing covariate: $I = \{i : X_i^* = 1\}$ under H_1, and $I = \{i : Z_i^* = 1\}$ under H_2.

Under the null hypothesis as well as under the alternatives, we generated the sample, estimated the model and tested goodness of fit. Two tests were performed: the test of Subsection 3.1 based on the simulation approximation (with 2000 realisations of the residual process) and the test of Subsection 3.2 based on the transformation. This was repeated 1000 times and empirical levels and powers of the tests on the nominal level of 0.05 were computed. Since the estimation of the model is highly time-consuming, we were able to carry out only 1000 repetitions in each situation, hence our results give only a broad image of the behaviour of the tests.

Table 1 reports the sizes of the tests in the above mentioned situations under H_0. It is seen that the tests maintain approximately their nominal significance levels which is not seriously affected by censoring. Table 2 confirms that the tests have good power against the alternatives H_1 and H_2 of missing

covariates. When censoring is present, the power decreases. There is no important difference between the two versions (simulation and transformation) of the test.

		$I = \{i : X_i > 0.5\}$		$I = \{i : Z_i > 0\}$	
		$n = 100$	$n = 200$	$n = 100$	$n = 200$
Without	Simulation	0.043	0.043	0.051	0.060
censoring	Transform.	0.067	0.054	0.073	0.047
Censoring	Simulation	0.044	0.049	0.052	0.047
U[0, 5] (31 %)	Transform.	0.054	0.064	0.053	0.058
Censoring	Simulation	0.056	0.071	0.045	0.044
U[0, 2.5] (51 %)	Transform.	0.041	0.056	0.013	0.009

Table 1: Empirical sizes of the two tests on the nominal level of 0.05.

		H_1		H_2	
		$n = 100$	$n = 200$	$n = 100$	$n = 200$
Without	Simul.	0.726	0.979	0.889	0.995
censoring	Transf.	0.739	0.972	0.887	0.994
Censoring U[0, 5]	Simul.	0.553	0.879	0.794	0.980
(H_1 25 %, H_2 24 %)	Transf.	0.553	0.871	0.777	0.971
Censoring U[0, 2.5]	Simul.	0.313	0.633	0.664	0.955
(H_1 45 %, H_2 42 %)	Transf.	0.304	0.621	0.661	0.945

Table 2: Empirical powers of the two tests on the nominal level of 0.05.

5 Generalisations

In this paper we used the idea of stratification. The tests based on the martingale residual processes $\hat{M}_i$ can be generalised in the following direction. Instead of the sum of the martingale residuals $\hat{M}_i(t) = N_i(t) - \hat{\Lambda}_i(t)$ over a stratum we can use sums of their transforms of the form

$$\Xi(t) = \sum_{i=1}^{n} \int_0^t H_i(s) d\hat{M}_i(s),$$

where H_i are vectors of some predictable processes. This type of tests was studied in [6]. A particularly important choice is $H_i(t) = Z_i(t)$ leading to the score process which is used for detection of departures from the proportional hazards assumption. The choice $H_i = 1_{\{i \in I\}}$ corresponds to the stratified martingale residual process presented in this paper.

References

[1] Andersen P.K., Borgan Ø., Gill R.D., Keiding N. (1993). *Statistical models based on counting processes.* Springer, New York.

[2] Andersen P.K., Gill R.D. (1982). *Cox's regression model for counting processes: a large sample study.* Ann. Statist. **10**, 1100 – 1120.

[3] Arjas E. (1988). *A graphical method for assessing goodness of fit in Cox's proportional hazards model.* J. Amer. Statist. Assoc. **83**, 204 – 212.

[4] Khmaladze E.V. (1981). *Martingale approach in the theory of goodness-of-fit tests.* Teor. Veroyatnost. i Primenen. **26**, 246 – 265. In Russian. English translation in Theory Probab. Appl. **26**, 240 – 257.

[5] Kraus D. (2004). *Goodness-of-fit inference for the Cox–Aalen additive-multiplicative regression model.* Statist. Probab. Lett. To appear.

[6] Kraus D. (2004). *Testing goodness of fit of hazard regression models.* In WDS 2004 Proceedings of Contributed Papers, Part I: Mathematics and Computer Sciences, Matfyzpress, Prague, 6 – 12.

[7] Lin D.Y., Wei L.J., Ying Z. (1993). *Checking the Cox model with cumulative sums of martingale-based residuals.* Biometrika **80**, 557 – 572.

[8] Lin D.Y., Ying Z. (1995). *Semiparametric analysis of general additive-multiplicative hazard models for counting processes.* Ann. Statist. **23**, 1712 – 1734.

[9] Marzec L., Marzec P. (1993). *Goodness of fit inference based on stratification in Cox's regression model.* Scand. J. Statist. **20**, 227 – 238.

[10] Marzec L., Marzec P. (1997). *Generalized martingale-residual processes for goodness-of-fit inference in Cox's type regression models.* Ann. Statist. **25**, 683 – 714.

[11] McKeague I.W., Utikal K.J. (1991). *Goodness-of-fit tests for additive hazards and proportional hazards models.* Scand. J. Statist. **18**, 177 – 195.

[12] Scheike T.H., Martinussen T. (2002). *A flexible additive multiplicative hazard model.* Biometrika **89**, 283 – 298.

[13] Scheike T.H., Zhang M.-J. (2002). *An additive-multiplicative Cox–Aalen regression model.* Scand. J. Statist. **29**, 75 – 88.

[14] Volf P. (1996). *Analysis of generalized residuals in hazard regression models.* Kybernetika **32**, 501 – 510.

Acknowledgement: The author acknowledges that this paper is an abbreviated version of his paper [5] published elsewhere. The research was partly supported by the GAČR Grants 201/02/0049 and 402/04/1294.

Address: D. Kraus, Department of Statistics, Charles University, Sokolovská 83, CZ-186 75 Prague, Czech Republic

E-mail: david.kraus@karlin.mff.cuni.cz

LINEÁRNÍ REGRESNÍ MODELY S RUŠIVÝMI PARAMETRY

Pavla Kunderová

Klíčová slova: Lineární regresní model, užitečné a rušivé parametry, BLUE, eliminační transformace.

Abstrakt: Cílem článku je uvést přehled nejdůležitějších výsledků, které byly získány při studiu lineárních regresních modelů, ve kterých je vektor parametrů prvního řádu rozdělen na podvektor užitečných parametrů a na podvektor rušivých parametrů.

1 Označení, úvod

Budeme užívat následující symboly: nechť $\mathcal{M}(\mathrm{A})$ označuje prostor vytvořený sloupci matice A. Je-li $\mathcal{M}(\mathrm{A}) \subset \mathcal{M}(\mathrm{B})$, V p.s.d., potom symbol P_A^V označuje projekční matici projektující na podprostor $\mathcal{M}(\mathrm{A})$ vzhledem k V- seminormě dané maticí V, $||\mathrm{x}||_= \sqrt{\mathrm{x}'\mathrm{V}\mathrm{x}}; \mathrm{M}_A^V = \mathrm{Y} - \mathrm{P}_A^V$. Obecně platí $\mathrm{P}_A^V = \mathrm{A}(\mathrm{A}'\mathrm{V}\mathrm{A})^-\mathrm{A}'\mathrm{V} + \mathrm{F}(\mathrm{Y} - \mathrm{V}^-\mathrm{V})$, kde F je libovolná matice vhodného typu, (viz [10], (2.14)).

Nechť $\mathrm{N}_{n,n}$ je p.d. (p.s.d.) matice a $\mathrm{A}_{m,n}$ libovolná matice, potom symbol $\mathrm{A}^-_{m(N)}$ označuje matici splňující vztah $\mathrm{A}\mathrm{A}^-_{m(N)}\mathrm{A} = \mathrm{A}$ a $\mathrm{N}\mathrm{A}^-_{m(N)}\mathrm{A} = [\mathrm{N}\mathrm{A}^-_{m(N)}\mathrm{A}]'$. Nechť

$$\mathrm{Y} = (\mathrm{X}, \mathrm{S}) \begin{pmatrix} \beta \\ \kappa \end{pmatrix} + \varepsilon, \ \beta \in R^{k_1}, \ \kappa \in R^{k_2}, \tag{1}$$

kde Y je n-rozměrný observační vektor, β je k_1-rozměrný vektor užitečných parametrů, κ je k_2-rozměrný vektor rušivých parametrů, X je daná $n \times k_1$ matice plánu příslušná užitečným parametrům, S je daná $n \times k_2$ matice plánu příslušná rušivým parametrům, ε je náhodný vektor chyb.
Předpokládejme, že
$E(\mathrm{Y}) = \mathrm{X}\beta + \mathrm{S}\kappa, \ \forall \beta \in R^{k_1}, \ \forall \kappa \in R^{k_2}$; $var(\mathrm{Y}) = \Sigma_\vartheta = \sum_{i=1}^p \vartheta_i V_i, \ \forall \vartheta = (\vartheta_1, \ldots, \vartheta_p)' \in \underline{\vartheta} \subset R^p$; $V_1, \ldots, V_p$ dané symetrické matice; $\underline{\vartheta} \subset R^p$ obsahuje otevřenou kouli v R^p; je-li $\vartheta \in \underline{\vartheta}$, je matice Σ_ϑ pozitivně semidefinitní; matice Σ_ϑ není funkcí vektoru $(\beta', \kappa')'$.

Je-li matice Σ_ϑ pozitivně definitní pro každé $\vartheta \in \underline{\vartheta}$ a $r(\mathrm{X}, \mathrm{S}) = k_1 + k_2 < n$, nazveme model (1) *regulární*, (viz [2], str.13).

S modelem typu (1) se setkáváme v mnohých oblastech. Někdy počet rušivých parametrů řádově převyšuje počet užitečných parametrů a to může vést k numerickým potížím.

2 Jednorozměrný lineární model s rušivými parametry

2.1 Regulární model

Tvrzení 2.1.1
V regulárním modelu (1) platí pro ϑ-LBLUE parametru $(\beta', \kappa')'$

$$\begin{pmatrix} \hat{\beta} \\ \hat{\kappa} \end{pmatrix} = \begin{pmatrix} (\mathrm{X}'\Sigma_\vartheta^{-1}\mathrm{M}_S^{\Sigma_\vartheta^{-1}}\mathrm{X})^{-1}\mathrm{X}'\Sigma_\vartheta^{-1}\mathrm{M}_S^{\Sigma_\vartheta^{-1}} \\ (\mathrm{S}'\Sigma_\vartheta^{-1}\mathrm{S})^{-1}\mathrm{S}'\Sigma_\vartheta^{-1}\mathrm{M}_X^{\Sigma_\vartheta^{-1}M_S^{\Sigma_\vartheta^{-1}}} \end{pmatrix} \mathrm{Y}. \tag{2}$$

$$\begin{aligned} var(\hat{\beta}) &= \mathrm{C}^{-1}, \\ var(\hat{\kappa}) &= (\mathrm{S}'\Sigma_\vartheta^{-1}\mathrm{S})^{-1} + (\mathrm{S}'\Sigma_\vartheta^{-1}\mathrm{S})^{-1}\mathrm{S}'\Sigma_\vartheta^{-1}\mathrm{X}\mathrm{C}^{-1}\mathrm{X}'\Sigma_\vartheta^{-1}\mathrm{S}(\mathrm{S}'\Sigma_\vartheta^{-1}\mathrm{S})^{-1}, \end{aligned}$$

kde $\mathrm{C} = \mathrm{X}'(\mathrm{M}_S\Sigma_\vartheta\mathrm{M}_S)^+\mathrm{X}$, $(\mathrm{M}_S\Sigma_\vartheta\mathrm{M}_S)^+ = \Sigma_\vartheta^{-1} - \Sigma_\vartheta^{-1}\mathrm{S}(\mathrm{S}'\Sigma_\vartheta^{-1}\mathrm{S})^{-1}\mathrm{S}'\Sigma_\vartheta^{-1}$.

Důkaz viz [6], Theorem 1

Následující tvrzení se týkají stejnoměrně nejlepších nevychýlených lineárních odhadů (UBLUE) parametrických funkcí v regresním modelu s rušivými parametry. Jsou dokázány v práci [6].

Tvrzení 2.1.2
V regulárním lineárním modelu (1) je statistika $\mathrm{g}'\mathrm{Y}$ UBLUE *odhad své střední hodnoty* $\mathrm{g}'(\mathrm{X}\beta + \mathrm{S}\kappa)$ *právě když*

$$\mathrm{g} \in Ker\left[\sum_{i=1}^{p} V_i M_{(X,S)} V_i\right] = Ker\left[\sum_{i=1}^{p} V_i M_S M_X^{M_S} V_i\right] = Ker\left[\sum_{i=1}^{p} V_i M_X M_S^{M_X} V_i\right]. \tag{3}$$

Tvrzení 2.1.3
V regulárním modelu (1) existuje UBLUE lineární funkce

$$\mathrm{f}'\begin{pmatrix} \beta \\ \kappa \end{pmatrix} = (\mathrm{f}_1', \mathrm{f}_2')\begin{pmatrix} \beta \\ \kappa \end{pmatrix} = \mathrm{f}_1'\beta + \mathrm{f}_2'\kappa$$

právě tehdy, když

$$\mathrm{f} \in \mathcal{M}\left[\begin{pmatrix} \mathrm{X}' \\ \mathrm{S}' \end{pmatrix}\mathrm{B}\right], \quad \text{kde} \quad \mathcal{M}(\mathrm{B}) = Ker\left[\sum_{i=1}^{p} V_i M_X M_S^{M_X} V_i\right].$$

Tvrzení 2.1.4
V regulárním lineárním modelu (1) existuje UBLUE lineární funktce $f_1'\beta$ užitečných parametrů právě tehdy, když

$$\mathrm{f}_1 \in \mathcal{M}(\mathrm{X}'\mathrm{M}_S\mathrm{C}), \quad \text{kde } \mathcal{M}(\mathrm{C}) = Ker\left[\sum_{i=1}^{p} \mathrm{V}_i\mathrm{M}_S\mathrm{M}_X^{M_S}\mathrm{V}_i\mathrm{M}_S\right].$$

2.2 Lineární model s rušivými parametry s podmínkou typu I

V praxi se vyskytují situace, kdy v regresním lineárním modelu s užitečnými a rušivými parametry jsou dána jistá omezení na užitečné parametry, na rušivé parametry se žádné podmínky nekladou.

Definice 2.2.1 (viz [2], str.57]) Je-li parametrickým prostorem parametru β místo R^{k_1} v lineárním modelu (1) pouze lineární varieta

$$\mathcal{B} = \{\beta : \beta \in R^{k_1}, \mathrm{b} + \mathrm{B}\beta = \mathrm{o}\}, \tag{4}$$

kde B je daná $q \times k_1$ matice a $\mathrm{b} \in \mathcal{M}(\mathrm{B})$ je daný q-rozměrný vektor, $r(\mathrm{B}) = q < k_1$, potom se model nazývá **model s podmínkami typu I na užitečné parametry**.

Tvrzení 2.2.2 *V regulárním modelu (1) s podmínkami (4) platí pro ϑ-LBLUE odhady parametrů β a κ*

$$\hat{\hat{\beta}} = [\mathrm{Y} - \mathrm{C}^{-1}\mathrm{B}'(\mathrm{B}\mathrm{C}^{-1}\mathrm{B}')^{-1}\mathrm{B}]\hat{\beta} - \mathrm{C}^{-1}\mathrm{B}'(\mathrm{B}\mathrm{C}^{-1}\mathrm{B})^{-1}\mathrm{b},$$

$$\hat{\hat{\kappa}} = \hat{\kappa} + (\mathrm{S}'\Sigma_\vartheta^{-1}\mathrm{S})^{-1}\mathrm{S}'\Sigma_\vartheta^{-1}\mathrm{X}\mathrm{C}^{-1}\mathrm{B}'(\mathrm{B}\mathrm{C}^{-1}\mathrm{B}')^{-1}(\mathrm{b} + \mathrm{B}\hat{\beta}), \mathrm{C} = \mathrm{X}'(\mathrm{M}_S\Sigma_\vartheta\mathrm{M}_S)^{+}\mathrm{X},$$

kde $\hat{\beta}, \hat{\kappa}$ jsou odhady v regulárním modelu bez podmínek (viz Tvrzení 2.1.1). Varianční matice odhadu $\hat{\hat{\beta}}$ je dána vztahem $var(\hat{\hat{\beta}}) = [\mathrm{M}_{B'}\mathrm{C}\mathrm{M}_{B'}]^{+}$.
Důkaz viz [7], Theorem 2

Tvrzení 2.2.3 *V regulárním modelu (1) s podmínkami (4) na užitečné parametry je statistika $\mathrm{g}'\mathrm{Y}$ UBLUE odhad své střední hodnoty právě tehdy, když*

$$\mathrm{g} \in \mathcal{K} = Ker\left(\sum_{i=1}^{p}\mathrm{V}_i\mathrm{M}_{(X,S)}\mathrm{V}_i + \sum_{i=1}^{p}\mathrm{V}_i\mathrm{M}_S\mathrm{P}^{M_S}_{X(X'M_SX)^{-1}B'}\mathrm{V}_i\right). \tag{5}$$

Důkaz viz [7], Theorem 3

Označení 2.2.4 Označme N takovou matici, že

$$\mathcal{M}(\mathrm{N}) = \mathcal{K} = Ker\left[\sum_{i=1}^{p}\mathrm{V}_i\mathrm{M}_{(X,S)}\mathrm{V}_i + \sum_{i=1}^{p}\mathrm{V}_i(\mathrm{M}_S\mathrm{P}^{M_S}_{X(X'M_SX^{-1}B')})\mathrm{V}_i\right].$$

Tvrzení 2.2.5 *V regulárním lineárním modelu* (1) *s podmínkami* (4) *má funkce*

$$\mathrm{f}'\begin{pmatrix}\beta\\ \kappa\end{pmatrix} = (\mathrm{f}_1', \mathrm{f}_2')\begin{pmatrix}\beta\\ \kappa\end{pmatrix}$$

svůj UBLUE odhad právě tehdy, když

$$\mathrm{f} = \begin{pmatrix}\mathrm{f}_1\\ \mathrm{f}_2\end{pmatrix} \in \mathcal{M}\begin{pmatrix}\mathrm{X}'\mathrm{N}, & \mathrm{B}'\\ \mathrm{S}'\mathrm{N}, & \mathrm{O}\end{pmatrix}. \tag{6}$$

Důkaz viz [7], Theorem 4

2.3 Lineární model s rušivými parametry s podmínkami typu II

Uvažujme stejný lineární model (1) jako v předchozích odstavcích, pouze s tou změnou, že vektor užitečných parametrů označíme symbolem β_1.

$$\mathrm{Y} = (\mathrm{X}, \mathrm{S}) \begin{pmatrix} \beta_1 \\ \kappa \end{pmatrix} + \varepsilon, \tag{7}$$

Předpokládejme opět, že jsou *splněny podmínky regularity.*

Existují situace, kdy do podmínek kladených na užitečné parametry vstupují ještě neobservovatelné parametry. Potom se vektor neznámých parametrů $(\beta_1', \kappa', \beta_2')'$ skládá ze tří podvektorů. Střední hodnota observačního vektoru je rovna $E(\mathrm{Y}) = \mathrm{X}\beta_1 + \mathrm{S}\kappa$, podmínky kladené na model obsahují užitečné parametry a vektor neobservovatelných parametrů β_2 dimenze k.

Studuje se lineární model (7) s podmínkami ve tvaru (viz [4],str.129])

$$\mathrm{b} + \mathrm{B}_1\beta_1 + \mathrm{B}_2\beta_2 = \mathrm{o}, \tag{8}$$

kde B_1 resp. B_2 jsou dané $q \times k_1$ resp. $q \times k$ matice a kde $\mathrm{b} \in \mathcal{M}(\mathrm{B}_1, \mathrm{B}_2)$ je daný q-rozměrný vektor. Tento model se nazývá **model nepřímého měření s podmínkami typu II**.

Předpokládejme, že $r(\mathrm{B}_1, \mathrm{B}_2) = q < k_1 + k, \quad r(\mathrm{B}_2) = k < q$. V práci [8] jsou dokázána následující tvrzení.

Tvrzení 2.3.1 *V regulárním modelu (7) s podmínkami (8) platí pro ϑ-LBLUE odhady parametrů β_1, κ a β_2 tyto vztahy*

$$\hat{\hat{\beta}}_1 = [\mathrm{Y} - \mathrm{C}^{-1}\mathrm{B}_1'\mathrm{Q}\mathrm{B}_1]\hat{\beta}_1 - \mathrm{C}^{-1}\mathrm{B}_1'\mathrm{Q}\mathrm{b}, \quad \hat{\hat{\beta}}_2 = -\mathrm{D}\mathrm{B}_1\hat{\beta}_1 - \mathrm{D}\mathrm{b},$$

$$\hat{\hat{\kappa}} = \hat{\kappa} + (\mathrm{S}'\Sigma_\vartheta^{-1}\mathrm{S})^{-1}\mathrm{S}'\Sigma_\vartheta^{-1}\mathrm{X}\mathrm{C}^{-1}\mathrm{B}_1'\mathrm{Q}(\mathrm{b} + \mathrm{B}_1\hat{\beta}_1),$$

kde $\hat{\beta}_1, \hat{\kappa}$ jsou odhady v regulárním modelu bez podmínek (viz Tvrzení 2.1.1),

$$\mathrm{C} = \mathrm{X}'(\mathrm{M}_S\Sigma_\vartheta\mathrm{M}_S)^+\mathrm{X}, \quad \mathrm{Q} = (\mathrm{M}_{B_2}\mathrm{B}_1\mathrm{C}^{-1}\mathrm{B}_1'\mathrm{M}_{B_2})^+.$$

$$\mathrm{D} = [\mathrm{B}_2'(\mathrm{B}_1\mathrm{C}^{-1}\mathrm{B}_1' + \mathrm{B}_2\mathrm{B}_2')^{-1}\mathrm{B}_2]^{-1}\mathrm{B}_2'(\mathrm{B}_1\mathrm{C}^{-1}\mathrm{B}_1' + \mathrm{B}_2\mathrm{B}_2')^{-1}.$$

Tvrzení 2.3.2 *V regulárním modelu (7) s podmínkami (8) na užitečné parametry a na neobservovatelné parametry je statistika* $\mathrm{g}'\mathrm{Y}$ *UBLUE odhadem své střední hodnoty právě když*

$$\mathrm{g} \in \mathcal{K} = Ker\left(\sum_{i=1}^{p}\mathrm{V}_i\mathrm{M}_{(X,S)}\mathrm{V}_i + \sum_{i=1}^{p}\mathrm{V}_i\mathrm{M}_S\mathrm{P}^{M_S}_{X(X'M_SX)^{-1}B_1'M_{B_2}}\mathrm{V}_i\right). \tag{9}$$

Označení 2.3.3 Nechť N je taková matice, že

$$\mathcal{M}(\mathrm{N}) = \mathcal{K} = Ker\left[\sum_{i=1}^{p}\mathrm{V}_i\mathrm{M}_{(X,S)}\mathrm{V}_i + \sum_{i=1}^{p}\mathrm{V}_i(\mathrm{M}_S\mathrm{P}^{M_S}_{X(X'M_SX)^{-1}B_1'M_{B_2}})\mathrm{V}_i\right].$$

Tvrzení 2.3.4 *V regulárním lineárním modelu (7) s podmínkami (8) má funkce*

$$\mathrm{f}_1'\beta_1 + \mathrm{f}_2'\beta_2 + \mathrm{f}_3'\kappa$$

svůj UBLUE odhad právě tehdy, když

$$\begin{pmatrix} \mathrm{f}_1 \\ \mathrm{f}_2 \\ \mathrm{f}_3 \end{pmatrix} \in \mathcal{M} \begin{pmatrix} \mathrm{X}'\mathrm{N}, & \mathrm{B}_1' \\ \mathrm{O}, & \mathrm{B}_2' \\ \mathrm{S}'\mathrm{N}, & \mathrm{O} \end{pmatrix}. \tag{10}$$

Všechna tvrzení v této kapitole uvedená byla vyslovena pro regulární lineární model s rušivými parametry. Obdobná tvrzení lze vyslovit i pro případ, kdy se regularita modelu nepředpokládá.

V praxi bývá počet rušivých parametrů někdy i řádově větší, než počet parametrů užitečných a výpočty pomocí uvedených vzorců mohou působit nemalé potíže.

Existují dva základní přístupy k problematice rušivých parametrů.
a) *přístup strukturální*: respektuje se struktura modelu a hledá se třída takových lineárních funkcí užitečných parametrů, jejichž odhad určený při zanedbání rušivých parametrů zůstává nestranný i v úplném modelu. Obdobně se požaduje, aby rozptyl odhadu funkce z této třídy byl stejný jak v modelu s rušivými parametry tak v modelu, kde rušivé parametry neuvažujeme. Určení takových tříd má velký význam pro praxi, při výpočtech odhadů lineárních funkcí užitečných parametrů z této třídy můžeme v modelu (1) vynechat matici S a zabývat se modelem

$$\mathrm{Y} \sim [\mathrm{X}\beta, \Sigma_\vartheta], \quad \beta \in R^{k_1}. \tag{11}$$

b) *přístup eliminační*: vhodnou transformací původního modelu dosáhneme toho, že se v novém modelu již rušivé parametry nevyskytují. Nesmíme ovšem tímto postupem ztratit informaci o užitečných parametrech, transformovaný model musí umožňovat stejně kvalitní odhady užitečných parametrů jako model původní.

Poznámka 2.3.5 Pro jednoduchost budeme modelu (1) říkat „velký“, modelu (11) budeme říkat „malý“.

2.4 Strukturální přístup k rušivým parametrům

Předpokládejme, že varianční matice Σ_ϑ vektoru pozorování je známá pozitivně semidefinitní matice, odhadujeme pouze parametry β, κ, tzv. parametry 1. řádu.

Tvrzení 2.4.1 *Lineární funkcionál $g'\beta$, $\beta \in R^{k_1}$, je velkém modelu (1) nestranně odhadnutelný (tj. existuje lineární funkce $l'Y$ taková, že $E(l'Y) = g'\beta, \forall\beta \in R^{k_1}$, $\forall\kappa \in R^{k_2}$) právě když* $\mathrm{g} \in \mathcal{M}(\mathrm{X}'\mathrm{M}_S)$.
Důkaz viz [9], Remark 2

Funkcionál $\mathrm{g}'\beta$ je v malém modelu (11) nestranně odhadnutelný právě když $\mathrm{g} \in \mathcal{M}(\mathrm{X}')$.

Označení 2.4.2 Označme symbolem $\mathcal{E}_a$ resp. $\mathcal{E}$ třídu všech lineárních funkcí $\mathrm{g}'\beta$, které jsou nestranně odhadnutelné v modelu (1) resp. v modelu (11).

Tvrzení 2.4.3 *Pro třídy* $\mathcal{E}_a,\ \mathcal{E}$ *platí*
$mathcalE_a \subset \mathcal{E}$,

$$\mathrm{p}E_a = \mathcal{E} \quad \mathrm{L}eftrightarrow \quad \mathcal{M}(\mathrm{X}) \cap \mathcal{M}(\mathrm{S}) = \{\mathrm{o}\}.$$

Důkaz [2], str.174

Označení 2.4.4 Označme symboly $\widehat{\mathrm{g}'\beta}_a$ resp. $\widehat{\mathrm{g}'\beta}$ Σ_ϑ-LBLUE odhady funkce $\mathrm{g}'\beta$ ve velkém modelu (1) resp. v malém modelu (11). Index a bude signalizovat, že uvažovaný model je „velký“, tj. s rušivými parametry. Předpokládejme, že

$$\mathcal{M}(\mathrm{X}, \mathrm{S}) \subset \mathcal{M}(\Sigma_\vartheta). \tag{12}$$

Tvrzení 2.4.5 *Za předpokladu (12) platí*

$$\widehat{\mathrm{g}'\beta} = \mathrm{g}'(\mathrm{X}'\Sigma_\vartheta^-\mathrm{X})^-\mathrm{X}'\Sigma_\vartheta^-\mathrm{Y},\ var[\widehat{\mathrm{g}'\beta}] = \mathrm{g}'(\mathrm{X}'\Sigma_\vartheta^-\mathrm{X})^-\mathrm{g},\ \mathrm{g} \in \mathcal{M}(\mathrm{X}'),$$

$$\widehat{\mathrm{g}'\beta}_a = \mathrm{g}'\left[(\mathrm{X}'\Sigma_\vartheta^-\mathrm{X})^-\mathrm{X}'\Sigma_\vartheta^- - (\mathrm{X}'\Sigma_\vartheta^-\mathrm{X})^-\mathrm{X}'\Sigma_\vartheta^-\mathrm{S}[\mathrm{S}'\Sigma_\vartheta^-\mathrm{M}_X^{\Sigma_\vartheta^-}\mathrm{S}]^-\mathrm{S}'\Sigma_\vartheta^-\mathrm{M}_X^{\Sigma_\vartheta^-}\right]\mathrm{Y}, \tag{13}$$
$$\mathrm{g} \in \mathcal{M}(\mathrm{X}'\mathrm{M}_S).$$

$$var[\widehat{\mathrm{g}'\beta}_a] = \mathrm{g}'[(\mathrm{X}'\Sigma_\vartheta^-\mathrm{X})^- + (\mathrm{X}'\Sigma_\vartheta^-\mathrm{X})^-\mathrm{X}'\Sigma_\vartheta^-\mathrm{S}[\mathrm{S}'\Sigma_\vartheta^-\mathrm{M}_X^{\Sigma_\vartheta^-}\mathrm{S}]^-\mathrm{S}'\Sigma_\vartheta^-\mathrm{X}(\mathrm{X}'\Sigma_\vartheta^-\mathrm{X})^-]\mathrm{g}.$$

Důkaz [2], str.177

Označení 2.4.6 Označme symbolem $\mathcal{E}_0$ takovou podtřídu třídy $\mathcal{E}_a$, která obsahuje všechny funkce $\mathrm{g}'\beta$, jejichž BLUE v modelu (1) má stejný rozptyl jako v modelu (11).

Tvrzení 2.4.7

$$\mathrm{g}'\beta \in \mathcal{E}_0 \quad \Leftrightarrow \quad \mathrm{g}'(\mathrm{X}'\Sigma_\vartheta^-\mathrm{X})^-\mathrm{X}'\Sigma_\vartheta^-\mathrm{S} = \mathrm{o}.$$

Důkaz [4], Věta V.2.1.6

Tvrzení 2.4.8
1. *Třída* $\mathcal{E}_0$ *je podprostorem* $\mathcal{E}_a$,
2. $\mathcal{E}_0 = \{\mathrm{g}'\beta, \beta \in R^{k_1} : \mathrm{g} \in \mathcal{M}(\mathrm{X}'\Sigma_\vartheta^-\mathrm{X}\mathrm{M}_{X'\Sigma_\vartheta^-S})\}$.

Důkaz viz [2], Theorem 4.1.4

Tvrzení 4.2.9 *Jiné vyjádření pro třídu* $\mathcal{E}_0$ *je* $\mathcal{E}_0 = \{\mathrm{g}'\beta : \widehat{\mathrm{g}'\beta} = \widehat{\mathrm{g}'\beta}_a\}$.
Důkaz [2], Theorem 4.1.6

Tvrzení 2.4.10

$$dim\mathcal{E}_0 = r(\mathrm{X}) - r(\mathrm{X}'\Sigma_\vartheta^-\mathrm{S}) = dim\mathcal{E}_a - \{r(\mathrm{X}'\Sigma_\vartheta^-\mathrm{S}) - dim[\mathcal{M}(\mathrm{X}) \cap \mathcal{M}(\mathrm{S})]\}.$$

Důkaz viz [2], Theorem 4.1.7

Z posledního tvrzení plyne, že

$$\mathcal{E}_0 = \mathcal{E}_a \quad \Leftrightarrow \quad r(\mathrm{X}'\Sigma_\vartheta^- \mathrm{S}) = \dim[\mathcal{M}(\mathrm{X}) \cap \mathcal{M}(\mathrm{S})].$$

Poznámka 2.4.11 To, zda daná lineární funkce užitečných parametrů patří do třídy $\mathcal{E}_0$, ověříme pomocí identity $\mathrm{AA}^-\mathrm{g} = \mathrm{g} \quad \Leftrightarrow \quad \mathrm{g} \in \mathcal{M}(\mathrm{A})$. V našem případě ověříme zda platí $\mathrm{X}'\Sigma_\vartheta^- \mathrm{XM}_{X'\Sigma^- S}[\mathrm{X}'\Sigma_\vartheta^- \mathrm{XM}_{X'\Sigma^- S}]^-\mathrm{g} = \mathrm{g}$.
Jestliže $\mathrm{g} \notin \mathcal{M}(\mathrm{X}'\Sigma_\vartheta^- \mathrm{XM}_{X'\Sigma_\vartheta^- S})$, je nutno při odhadu funkce $\mathrm{g}'\beta, \beta \in R^{k_1}$, uvážit i rušivé parametry, tj. užít odhad (13).

Místo podmínky (12) je možno předpokládat splnění

$$\mathcal{M}(\mathrm{S}) \ \subset \mathcal{M}(\Sigma_\vartheta + \mathrm{XX}'),$$

potom ve vzorcích pro odhady a ve vyjádření třídy $\mathcal{E}_0$ vystupuje místo matice Σ_ϑ matice $\mathrm{T} = \Sigma_\vartheta + \mathrm{XX}'$.

Nerozřešenou úlohou strukturálního přístupu k rušivým parametrům je problém invariance podprostoru $\mathcal{E}_0$ na parametry 2. řádu ϑ, $\Sigma_\vartheta = \sum_{i=1}^p \vartheta_i \mathrm{V}_i$.

Zajímavé problémy se objevují v souvislosti s tzv. podparametrizováním modelu (kdy je ve skutečnosti správný velký model, ale odhady užitečných parametrů počítáme v malém modelu) nebo přeparametrizováním modelu (ve skutečnosti platí malý model, ale užíváme vzorce pro odhady užitečných parametrů z velkého modelu) v případě, kdy lineární funkce $\mathrm{g}'\beta$ nepatří do třídy $\mathcal{E}_0$. Řešení jsou uvedena v připravované monografii Fišerová, E., Kubáček,L., Kunderová,P.:*Linear statistical models: regularity and singularities.*

2.5 Eliminační přístup

Uvažujme model (1) s rušivými parametry kde Σ_ϑ je známá.

a) Třída $\mathcal{T}_1$ eliminačních matic

Nejprve budeme studovat třídu eliminačních matic

$$\mathcal{T}_1 = \{\mathrm{T} : \mathrm{TX} = \mathrm{X}, \mathrm{TS} = \mathrm{O}\},$$

kde T je vhodná matice příslušného rozměru. Volbou matice $\mathrm{T} \in \mathcal{T}_1$ dostaneme transformovaný model $(\mathrm{TY}, \mathrm{X}\beta, \mathrm{T}\Sigma_\vartheta \mathrm{T}')$, ve kterém se nezměnila matice plánu příslušná užitečným parametrům. Úlohou je najít takovou matici $\mathrm{T} \in \mathcal{T}_1$, jejíž užití nezpůsobí ztrátu informace o užitečných parametrech.

Tvrzení 2.5.1 *Eliminační transformace* $\mathrm{T} \in \mathcal{T}_1$ *v modelu (1) existuje právě tehdy, když* $\mathcal{M}(\mathrm{X}) \cap \mathcal{M}(\mathrm{S}) = \{\mathrm{o}\}$.
Důkaz viz [2], Theorem 4.2.1

V následujících úvahách budeme předpokládat splnění podmínky $\mathcal{M}(\mathrm{X}) \cap \mathcal{M}(\mathrm{S}) = \{\mathrm{o}\}$. Potom jsou funkce $\mathrm{g}'\beta, \mathrm{g} \in \mathcal{M}(\mathrm{X}')$, nestranně odhadnutelné ve velkém i malém modelu, protože $\mathcal{M}(\mathrm{X}') = \mathcal{M}(\mathrm{X}'\mathrm{M}_S)$.

Tvrzení 2.5.2 *Třídu* $\mathcal{T}_1$ *eliminačních matic lze vyjádřit ve tvaru*

$$\mathcal{T}_1 = \{\mathrm{T} = (\mathrm{X}, \mathrm{O})(\mathrm{X}, \mathrm{S})^- + \mathrm{Z} - \mathrm{Z}(\mathrm{X}, \mathrm{S})(\mathrm{X}, \mathrm{S})^- : \mathrm{Z} \text{ libovolná}\}.$$

Zvolíme-li za matici $(\mathrm{X},\mathrm{S})^-$ *matici* $[(\mathrm{X},\mathrm{S})'\mathrm{W}^{-1}(\mathrm{X},\mathrm{S})]^-(\mathrm{X},\mathrm{S})'\mathrm{W}^{-1}$, W p.d., *dostaneme třídu*

$$\mathcal{T}_1 = \{(\mathrm{Y}-\mathrm{Z})\mathrm{P}_X^{(M_S W M_S)^+} + \mathrm{Z}\mathrm{M}_S^{(M_X W M_X)^+} : \mathrm{Z} \text{ libovolná}\}.$$

Položíme-li $\mathrm{Z}=\mathrm{Y}$, *dostaneme transformační matici*

$$\mathrm{T} = \mathrm{M}_S^{(M_X W M_X)^+} = \mathrm{Y} - \mathrm{S}[\mathrm{S}'(\mathrm{M}_X\mathrm{W}\mathrm{M}_X)^+\mathrm{S}]^-\mathrm{S}'(\mathrm{M}_X\mathrm{W}\mathrm{M}_X)^+.$$

Důkaz viz [2], Theorem 4.2.3 nebo [4], Věta V.2.2.1

Definice 2.5.3 *Eliminační matice ze třídy* $\mathcal{T}_1$*, pro které s pravděpodobností 1 platí*

$$\mathrm{X}\left[(\mathrm{X}')_{m(T\Sigma_\vartheta T')}\right]'\mathrm{TY} = (\mathrm{X},\mathrm{O})\left\{[(\mathrm{X},\mathrm{S})']^-_{m(\Sigma_\vartheta)}\right\}'\mathrm{Y}, \tag{14}$$

se nazývají optimální vzhledem k užitečným parametrům.

Užití optimální transformace T z předchozí definice zaručuje, že odhad BLUE nevychýleně odhadnutelné funkce $\mathrm{g}'\beta$ v modelu (1) je s pravděpodobností 1 identický s odhadem v transformovaném modelu $(\mathrm{TY},\mathrm{X}\beta,\mathrm{T}\Sigma_\vartheta\mathrm{T}')$.

O optimální transformační matici lze vyslovit následující dvě věty (viz [2], Theorem 4.2.8, Theorem 4.2.9)

Tvrzení 2.5.4 *Nechť* W *je libovolná, pevně zvolená* $n\times n$ *pozitivně definitní matice. Eliminační matice* T *je v modelu (1) optimální vzhledem k užitečným parametrům právě když*

$$\mathrm{M}^{W^{-1}}_{(X,S)}\Sigma_\vartheta\mathrm{T}'(\mathrm{X}')^-_{m(T\Sigma_\vartheta T')}\mathrm{X}' = \mathrm{O}. \tag{15}$$

Tvrzení 2.5.5 *Eliminační matice* T, *splňující vztah* $\mathrm{T}=\mathrm{Y}-\mathrm{SC}$, *kde* C *je libovolná* $k_2\times n$ *matice pro kterou platí* $\mathrm{CX}=\mathrm{O}$ & $\mathrm{CS}=\mathrm{Y}$, *je ve velkém modelu (1) optimální vzhledem k užitečným parametrům.*

Poznámka 2.5.6 (viz [2], Remark 4.2.10, Corollary 4.2.15)

a) Pro libovolnou pozitivně definitní matici W je matice $\mathrm{M}_S^{(M_X W M_X)^+}$ optimální eliminační maticí vzhledem k užitečným parametrům.
b) Platí-li $\mathrm{T}\Sigma_\vartheta(\mathrm{Y}-\mathrm{T})'=\mathrm{O}$, je T optimální vzhledem k užitečným parametrům.

Velký model (1) lze pro libovolnou matici $\mathrm{T}\in\mathcal{T}_1$ vyjádřit v ekvivalentním tvaru (viz [2], Lemma 4.2.12)

$$\left[\begin{pmatrix}\mathrm{T}\\ \mathrm{Y}-\mathrm{T}\end{pmatrix}\mathrm{Y}, \begin{pmatrix}\mathrm{X}, & \mathrm{O}\\ \mathrm{O}, & \mathrm{S}\end{pmatrix}\begin{pmatrix}\beta\\ \kappa\end{pmatrix}, \begin{pmatrix}\Sigma_{11}, & \Sigma_{12}\\ \Sigma_{21}, & \Sigma_{22}\end{pmatrix}\right],$$

$$\Sigma_{11} = var(\mathrm{TY}),\quad \Sigma_{12} = cov[\mathrm{TY},(\mathrm{Y}-\mathrm{T})\mathrm{Y}] = \Sigma_{21}',\quad \Sigma_{22} = var[(\mathrm{Y}-\mathrm{T})\mathrm{Y}]. \tag{16}$$

Pomocí tohoto modelu lze potom určit BLUE vektorové funkce užitečných parametrů, jak uvádí následující věta (viz [2], Theorem 4.2.13)

Tvrzení 2.5.7 *Pro BLUE vektorové funkce* $\mathrm{X}\beta$ *v modelu (1), kde* $\mathcal{M}(\mathrm{X}) \cap \mathcal{M}(\mathrm{S}) = \{\mathrm{o}\}$, *platí*

$$\widehat{\mathrm{X}\beta} = \mathrm{X}[(\mathrm{X}')^-_{m(A)}]' \left(\mathrm{TY} - \Sigma_{12}\Sigma_{22}^-\{\mathrm{Y} - \mathrm{S}[(\mathrm{S}')^-_{m(\Sigma_{22})}]'\}(\mathrm{Y} - \mathrm{T})\mathrm{Y}\right),$$

kde $\Sigma_{ij}, i, j = 1, 2$ jsou dány vztahy (16) a kde

$$\mathrm{A} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^-[\Sigma_{22} - \Sigma_{22}(\mathrm{S}')^-_{m(\Sigma_{22})}\mathrm{S}']\Sigma_{22}^-\Sigma_{21},$$

Pro odhad vektorové funkce $\mathrm{X}\beta$ byla dokázána následující věta (viz [2], Theorem 4.2.21)

Tvrzení 2.5.8 *Matice* $\mathrm{T} = \mathrm{XX}' \left[(\mathrm{XX}' + \mathrm{SS}')^-_{m(\Sigma_\vartheta)}\right]'$, *je jedinou optimální eliminační maticí takovou, že* TY *je v modelu (1) BLUE odhadem vektorové funkce* $\mathrm{X}\beta$ *skoro všude.*

Eliminační transformace, které jsou optimální vzhledem k parametrům prvního i druhého řádu

Uvažujme model (1) ve kterém má varianční matice speciální strukturu $\Sigma_\vartheta = \sum_{i=1}^p \vartheta_i \mathrm{V}_i,\ \vartheta \in \underline{\vartheta}$. Označme

$$\mathcal{T}_0 = \{\mathrm{T} : \mathrm{T} = \mathrm{Y} - \mathrm{SC},\ \mathrm{CX} = \mathrm{O},\ \mathrm{CS} = \mathrm{Y}\}.$$

Podle Tvrzení 2.5.5 je třída $\mathcal{T}_0$ eliminačních matic optimální vzhledem k užitečným parametrům. Tato třída má dvě další výhodné vlastnosti. Matice $\mathrm{Y} - \mathrm{SC}$ nezávisí na varianční matici Σ_ϑ, tedy T je sdruženě (vzhledem k parametrům druhého řádu) optimální vzhledem k užitečným parametrům prvního řádu. Protože $\mathrm{CX} = \mathrm{O},\ \mathrm{CS} = \mathrm{Y}$, je náhodný vektor CY nevychýleným odhadem vektoru rušivých parametrů κ. Transformace $(\mathrm{Y} - \mathrm{SC})\mathrm{Y} = \mathrm{Y} - \mathrm{S}\hat{\kappa}$ je proto nejpřirozenější eliminační procedurou.

Tvrzení 2.5.9 *Třídu* $\mathcal{T}_0$ *lze zapsat v ekvivalentním tvaru*

$$\mathcal{T}_0 = \{\mathrm{M}_S^{(M_X W M_X)^+} - \mathrm{SVM}_X^{W^{-1}}\mathrm{M}_S^{(M_X W M_X)^+}\},$$

kde V *je libovolná* $k_2 \times n$ *matice*, W *libovolná* $n \times n$ *p.d. matice.*
Důkaz viz [2], Theorem 4.3.5

Tvrzení 2.5.10 *Označme* $\Sigma_0 = \sum_{i=1}^p \vartheta_{0,i}\mathrm{V}_i$.
a) V regulárním modelu (1) je funkce $\mathrm{f}'\vartheta = \sum_{i=1}^p \mathrm{f}_i\vartheta_i,\ \vartheta \in \underline{\vartheta}$, *nevychýleně, kvadraticky a invariantně odhadnutelná (t.j. odhad je tvaru* $\mathrm{Y}'\mathrm{AY}$, *kde* $\mathrm{A}_{n,n}$ *je symetrická matice, odhad je invariantní na změnu vektoru* β*) právě tehdy, když* $\mathrm{f} \in \mathcal{M}(\mathrm{S}_{(M_{(X,S)}\Sigma_0 M_{(X,S)})^+})$, kde

$$\{\mathrm{S}_{(M_{(X,S)}\Sigma_0 M_{(X,S)})^+}\}_{i,j} = Tr[(\mathrm{M}_{(X,S)}\Sigma_0\mathrm{M}_{(X,S)})^+\mathrm{V}_i(\mathrm{M}_{(X,S)}\Sigma_0\mathrm{M}_{(X,S)})^+\mathrm{V}_j],$$

$$i, j = 1, \ldots, p.$$

b) Splňuje-li funkce $\mathrm{f}'\vartheta$ *podmínku z a), potom pro odhad* ϑ_0*-MINQUE funkce* $\mathrm{f}'\vartheta$ *platí*

$$\widehat{\mathrm{f}'\vartheta} = \sum_{i=1}^{p} \lambda_i \mathrm{Y}'(\mathrm{M}_{(X,S)}\Sigma_0\mathrm{M}_{(X,S)})^{+}\mathrm{V}_i(\mathrm{M}_{(X,S)}\Sigma_0\mathrm{M}_{(X,S)})^{+}\mathrm{Y},$$

kde vektor λ *je řešením soustavy rovnic* $\mathrm{S}_{(M_{(X,S)}\Sigma_0 M_{(X,S)})^{+}}\lambda = \mathrm{f}$.
Důkaz viz [4], Důsledek V.2.2.4

Poznámka 2.5.11 Matice $S_{(M_{(X,S)}\Sigma_0 M_{(X,S)})^{+}}$ se nazývá kriteriální matice pro odhadnutelnost funkce $\mathrm{f}'\vartheta$.

Poznámka 2.5.12 V [4], str. 153 je dokázáno, že transformační matice $\mathrm{T} = \mathrm{M}_S^{(M_X\Sigma M_X)^{+}}$ je jedna z možných eliminačních matic, které eliminují rušivé parametry a současně můžeme v eliminovaném modelu určit stejně přesně (tj. se stejnou disperzí) jak odhady funkce parametru β tak odhady funkce parametru ϑ, přičemž třída nestranně odhadnutelných funkcí se eliminační transformací nezmění. Podstatnou roli zde hraje předpoklad $\mathcal{M}(\mathrm{X}) \cap \mathcal{M}(\mathrm{S}) = \{o\}$ a skutečnost, že odhad funkcí parametru ϑ realizujeme pomocí ϑ-MINQUE metody.

V knize [2] (Theorem 4.3.7, Theorem 4.3.8) jsou dokázána následující zajímavá tvrzení

Tvrzení 2.5.13 *Lineární funkce* $\mathrm{f}'\vartheta$ *vektorového parametru* $\vartheta \in \underline{\vartheta} \subset R^p$, *která je nevychýleně a invariantně odhadnutelná v modelu (1) před provedením eliminační transformace, je nevychýleně a invariantně odhadnutelná v transformovaném modelu*

$$\left(\mathrm{TY}, \mathrm{X}\beta, \sum_{i=1}^{p} \vartheta_i \mathrm{TV}_i\mathrm{T}'\right), \tag{17}$$

jestliže

$$\mathrm{T} \in \mathcal{T}^{\star} = \{\mathrm{M}_S^{(M_X W M_X)^{+}} : \mathrm{W} \text{ libovolná } n \times n \text{ p.d. matice}\}.$$

Tvrzení 2.5.14 *Nechť* $\mathrm{T} \in \mathcal{T}^{\star}$ *a nechť* $\mathrm{f}'\vartheta$, $\vartheta \in \underline{\vartheta}$ *je nevychýleně a invariantně odhadnutelná funkce. Potom je* ϑ_0*-MINQUE odhad funkce* $f'\vartheta$ *v modelu (1) identický s odhadem* ϑ_0*-MINQUE stejné funkce v modelu (17) po eliminaci.*

b) Třída $\mathcal{T}_2$ eliminačních matic

Uvažujme lineární model (1) s rušivými parametry Jestliže budeme brát transformační matice z obecnější třídy $\mathcal{T}_2 = \{\mathrm{T} : \mathrm{TS} = \mathrm{O}\}$, tj. bude-li po transformaci změněna matice plánu příslušná užitečným parametrům, lze získat některé zajímavé výsledky, (zatím byly dokázány pro regulární model).

Po transformaci modelu (1) transformační maticí $\mathrm{T} \in \mathcal{T}_2$ dostaneme lineární model

$$\mathrm{TY} \ \sim \ [\mathrm{TX}\beta, \mathrm{T}\Sigma_\vartheta\mathrm{T}']. \tag{18}$$

Obecné řešení maticové rovnice $\mathrm{TS} = \mathrm{O}$ má tvar $\mathrm{T} = \mathrm{A}(\mathrm{Y} - \mathrm{SS}^-)$, kde A je libovolná matice odpovídajícího typu, S^- je nějaká g-inverze matice S. Zvolíme-li $\mathrm{S}^- = (\mathrm{S}'\mathrm{WS})^-\mathrm{S}'\mathrm{W}$, kde W je libovolná p.s.d. matice taková, že

$$\mathcal{M}(\mathrm{S}') = \mathcal{M}(\mathrm{S}'\mathrm{WS}), \tag{19}$$

potom $\mathrm{T} = \mathrm{AM}_S^W$, kde M_S^W je daná jednoznačně.

Nejprve uvažujme transformační matici $\mathrm{T} = \mathrm{M}_S^W$, t.j. lineární model

$$\mathrm{M}_S^W\mathrm{Y} \ \sim \ [\mathrm{M}_S^W\mathrm{X}\beta, \mathrm{M}_S^W\Sigma_\vartheta(\mathrm{M}_S^W)'], \text{ kde } \Sigma_\vartheta \text{ je regulární.} \tag{20}$$

Lze dokázat, že $\mathcal{M}(\mathrm{M}_S) = \mathcal{M}([\mathrm{M}_S^W]')$, t.j. $\mathcal{M}(\mathrm{X}'\mathrm{M}_S) = \mathcal{M}(\mathrm{X}'(\mathrm{M}_S^W)')$, tedy třídy odhadnutelných funkcí $\mathrm{g}'\beta$ v modelu (1) a v modelu (20) jsou identické.

Tvrzení 2.5.15 *Odhad ϑ-LBLUE odhadnutelné funkce $\mathrm{g}'\beta$, $\mathrm{g} \in \mathcal{M}(\mathrm{X}'\mathrm{M}_S)$ v modelu (20) je stejný jako v regulárním modelu (1).*
Důkaz viz [9], Theorem 3

Tvrzení 2.5.16 *Lineární funkce $f'\vartheta$ vektorového parametru $\vartheta \in \underline{\vartheta} \subset R^p$, která je nevychýleně odhadnutelná v regulárním modelu (1) před eliminací je nevychýleně odhadnutelná i v modelu (20).*
Důkaz viz [9], Theorem 4

Tvrzení 2.5.17 *Nechť $\mathrm{f}'\vartheta$, $\vartheta \in \underline{\vartheta}$ je nevychýleně odhadnutelná funkce. Potom jsou odhad ϑ_0-*MINQUE *v modelu* (1) *a odhad ϑ_0-*MINQUE *v modelu* (20) *po eliminaci shodné.*
Důkaz viz [9], Theorem 5

Uvažujme následující transformovaný lineární model

$$\mathrm{AM}_S^{(M_X\Sigma_\vartheta M_X)^+}\mathrm{Y} \ \sim \ [\mathrm{AX}\beta, \mathrm{AM}_S^{(M_X\Sigma_\vartheta M_X)^+}\Sigma_\vartheta(\mathrm{M}_S^{(M_X\Sigma_\vartheta M_X)^+})'\mathrm{A}'], \Sigma_\vartheta \text{ p.d.}, \tag{21}$$

kde A je taková matice, že

$$\mathcal{M}(\mathrm{X}'\mathrm{A}') = \mathcal{M}(\mathrm{X}'\mathrm{M}_S). \tag{22}$$

Platí

$$E[\mathrm{AP}_X^{(M_S\Sigma_\vartheta M_S)^+}\mathrm{Y}]$$
$$= \mathrm{AX}(\mathrm{X}'[\mathrm{M}_S\Sigma_\vartheta\mathrm{M}_S]^+\mathrm{X})^-\mathrm{X}'[\mathrm{M}_S\Sigma_\vartheta\mathrm{M}_S]^+(\mathrm{X}\beta + \mathrm{S}\kappa) = \mathrm{AX}\beta.$$

Tvrzení 2.5.18 $\mathrm{AP}_X^{(M_S\Sigma_\vartheta M_S)^+}\mathrm{Y}$ *je nejlepším odhadem své střední hodnoty.*
Důkaz viz [9], Lemma 2

Tvrzení 2.5.19 *V modelu (21) tvoří odhady* $\mathrm{AP}_X^{(M_S\Sigma_\vartheta M_S)^+}\mathrm{Y}$, *kde* A *je libovolná matice taková, že $\mathcal{M}(\mathrm{X}'\mathrm{A}') = \mathcal{M}(\mathrm{X}'\mathrm{M}_S)$, třídu všech optimálních odhadů vektorové funkce* $\mathrm{AX}\beta$.
Důkaz viz [9], Theorem 6

V případě transformací volených ze třídy $\mathcal{T}_2$ byla tvrzení vyslovena pro regulární varianční matici. Také zde zbývá dokázat tvrzení pro obecný případ. Pro omezený rozsah článku nejsou uvedena tvrzení o rušivých parametrech v multivariátním lineárním modelu.

Reference

[1] Kubáček L. (1960) *Foundation of estimation theory.* Elsevier, Amsterdam, Oxford, New York, Tokyo.

[2] Kubáček L., Kubáčková L., Volaufová J. (1995). *Statistical models with linear structures.* Veda, Publishing House of the Slovak Academy of Sciences, Bratislava.

[3] Kubáčková L., Kubáček L. (1990). *Elimination transformation of an observation vector preserving information on the first and second order parameters.* Technical Report, No 11. Institute of Geodesy, University of Stuttgart, 1 – 71.

[4] Kubáček L., Kubáčková L. (2000). *Statistika a metrologie.* Univerzita Palackého v Olomouci - vydavatelství.

[5] Kunderová P. (2000). *Linear models with nuisance parameters and deformation measurement.* Acta Univ. Palacki. Olomuc., Mathematica **39**, 95 – 105.

[6] Kunderová P. (2001). *Locally best and uniformly best estimators in linear model with nuisance parameters.* Tatra Mt. Math. Publ. **22**, 27 – 36.

[7] Kunderová P. (2001) *Regular linear model with the nuisance parameters with constraints of the type I.* Acta Univ. Palacki. Olomuc. Mathematica **40**, 151 – 159.

[8] Kunderová P. (2002) *Regular linear model with nuisance parameters with constraints of the type II.* Folia Fac. Sci. Nat. Univ. Masarykianae Brunensis, Mathematica **11**, 151 – 162.

[9] Kunderová P. (2003) *Eliminating transformations for nuisance parameters in linear model.* Acta Univ. Palacki. Olomuc., Mathematica **42**, 59 – 68.

[10] Nordström K., Fellman J. (1990) *Characterizations and dispersion–matrix robustness of efficiently estimable parametric functionals in linear models with nuisance parameters.* Linear Algebra and its Applications **127**, 341 – 361.

Adresa: P. Kunderová, Katedra matematické analýzy a aplikací matematiky, PřF UP Olomouc, Tomkova 40, 779 00 Olomouc

E-mail: `kunderov@inf.upol.cz`

ASYMPTOTIKA ODHADŮ POMOCÍ EMPIRICKÉHO ROZDĚLENÍ

Petr Lachout

Klíčová slova: Data, odhad, empirické rozdělení pravděpodobnosti, statistika, asymptotika odhadu.

Abstrakt: Příspěvek se zabývá asymptotikou odhadů, a to jak bodových, tak intervalových. Jsou uvažovány odhady, které lze reprezentovat jako funkcionály na pravděpodobnostních měrách. Pokud je tento funkcionál slabě spojitý, pak je příslušný bodový odhad konzistentní. Intervalový odhad můžeme získat využití centrální limitní věty, delta věty nebo konvergence procesů.

1 Úvod a značení

Příspěvek se zabývá podmínkami, které zaručují vhodné asymptotické chování odhadů. Obecný postup je takový, že z naměřených údajů napočteme nějakou statistiku, u které vyšetřujeme její asymptotické chování. Potřebuje, aby statistika buď konvergovala k teoretické hodnotě parametru, nebo k nějakému rozdělení pravděpodobnosti. V prvém případě mluvíme o bodovém odhadu a v druhém případě o odhadu intervalovém, potažmo o testování hypotéz. Podívejme se na tuto situaci obecně.

Abychom mohli lépe popsat proces odhadování a testování a abychom do našich úvah zahrnuli, co nejvíce požívaných statistik, volíme poněkud nezvyklou strukturu dat. Naměřená data uvažujeme jako pole indexované dvěma indexy $x_{k,n}$, $k = 1, 2, \ldots, K_n$, $n \in \mathbb{N}$. Jednotlivá pozorování budou hodnoty z prostoru $\mathcal{X} \subset \mathbb{R}^q$, kde $q \in \mathbb{N}$. Pro zjednodušení zápisu budeme symbolem $x_{\bullet,n}$ označovat řádek dat $x_{1,n}, x_{2,n}, \ldots, x_{K_n,n}$.

Pod statistikou rozumíme posloupnost funkcí $S = (S_n, n \in \mathbb{N})$, kde pro každé $n \in \mathbb{N}$ je $S_n : \mathcal{X}^{K_n} \to \mathbb{R}$. Většina používaných statistik však má speciální strukturu. Aby jsme ji mohli lépe popsat a využít, označme

- $\mathcal{M}_1^+(\mathcal{W})$ – všechny borelovské pravděpodobnosti na prostoru $\mathcal{W}$;
- $\mathcal{D}(\mathcal{W})$ – všechny diskrétní pravděpodobnosti na prostoru $\mathcal{W}$, tj. jsou neseny konečně body.

Uvažují se převážně statistiky typu

$$S_n(x_{\bullet,n}) = \tilde{S}(\mathsf{P}_n), \tag{1}$$

kde $\tilde{S} : \mathcal{D}(\mathcal{X}) \to \mathbb{R}$ a $\mathsf{P}_n \in \mathcal{D}(\mathcal{X})$ je empirické rozdělení pravděpodobnosti $x_{\bullet,n}$ definované

$$\mathsf{P}_n(A) = \frac{1}{K_n} \#\left\{k : x_{k,n} \in A,\ k \in \{1, 2, \ldots, K_n\}\right\} \text{ pro každé } A \subset \mathcal{X}. \tag{2}$$

2 Příklady

Uveďme si několik motivačních příkladů.

Příklad 1: Pozorujeme posloupnost reálných čísel $x_1, x_2, x_3, \ldots$. a jako statistiku používáme průměr

$$\frac{1}{n}\sum_{k=1}^{n} x_k. \tag{3}$$

V naší terminologii přepíšeme takto

$$\begin{aligned} K_n &= n, && (4)\\ x_{k,n} &= x_k, && (5)\\ S_n(x_{\bullet,n}) &= \int_{\mathbb{R}} t \ \mathsf{dP}_n(t). && (6) \end{aligned}$$

♣

Příklad 2: Pozorujeme posloupnost reálných čísel $x_1, x_2, x_3, \ldots$. a jako statistiku používáme průměr pozorování transformovaných funkcí $\varphi : \mathbb{R} \to \mathbb{R}$

$$\frac{1}{n}\sum_{k=1}^{n} \varphi(x_k). \tag{7}$$

V naší terminologii přepíšeme takto

$$\begin{aligned} K_n &= n, && (8)\\ x_{k,n} &= x_k, && (9)\\ S_n(x_{\bullet,n}) &= \int_{\mathbb{R}} \varphi(t) \ \mathsf{dP}_n(t). && (10) \end{aligned}$$

♣

Příklad 3: Pozorujeme posloupnost reálných čísel $x_1, x_2, x_3, \ldots$. a pro každé $n \in \mathbb{N}$ uspořádejme prvních n podle velikosti $x_{[1:n]} \le x_{[2:n]} \le \cdots \le x_{[n:n]}$. Pro pevné $\alpha \in (0,1)$ používáme empirický kvantil

$$x_{[\lceil \alpha n \rceil : n]}. \tag{11}$$

V naší terminologii přepíšeme takto

$$\begin{aligned} K_n &= n, && (12)\\ x_{k,n} &= x_k, && (13)\\ S_n(x_{\bullet,n}) &= \inf\{t \in \mathbb{R} : \mathsf{P}_n((-\infty, t)) \ge \alpha\}. && (14) \end{aligned}$$

♣

Příklad 4: Pozorujeme posloupnost reálných čísel $x_1, x_2, x_3, \ldots$. a pro každé $n \in \mathbb{N}$ uspořádejme prvních n podle velikosti $x_{[1:n]} \leq x_{[2:n]} \leq \cdots \leq x_{[n:n]}$. Pro pevné $\alpha \in (0,1)$ používáme useknutý průměr

$$\frac{1}{(1-2\alpha)n} \sum_{k=\lceil \alpha n \rceil}^{\lfloor (1-\alpha)n \rfloor} x_{[k:n]}. \tag{15}$$

V naší terminologii přepíšeme takto

$$\begin{aligned} K_n &= n, && (16)\\ x_{k,n} &= x_k, && (17)\\ S_n(x_{\bullet,n}) &= \frac{1}{1-2\alpha} \int_{\mathbb{R}} t \mathbb{I}_{[\mathsf{P}_n((-\infty,t]) \geq \alpha, \mathsf{P}_n((-\infty,t)) \leq (1-\alpha)]} \, \mathsf{dP}_n(t). && (18) \end{aligned}$$

♣

Příklad 5: Pozorujeme posloupnost reálných čísel $x_1, x_2, x_3, \ldots$. a pro každé $n \in \mathbb{N}$ uspořádejme prvních n podle velikosti $x_{[1:n]} \leq x_{[2:n]} \leq \cdots \leq x_{[n:n]}$. Pro lebesgueovsky integrovatelnou váhovou funkci $c : [0,1] \to \mathbb{R}$ s $\int_0^1 c(t) \, \mathsf{d}t \neq 0$ používáme vážený průměr

$$\left(\sum_{k=1}^{n} c\left(\frac{k}{n}\right) \right)^{-1} \sum_{k=1}^{n} c\left(\frac{k}{n}\right) x_{[k:n]}. \tag{19}$$

V naší terminologii přepíšeme takto

$$\begin{aligned} K_n &= n, && (20)\\ x_{k,n} &= x_k, && (21)\\ S_n(x_{\bullet,n}) &= \gamma(c, \mathsf{P}_n)^{-1} \int_{\mathbb{R}} c\Big(\mathsf{P}_n((-\infty,t])\Big) t \, \mathsf{dP}_n(t), && (22) \end{aligned}$$

kde $\gamma(c, \mathsf{P}_n) = \int_{\mathbb{R}} c\Big(\mathsf{P}_n((-\infty,t])\Big) \, \mathsf{dP}_n(t)$.

♣

Příklad 6: Pozorujeme posloupnost reálných čísel $x_1, x_2, x_3, \ldots$. a jako statistiku používáme empirický moment

$$\frac{1}{n-h} \sum_{k=1}^{n-h} x_k x_{k+h}, \text{ kde } h \in \mathbb{N}. \tag{23}$$

V naší terminologii přepíšeme takto

$$\begin{aligned} K_n &= n-h, && (24)\\ x_{k,n} &= (x_k, x_{k+h}), && (25)\\ S_n(x_{\bullet,n}) &= \int_{\mathbb{R}^2} t_1 t_2 \, \mathsf{dP}_n(t_1, t_2). && (26) \end{aligned}$$

♣

Příklad 7: Pozorujeme posloupnost reálných čísel $x_1, x_2, x_3, \ldots$. a používáme U-statistiku

$$\frac{1}{n^q} \sum_{k_1,k_2,\ldots,k_q=1}^{n} u(x_{k_1}, x_{k_2}, \ldots, x_{k_q}). \tag{27}$$

V naší terminologii přepíšeme takto

$$\begin{aligned} K_n &= n, && (28)\\ x_{k,n} &= x_k, && (29)\\ S_n(x_{\bullet,n}) &= \int_{\mathbb{R}^q} u(t_1, t_2, \ldots, t_q)\ \mathsf{dP}_n(t_1)\ \mathsf{dP}_n(t_2) \ldots\ \mathsf{dP}_n(t_q) && (30)\\ &= \int_{\mathbb{R}^q} u(t_1, t_2, \ldots, t_q)\ \mathsf{dP}_n^{\otimes q}(t_1, t_2, \ldots, t_q). && (31) \end{aligned}$$

♣

Příklad 8: Pozorujeme posloupnost dvojic reálných čísel (x_1, y_1), (x_2, y_2), $(x_3, y_3), \ldots$. Počítáme M-odhad koeficientů lineární regrese

$$(\hat\alpha_n, \hat\beta_n) \in \operatorname{argmin} \left\{ \sum_{k=1}^{n} \rho\left(y_k - a - b x_k\right) :: a, b \in \mathbb{R} \right\}. \tag{32}$$

V naší terminologii přepíšeme takto

$$\begin{aligned} K_n &= n, && (33)\\ x_{k,n} &= (x_k, y_k), && (34)\\ S_n(x_{\bullet,n}) &\in \operatorname{argmin} \left\{ \int_{\mathbb{R}} \rho\left(s - a - bt\right)\ \mathsf{dP}_n(t,s) :: a, b \in \mathbb{R} \right\}. && (35) \end{aligned}$$

♣

Příklad 9: Pozorujeme posloupnost dvojic reálných čísel (x_1, y_1), (x_2, y_2), $(x_3, y_3), \ldots$. Počítáme M-odhad koeficientů lineární regrese

$$(\hat\alpha_n, \hat\beta_n) \in \operatorname{argmin} \left\{ \sum_{k=1}^{n} \rho\left(y_k - a - b x_k\right) :: a, b \in \mathbb{R} \right\}. \tag{36}$$

Dále se zabýváme testování, zda jsou splněny předpoklady modelu. Konkrétně, zda rozdělení chyb odpovídá předpokladům. Používáme statistiku

$$\sum_{k=1}^{n} \rho\left(y_k - \hat\alpha_n - \hat\beta_n x_k\right). \tag{37}$$

V naší terminologii přepíšeme takto

$$\begin{aligned} K_n &= n, && (38)\\ x_{k,n} &= y_k - \hat{\alpha}_n - \hat{\beta}_n x_k, && (39)\\ S_n(x_{\bullet,n}) &= \int_{\mathbb{R}} \rho(t)\ \mathsf{dP}_n(t). && (40) \end{aligned}$$

♣

3 Konvergence

V této kapitole uvažujeme data ve tvaru

- $\mathsf{X} = (x_{k,n}, k = 1, 2, \ldots, K_n, n \in \mathbb{N})$;
- P_n označuje empirické rozdělení řádku $x_{\bullet,n}$.

Budeme se zabývat statistikou tvaru $S : \mathcal{D}(\mathcal{X}) \to \mathbb{R}$ a konvergentností posloupnosti $S(\mathsf{P}_n)$. Jedná se vlastně o vyšetřování hromadných bodů dané statistiky pro daná data, což je

$$\begin{aligned} \mathsf{Ls}\,(S, \mathsf{X}) &= \{\text{všechny hromadné body posloupnosti } S(\mathsf{P}_n), n \in \mathbb{N}\}\\ &= \left\{ \limsup_{k\to+\infty} S(\mathsf{P}_{n_k}), \liminf_{k\to+\infty} S(\mathsf{P}_{n_k}) : n_k \uparrow +\infty \right\}\\ &= \left\{ \limsup_{k\to+\infty} S(\mathsf{P}_{n_k}) : n_k \uparrow +\infty \right\} && (41)\\ &= \left\{ \liminf_{k\to+\infty} S(\mathsf{P}_{n_k}) : n_k \uparrow +\infty \right\}\\ &= \left\{ \lim_{k\to+\infty} S(\mathsf{P}_{n_k}) : n_k \uparrow +\infty,\ \lim_{k\to+\infty} S(\mathsf{P}_{n_k}) \text{ existuje} \right\}. \end{aligned}$$

Konvergenci v pravděpodobnosti dokážeme vyšetřit, pokud lze naši statistiku rozšířit tak, aby $S : \mathcal{M}_1^+(\mathbb{R}^q) \to \mathbb{R}$ a aby byla spojitá vzhledem ke slabé konvergenci měr, a když data splňují podmínku těsnosti, tj.

$$\begin{aligned} &\forall \varepsilon > 0\, \exists \Delta > 0 \text{ tak, že } \forall n \in \mathbb{N} && (42)\\ &\mathsf{P}_n\left(\{x \in \mathcal{X} : \|x\| \le \Delta\}\right) =\\ &\quad = \frac{1}{K_n} \#\left\{k : \|x_{k,n}\| \le \Delta,\ k \in \{1, 2, \ldots, K_n\}\right\} \ge 1 - \varepsilon. \end{aligned}$$

Věta 3.1. *Nechť $S : \mathcal{M}_1^+(\mathbb{R}^q) \to \mathbb{R}$ je spojitá vzhledem ke slabé konvergenci měr. Když data splňují podmínku těsnosti (42), pak*

$$\mathsf{Ls}\,(S, \mathsf{X}) = \left\{ S(\mathsf{P}) : \textit{existuje } n_k \uparrow +\infty \textit{ tak, že } \mathsf{P}_{n_k} \xrightarrow[k\to+\infty]{w} \mathsf{P} \right\}.$$

Důkaz:

1. Pokud $\mathsf{P}_{n_k} \xrightarrow[k\to+\infty]{w} \mathsf{P}$, pak ze slabé spojitosti statistiky S vyplývá, že $\lim_{k\to+\infty} S(\mathsf{P}_{n_k}) = S(\mathsf{P})$. Tudíž $S(\mathsf{P}) \in \mathsf{Ls}\,(S, \mathsf{X})$.

2. Pokud $s \in \mathsf{Ls}\,(S, \mathsf{X})$, pak existuje podposloupnost taková, že $s = \lim_{k\to+\infty} S(\mathsf{P}_{n_k})$.

   Posloupnost P_{n_k}, $k \in \mathbb{N}$ je těsná, neboť těsnost předpokládáme.

   Pak podle Prochorovovy věty existuje podposloupnost $\mathsf{P}_{n_{k_j}}$, $j \in \mathbb{N}$ a pravděpodobnostní míra P tak, že $\mathsf{P}_{n_{k_j}} \xrightarrow[j\to+\infty]{w} \mathsf{P}$.

   Ze slabé spojitosti statistiky S vyplývá, že $s = S(\mathsf{P})$.

Q.E.D.

Nyní si uvědomme, co věta říká pro data, která jsou realizací náhodného schematu $X_{k,n}$, $k = 1, 2, \ldots, K_n$, $n \in \mathbb{N}$; tj.

$$x_{k,n} = X_{k,n}(\omega),\ k = 1, 2, \ldots, K_n,\ n \in \mathbb{N}. \tag{43}$$

Věta 3.2. *Když* $S : \mathcal{M}_1^+(\mathbb{R}^q) \to \mathbb{R}$ *je spojitá vzhledem ke slabé konvergenci měr a pro skoro všechna* $\omega \in \Omega$ *data (43) splňují podmínku těsnosti (42) a* $\mathsf{P}_n(\omega) \xrightarrow[n\to+\infty]{w} \mathsf{P}$, *kde* P *je nenáhodná, pak* $S(\mathsf{P}_n(\omega)) \xrightarrow[n\to+\infty]{} S(\mathsf{P})$ *pro skoro všechna* $\omega \in \Omega$.

Tato situace nastává například pro $X_{k,n} = X_k$, kde X_k, $k \in \mathbb{N}$ jsou i.i.d. nebo ergodické (silně) stacionární posloupnost, viz [4], [2].

Použití věty je však omezeno předpokladem slabé spojitosti funkcionálu S. Uvědomme si, že průměr (3), medián (11), useknutý průměr (15), atd. nejsou reprezentovány slabě spojitým funkcionálem. Naproti tomu průměr transformovaných dat (7), při φ spojité omezené funkci, a vážený průměr (19), při c spojité omezené funkci, jsou reprezentovány slabě spojitým funkcionálem.

4 Rychlost konvergence

V této kapitole pokračujeme v diskusi modelu představeném v předchozích kapitolách. Opět uvažujeme statistiku $S : \mathcal{M}_1^+(\mathbb{R}^q) \to \mathbb{R}$, která je slabě spojitá, a data, která jsou těsná. Navíc předpokládáme, že $\mathsf{P}_n \xrightarrow[n\to+\infty]{w} \mathsf{P}$. Z předchozí kapitoly víme, že potom $\mathsf{Ls}\,(S, \mathsf{X}) = \{S(\mathsf{P})\}$. Zajímá nás jak rychle konverguje k nule rozdíl

$$T(\mathsf{P}_n, \mathsf{P}) = S(\mathsf{P}_n) - S(\mathsf{P}), \quad n \in \mathbb{N}. \tag{44}$$

To znamená, že hledáme $\tau_n > 0$, $\tau_n \to +\infty$ tak, aby $\tau_n T(\mathsf{P}_n, \mathsf{P}) = \mathcal{O}(1)$. Nelze očekávat nenulovou limitu této normalizace. Typicky je třeba očekávat oscilace, plyne to ze zákona iterovaného logaritmu, viz [1], [4].

Tudíž se budeme zabývat pouze daty, které jsou realizací nějakého náhodného schematu (43). Potom $\tau_n T(\mathsf{P}_n, \mathsf{P})$ je náhodná veličina. Její rozdělení pravděpodobnosti budeme označovat $\mathcal{L}\left(\tau_n T(\mathsf{P}_n, \mathsf{P})\right)$

Pro dané $\tau = (\tau_n, n \in \mathbb{N})$ nás proto zajímají hromadné body

$$\begin{aligned} \mathsf{Ls}_2\left(S, \mathsf{X}, \tau\right) &= \left\{\text{hromadné body posloupnosti } \mathcal{L}\left(\tau_n T(\mathsf{P}_n, \mathsf{P})\right), n \in \mathbb{N}\right\} \\ &= \left\{\mu :: \exists n_k \uparrow +\infty,\ \mathcal{L}\left(\tau_{n_k} T(\mathsf{P}_{n_k}, \mathsf{P})\right) \xrightarrow[k\to+\infty]{w} \mu\right\}. \end{aligned} \tag{45}$$

K vyšetření konvergence můžeme využít následující.

4.1 Centrální limitní věta

Pokud uvažovaná statistika je tvaru $S(Q) = \int_{\mathbb{R}^q} f(t)\ \mathsf{d}Q(t)$, pak

$$\tau_n T(\mathsf{P}_n, \mathsf{P}) = \tau_n \sum_{k=1}^{K_n} (f(X_{k,n}) - S(\mathsf{P})). \tag{46}$$

Tento tvar je výhodný pokud data splňují „silný zákon velkých čísel“ a „centrální limitní větu“ pro funkci f. Například pokud $X_{k,n} = X_k$, $k \in \mathbb{N}$, kde X_k, $k \in \mathbb{N}$ jsou i.i.d., to znamená požadovat $\mathsf{E}\left[f(X_1)^2\right] < +\infty$.

4.2 Delta věta

Pokud uvažovaná statistika má Hadamardovu derivaci DS v bodě P, tj. pro každé $Q_n \in \mathcal{M}_1^+(\mathbb{R}^q)$, $n \in \mathbb{N}$, $Q_n \xrightarrow[n\to+\infty]{w} Q$ existuje limita

$$n\left(S\left(P + \frac{1}{n}(Q_n - P)\right) - S(P)\right) \xrightarrow[n\to+\infty]{} DS(P; Q), \tag{47}$$

pak můžeme využít „delta větu“.

Delta věta zaručuje, že když S má Hadamardovu derivaci v bodě P,

$$\tau_n \xrightarrow[n\to+\infty]{} +\infty, \quad \tau_n(\mathsf{P}_n - \mathsf{P}) \xrightarrow[n\to+\infty]{\mathcal{D}} \eta, \tag{48}$$

pak

$$\tau_n T(\mathsf{P}_n, \mathsf{P}) \xrightarrow[n\to+\infty]{\mathcal{D}} DS(P; \eta). \tag{49}$$

4.3 Konvergence procesů

Pokud lze psát $\tau_n T(\mathsf{P}_n, \mathsf{P}) = Z\Big(\tau_n(\mathsf{P}_n(f) - \mathsf{P}(f)), f \in \mathcal{F}\Big)$, kde $\mathcal{F}$ je daný systém reálných funkcí definovaných na $\mathcal{X}$, přičemž $\mathsf{P}(f) = \int_{\mathcal{X}} f\ \mathsf{d}\mathsf{P}$.

Věta 4.1. *Když* $\mathcal{Y}(\mathcal{F}) \subset \mathbb{R}^{\mathcal{F}}$ *je topologický prostor funkcí,*

$$\Big(\tau_n(\mathsf{P}_n(f) - \mathsf{P}(f)), f \in \mathcal{F}\Big) \in \mathcal{Y}(\mathcal{F}), \tag{50}$$

$$\Big(\tau_n(\mathsf{P}_n(f) - \mathsf{P}(f)), f \in \mathcal{F}\Big) \xrightarrow[n\to+\infty]{\mathcal{D}} \Big(H(f), f \in \mathcal{F}\Big) \ v\ \mathcal{Y}(\mathcal{F}), \tag{51}$$

a $Z : \mathbb{R}^{\mathcal{F}} \to \mathbb{R}$ *je spojitá funkce, pak*

$$Z\Big(\tau_n(\mathsf{P}_n(f) - \mathsf{P}(f)), f \in \mathcal{F}\Big) \xrightarrow[n\to+\infty]{\mathcal{D}} Z\Big(H(f), f \in \mathcal{F}\Big) \ v\ \mathbb{R}. \tag{52}$$

Jako prostor $\mathcal{Y}(\mathcal{F})$ může sloužit $l^{+\infty}[\mathcal{F}]$ prostor všech omezených funkcí na $\mathcal{F}$, prostor spojitých funkcí případně Skorochodův prostor D. Obecnou slabou konvergenci pravděpodobnostních měr a s ní spojenou konvergenci procesů v distribuci je možno nalézt v [2], [5] a některé speciální případy v [3], [4].

Jako množiny $\mathcal{F}$ je možno použít $(\mathbb{I}_{[\bullet<x]}, x \in \mathbb{R}^q)$, což vede na distribuční funkce. Dále je možno použít množinu všech spojitých funkcí, či nějakou jejich podmnožinu. Například polynomy nejvýše stupně 5, atd.

Reference

[1] Doukhan P. (1994). *Mixing: properties and examples.* Lecture Notes in Statistics **85**, Springer-Verlag, Berlin.

[2] Hoffmann-Jørgensen J. (1994). *Probability with a view towards to statistics I,II.* Chapman and Hall, New York.

[3] Jurečková J., Sen P.K. (1996). *Robust statistical procedures.* John Wiley & Sons, Inc., New York.

[4] Štěpán J. (1987). *Teorie pravděpodobnosti.* Academia, Praha.

[5] van der Vaart A.W., Wellner J.A. (1996). *Weak convergence and empirical processes.* Springer, New York.

Poděkování: Příspěvek vznikl za podpory výzkumného záměru MŠMT: MSM 113200008 a grantové agentury České republiky grantu č. 201/03/1027.

Adresa: P. Lachout, KPMS MFF UK, Sokolovská 83, 186 75 Praha 8,
a ÚTIA AV ČR, Pod Vodárenskou věží 4, 182 08 Praha 8

E-mail: `lachout@karlin.mff.cuni.cz`

STATISTICAL ANALYSIS OF GEODETICAL MEASUREMENTS

Jaroslav Marek, Eva Fišerová

Keywords: Two stage regression models, uncertainty of the type A and B, unbiased estimator, confidence domain.

Abstract: Geodetical measurements when geodetical network coordinates are supposed to be stochastic can be modelled by twostage regression model. Two types of estimators of model parameters and their confidence domains are presented in the paper.

Introduction

Staking out points or determining coordinates of given points are typical geodetical problems. In such cases some other points, which coordinates are known, from the government geodetical network are chosen. Then unknown coordinates are calculated on the basis of measured distances and angles between these geodetical points and our determining ones.

The mentioned process of the experiment can be modelled if geodetical network coordinates are supposed to be stochastic — i.e., they are inaccurate — by the twostage linear model. The first stage concerns the inaccuracy in determining of government geodetical network coordinates and these inaccuracy will be called the uncertainty of the type B. The second stage is connected with measurements of distances and angles. The inaccuaracy in the second stage will be called uncertainty of the type A.

The aim of the paper is to compare standard estimators, **H**–optimum estimators in the twostage model and to find their confidence domains.

1 Model of connecting measurements

Let $\mathbf{A}$ be an $m \times n$ matrix. Let $\mathcal{M}(\mathbf{A}) = \{\mathbf{Au} : \mathbf{u} \in \mathbb{R}^n\} \subset \mathbb{R}^m$ denote the column space of the matrix $\mathbf{A}$. Let the symbol $\mathbf{A}^-$ means a g–inverse of the matrix $\mathbf{A}$.

Definition 1.1. The model of connecting measurements will be called random vector $\mathbf{Y} = (\mathbf{Y}_1', \mathbf{Y}_2')'$, with the mean values and the covariance matrix:

$$\begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} \sim \left[\begin{pmatrix} \mathbf{X}_1, & \mathbf{0} \\ \mathbf{D}, & \mathbf{X}_2 \end{pmatrix} \begin{pmatrix} \Theta \\ \beta \end{pmatrix}, \begin{pmatrix} \mathbf{\Sigma}_{1,1}, & \mathbf{0} \\ \mathbf{0}, & \mathbf{\Sigma}_{2,2} \end{pmatrix} \right],$$

where $\mathbf{X}_1, \mathbf{D}, \mathbf{X}_2$ are known $n_1 \times k_1, n_2 \times k_1, n_2 \times k_2$ matrices, such that $\mathcal{M}(\mathbf{D}') \subset \mathcal{M}(\mathbf{X}_1')$; Θ, β are unknown k_1 and k_2-dimensional vectors; $\mathbf{\Sigma}_{1,1}$ and $\mathbf{\Sigma}_{2,2}$ are known covariance matrices of vectors $\mathbf{Y}_1$ and $\mathbf{Y}_2$.

In this model the parameter Θ is estimated on the basis of the vector $\mathbf{Y}_1$ of the first stage and parameter β on the basis of vectors $\mathbf{Y}_2 - \mathbf{D}\hat{\Theta}$ and $\hat{\Theta}$. The results of measurements from the second stage (this means $\mathbf{Y}_2$) we cannot use for the change of the estimator $\hat{\Theta}$.

The parametric space of this model is

$$\underline{\Theta} = \{(\Theta', \beta')' : \mathbf{B}\beta + \mathbf{C}\Theta + \mathbf{a} = \mathbf{0}\},$$

where $\mathbf{B}, \mathbf{C}$ are $q \times k_2, q \times k_1$ matrices, $\mathbf{a}$ is a q-dimensional vector and the rank of the matrix $\mathbf{B}$ is $r(\mathbf{B}) = q < k_2$.

The vector Θ is the parameter of the first stage (connecting), the vector β is the parameter of the second stage (connected). In the second stage we have the unbiased estimator $\widehat{\Theta} = (\mathbf{X}_1'\mathbf{\Sigma}_{1,1}^{-1}\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{\Sigma}_{1,1}^{-1}\mathbf{Y}_1$ from the first stage and its covariance matrix $\mathrm{Var}(\widehat{\Theta}) = (\mathbf{X}_1'\mathbf{\Sigma}_{1,1}^{-1}\mathbf{X}_1)^{-1}$ at our disposal.

The vector $\mathbf{Y}$ will be supposed to be normally distributed.

Definition 1.2. The model in Definition 1.1 in this parametric space $\underline{\Theta}$ is regular provided $r(\mathbf{X}_1) = k_1$, $r(\mathbf{X}_2) = k_2$, $\mathbf{\Sigma}_{1,1}, \mathbf{\Sigma}_{2,2}$ are positively definite matrices and $r(\mathbf{B}) = q$.

Lemma 1.1. *The class $\widetilde{\mathcal{U}}_\beta$ of all linear unbiased estimators $\widetilde{\widetilde{\beta}}$ of the parameter β in the model from Definition 1.1 based on vectors $\mathbf{Y}_2 - \mathbf{D}\hat{\Theta}$ and $\hat{\Theta}$, and satisfaying the (random) condition $\mathbf{B}\widetilde{\widetilde{\beta}} + \mathbf{C}\widehat{\Theta} + \mathbf{a} = \mathbf{0}$ is*

$$\begin{aligned}\widetilde{\mathcal{U}}_\beta = \{ & \left[\mathbf{I} - \mathbf{B}^-\mathbf{B}\right]\left[\mathbf{X}_2^- + \mathbf{W}_1(\mathbf{I} - \mathbf{X}_2\mathbf{X}_2^-) + \mathbf{W}_2\mathbf{B}\mathbf{X}_2^-\right](\mathbf{Y}_2 - \mathbf{D}\widehat{\Theta}) \\ & + \left[-\mathbf{B}^- + (\mathbf{I} - \mathbf{B}^-\mathbf{B})\mathbf{W}_2\right]\mathbf{C}\widehat{\Theta} + (\mathbf{I} - \mathbf{B}^-\mathbf{B})\mathbf{W}_2\mathbf{a} - \mathbf{B}^-\mathbf{a}, \\ & \mathbf{W}_1 \text{ an arbitrary } k_2 \times n_2 \text{ matrix}, \mathbf{W}_2 \text{ an arbitrary } k_2 \times q \text{ matrix}, \\ & \mathbf{X}_2^- \text{ and } \mathbf{B}^- \text{ are arbitrary but fixed } g\text{–inverses} \}.\end{aligned}$$

Proof [1], p. 647. □

Corollary 1.1. *The covariance matrix of the estimator $\widetilde{\widetilde{\beta}}$ is*

$$\begin{aligned}\mathrm{Var}(\widetilde{\widetilde{\beta}}) \;=\; & (\mathbf{I} - \mathbf{B}^-\mathbf{B})\left[\mathbf{X}_2^- + \mathbf{W}_1(\mathbf{I} - \mathbf{X}_2\mathbf{X}_2^-) + \mathbf{W}_2\mathbf{B}\mathbf{X}_2^-\right]\mathbf{\Sigma}_{2,2} \\ & \times \left[\mathbf{X}_2^- + \mathbf{W}_1(\mathbf{I} - \mathbf{X}_2\mathbf{X}_2^-) + \mathbf{W}_2\mathbf{B}\mathbf{X}_2^-\right]'(\mathbf{I} - \mathbf{B}^-\mathbf{B})' \\ & + \{(\mathbf{I} - \mathbf{B}^-\mathbf{B})[-\mathbf{X}_2^-\mathbf{D} - \mathbf{W}_1(\mathbf{I} - \mathbf{X}_2\mathbf{X}_2^-)\mathbf{D} - \mathbf{W}_2\mathbf{B}\mathbf{X}_2^-\mathbf{D} \\ & + \mathbf{W}_2\mathbf{C}] - \mathbf{B}^-\mathbf{C}\}(\mathbf{X}_1'\mathbf{\Sigma}_{1,1}^{-1}\mathbf{X}_1)^{-1}\{(\mathbf{I} - \mathbf{B}^-\mathbf{B}) \\ & \times [-\mathbf{X}_2^-\mathbf{D} - \mathbf{W}_1(\mathbf{I} - \mathbf{X}_2\mathbf{X}_2^-)\mathbf{D} - \mathbf{W}_2\mathbf{B}\mathbf{X}_2^-\mathbf{D} + \mathbf{W}_2\mathbf{C}] - \mathbf{B}^-\mathbf{C}\}'.\end{aligned}$$

Definition 1.3. The least squares estimator of the parameter β obtained under the condition $\mathbf{\Sigma}_{1,1} = \mathbf{0}$ ($\Rightarrow \mathrm{Var}(\widehat{\Theta}) = \mathbf{0}$) is called the standard estimator if in this estimator the vector Θ is substituted by $\widehat{\Theta}$.

Theorem 1.1. *The standard estimator $\hat{\beta}$ of the parameter β is given as*

$$\begin{aligned}\hat{\beta} \;=\; & (\mathbf{X}_2'\mathbf{\Sigma}_{2,2}^{-1}\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{\Sigma}_{2,2}^{-1}(\mathbf{Y}_2 - \mathbf{D}\widehat{\Theta}) \\ & - (\mathbf{X}_2'\mathbf{\Sigma}_{2,2}^{-1}\mathbf{X}_2)^{-1}\mathbf{B}'[\mathbf{B}(\mathbf{X}_2'\mathbf{\Sigma}_{2,2}^{-1}\mathbf{X}_2)^{-1}\mathbf{B}']^{-1} \\ & \times \{\mathbf{a} + \mathbf{C}\widehat{\Theta} + \mathbf{B}(\mathbf{X}_2'\mathbf{\Sigma}_{2,2}^{-1}\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{\Sigma}_{2,2}^{-1}(\mathbf{Y}_2 - \mathbf{D}\widehat{\Theta})\},\end{aligned}$$

whereas this estimator is unbiased, it means $\mathbf{E}(\hat{\beta}) = \beta$.

Proof [3], p. 72. □

Theorem 1.2. *If* $\mathrm{Var}(\widehat{\Theta}) \neq \mathbf{0}$ *then the covariance matrix of the standard estimator* $\hat{\beta}$ *is formed by "uncertainty A" and "uncertainty B":*

$$\mathrm{Var}(\hat{\beta}) = \underbrace{\mathrm{Var}_0(\hat{\beta})}_{\substack{\textit{uncertainty}\\ \textit{type A}}} + \underbrace{\mathbf{A}\mathrm{Var}(\widehat{\Theta})\mathbf{A}'}_{\substack{\textit{uncertainty}\\ \textit{type B}}},$$

$$\begin{aligned}
\textit{where } \mathrm{Var}_0(\hat{\beta}) &= (\mathbf{X}_2'\mathbf{\Sigma}_{2,2}^{-1}\mathbf{X}_2)^{-1} - (\mathbf{X}_2'\mathbf{\Sigma}_{2,2}^{-1}\mathbf{X}_2)^{-1}\mathbf{B}'[\mathbf{B}(\mathbf{X}_2'\mathbf{\Sigma}_{2,2}^{-1}\mathbf{X}_2)^{-1}\mathbf{B}']^{-1} \\
&\quad \times \mathbf{B}(\mathbf{X}_2'\mathbf{\Sigma}_{2,2}^{-1}\mathbf{X}_2)^{-1}, \\
\mathbf{A} &= \langle\{\mathbf{I} - (\mathbf{X}_2'\mathbf{\Sigma}_{2,2}^{-1}\mathbf{X}_2)^{-1}\mathbf{B}'[\mathbf{B}(\mathbf{X}_2'\mathbf{\Sigma}_{2,2}^{-1}\mathbf{X}_2)^{-1}\mathbf{B}']^{-1}\mathbf{B}\} \\
&\quad \times (\mathbf{X}_2'\mathbf{\Sigma}_{2,2}^{-1}\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{\Sigma}_{2,2}^{-1}\mathbf{D} - (\mathbf{X}_2'\mathbf{\Sigma}_{2,2}^{-1}\mathbf{X}_2)^{-1} \\
&\quad \times \mathbf{B}'[\mathbf{B}(\mathbf{X}_2'\mathbf{\Sigma}_{2,2}^{-1}\mathbf{X}_2)^{-1}\mathbf{B}']^{-1}\mathbf{C}\rangle\,.
\end{aligned}$$

Proof [3], p. 74. □

Definition 1.4. Let $\mathbf{H}$ be a given $k_2 \times k_2$ positive semidefinite matrix. The estimator $\widetilde{\widetilde{\beta}}$ from the class $\widetilde{\mathcal{U}}_\beta$ is $\mathbf{H}$-optimal if it minimizes the function

$$\phi(\widetilde{\widetilde{\beta}}) = \mathrm{Tr}[\mathbf{H}\mathrm{Var}(\widetilde{\widetilde{\beta}})], \quad \widetilde{\widetilde{\beta}} \in \widetilde{\mathcal{U}}_\beta\,.$$

Here the symbol $\mathrm{Tr}[\mathbf{H}\mathrm{Var}(\widetilde{\widetilde{\beta}})]$ means the trace of $[\mathbf{H}\mathrm{Var}(\widetilde{\widetilde{\beta}})]$.

Theorem 1.3. *If the estimator* $\widetilde{\widetilde{\beta}}$ *from the class* $\widetilde{\mathcal{U}}_\beta$ *is* $\mathbf{H}$*-optimal, then matrices* $\mathbf{W}_1, \mathbf{W}_2$ *in Lemma 1.1 are solutions of the following equation*

$$\mathbf{U}_1 \begin{pmatrix} \mathbf{W}_1, & \mathbf{W}_2 \end{pmatrix} \begin{pmatrix} \mathbf{V}_1, & \mathbf{T}_1 \\ \mathbf{V}_2, & \mathbf{T}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{P}_1, & \mathbf{P}_2 \end{pmatrix},$$

where

$$\begin{aligned}
\mathbf{U}_1 &= [\mathbf{I} - \mathbf{B}'(\mathbf{B}^-)']\mathbf{H}[\mathbf{I} - \mathbf{B}^-\mathbf{B}], \\
\mathbf{V}_1 &= (\mathbf{I} - \mathbf{X}_2\mathbf{X}_2^-)[\mathbf{\Sigma}_{2,2} + \mathbf{D}(\mathbf{X}_1'\mathbf{\Sigma}_{1,1}^{-1}\mathbf{X}_1)^{-1}\mathbf{D}'](\mathbf{I} - (\mathbf{X}_2^-)'\mathbf{X}_2'), \\
\mathbf{V}_2 &= \mathbf{B}\mathbf{X}_2^-[\mathbf{\Sigma}_{2,2} + \mathbf{D}(\mathbf{X}_1'\mathbf{\Sigma}_{1,1}^{-1}\mathbf{X}_1)^{-1}\mathbf{D}'][\mathbf{I} - (\mathbf{X}_2^-)'\mathbf{X}_2'] \\
&\quad - \mathbf{C}(\mathbf{X}_1'\mathbf{\Sigma}_{1,1}^{-1}\mathbf{X}_1)^{-1}\mathbf{D}'[\mathbf{I} - (\mathbf{X}_2^-)'\mathbf{X}_2'],
\end{aligned}$$

$$
\begin{aligned}
\mathbf{P}_1 &= -[\mathbf{I}-\mathbf{B}'(\mathbf{B}^-)']\mathbf{H}[\mathbf{I}-\mathbf{B}^-\mathbf{B}]\mathbf{X}_2^-[\mathbf{\Sigma}_{2,2}+\mathbf{D}(\mathbf{X}_1'\mathbf{\Sigma}_{1,1}^{-1}\mathbf{X}_1)^{-1}\mathbf{D}' \\
&\quad \times [\mathbf{I}-(\mathbf{X}_2^-)'\mathbf{X}_2'] - [\mathbf{I}-\mathbf{B}'(\mathbf{B}^-)']\mathbf{H}\mathbf{B}^-\mathbf{C}(\mathbf{X}_1'\mathbf{\Sigma}_{1,1}^{-1}\mathbf{X}_1)^{-1}\mathbf{D}'] \\
&\quad \times [\mathbf{I}-(\mathbf{X}_2^-)'\mathbf{X}_2'], \\
\mathbf{T}_1 &= [\mathbf{I}-(\mathbf{X}_2^-)'\mathbf{X}_2'\{[\mathbf{\Sigma}_{2,2}+\mathbf{D}(\mathbf{X}_1'\mathbf{\Sigma}_{1,1}^{-1}\mathbf{X}_1)^{-1}\mathbf{D}'] \\
&\quad \times (\mathbf{X}_2^-)'\mathbf{B}' - \mathbf{D}(\mathbf{X}_1'\mathbf{\Sigma}_{1,1}^{-1}\mathbf{X}_1)^{-1}\mathbf{C}'\}, \\
\mathbf{T}_2 &= \mathbf{B}\mathbf{X}_2^-[\mathbf{\Sigma}_{2,2}+\mathbf{D}(\mathbf{X}_1'\mathbf{\Sigma}_{1,1}^{-1}\mathbf{X}_1)^{-1}\mathbf{D}'](\mathbf{X}_2^-)'\mathbf{B}' + \mathbf{C}(\mathbf{X}_1'\mathbf{\Sigma}_{1,1}^{-1}\mathbf{X}_1)^{-1}\mathbf{C}' \\
&\quad - \mathbf{C}(\mathbf{X}_1'\mathbf{\Sigma}_{1,1}^{-1}\mathbf{X}_1)^{-1}\mathbf{D}'(\mathbf{X}_2^-)'\mathbf{B}' - \mathbf{B}\mathbf{X}_2^-\mathbf{D}(\mathbf{X}_1'\mathbf{\Sigma}_{1,1}^{-1}\mathbf{X}_1)^{-1}\mathbf{C}', \\
\mathbf{P}_2 &= -[\mathbf{I}-\mathbf{B}'(\mathbf{B}^-)']\mathbf{H}[\mathbf{I}-\mathbf{B}^-\mathbf{B}]\mathbf{X}_2^-[\mathbf{\Sigma}_{2,2}+\mathbf{D}(\mathbf{X}_1'\mathbf{\Sigma}_{1,1}^{-1}\mathbf{X}_1)^{-1}\mathbf{D}'](\mathbf{X}_2^-)'\mathbf{B}' \\
&\quad + [\mathbf{I}-\mathbf{B}'(\mathbf{B}^-)']\mathbf{H}\mathbf{B}^-\mathbf{C}(\mathbf{X}_1'\mathbf{\Sigma}_{1,1}^{-1}\mathbf{X}_1)^{-1}[\mathbf{C}'-\mathbf{D}'(\mathbf{X}_2^-)'\mathbf{B}'] \\
&\quad + [\mathbf{I}-\mathbf{B}'(\mathbf{B}^-)']\mathbf{H}[\mathbf{I}-\mathbf{B}^-\mathbf{B}]\mathbf{X}_2^-\mathbf{D}(\mathbf{X}_1'\mathbf{\Sigma}_{1,1}^{-1}\mathbf{X}_1)^{-1}\mathbf{C}'.
\end{aligned}
$$

Proof [1], p. 653. □

2 Confidence domain

Lemma 2.1. *The cross covariance matrix for standard estimator is*

$$
\mathrm{Var}\begin{pmatrix} \widehat{\Theta} \\ \hat{\beta} \end{pmatrix} = \begin{pmatrix} \mathrm{Var}(\widehat{\Theta}), & \mathrm{cov}(\widehat{\Theta},\hat{\beta}) \\ \mathrm{cov}(\hat{\beta},\widehat{\Theta}), & \mathrm{Var}(\hat{\beta}) \end{pmatrix}
$$

where

$$
\begin{aligned}
\mathrm{cov}(\widehat{\Theta},\hat{\beta}) &= \mathrm{Var}(\widehat{\Theta})\{-\mathbf{D}'\mathbf{\Sigma}_{2,2}^{-1}\mathbf{X}_2(\mathbf{X}_2'\mathbf{\Sigma}_{2,2}^{-1}\mathbf{X}_2)^{-1} - \mathbf{C}'[\mathbf{B}(\mathbf{X}_2'\mathbf{\Sigma}_{2,2}^{-1}\mathbf{X}_2)^{-1}\mathbf{B}']^{-1} \\
&\quad \times \mathbf{B}(\mathbf{X}_2'\mathbf{\Sigma}_{2,2}^{-1}\mathbf{X}_2)^{-1} + \mathbf{D}'\mathbf{\Sigma}_{2,2}^{-1}\mathbf{X}_2(\mathbf{X}_2'\mathbf{\Sigma}_{2,2}^{-1}\mathbf{X}_2)^{-1}\mathbf{B}' \\
&\quad \times [\mathbf{B}(\mathbf{X}_2'\mathbf{\Sigma}_{2,2}^{-1}\mathbf{X}_2)^{-1}\mathbf{B}']^{-1}\mathbf{B}(\mathbf{X}_2'\mathbf{\Sigma}_{2,2}^{-1}\mathbf{X}_2)^{-1}\}\,, \\
\mathrm{cov}(\hat{\beta},\widehat{\Theta}) &= \mathrm{cov}(\widehat{\Theta},\hat{\beta})\,.
\end{aligned}
$$

Proof It follows from Theorems 1.1 and from the fact that according to Definition 1.1 vectors $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are uncorrelated. □

Lemma 2.2. *The cross covariance matrix for estimator* $\widetilde{\widetilde{\beta}}$ *is*

$$
\mathrm{Var}\begin{pmatrix} \widehat{\Theta} \\ \widetilde{\widetilde{\beta}} \end{pmatrix} = \begin{pmatrix} \mathrm{Var}(\widehat{\Theta}), & \mathrm{cov}(\widehat{\Theta},\widetilde{\widetilde{\beta}}) \\ \mathrm{cov}(\widetilde{\widetilde{\beta}},\widehat{\Theta}), & \mathrm{Var}(\widetilde{\widetilde{\beta}}) \end{pmatrix}
$$

where

$$
\begin{aligned}
\mathrm{cov}(\widehat{\Theta},\widetilde{\widetilde{\beta}}) &= \mathrm{Var}(\widehat{\Theta})\{\mathbf{C}'[-\mathbf{B}^- + (\mathbf{I}-\mathbf{B}^-\mathbf{B})\mathbf{W}_2]' \\
&\quad -\mathbf{D}'[\mathbf{X}_2^- + \mathbf{W}_1(\mathbf{I}-\mathbf{X}_2\mathbf{X}_2^-) + \mathbf{W}_2\mathbf{B}\mathbf{X}_2^-]'(\mathbf{I}-\mathbf{B}^-\mathbf{B})'\}\,, \\
\mathrm{cov}(\widetilde{\widetilde{\beta}},\widehat{\Theta}) &= \mathrm{cov}(\widehat{\Theta},\widetilde{\widetilde{\beta}})\,.
\end{aligned}
$$

Proof It follows Definition 1.1 and Lemma 1.1. □

Theorem 2.1. *Let* $\mathbf{h}(\Theta,\beta)$ *be* s*–dimensional differentiable function. Then the* $(1-\alpha)$*–confidence ellipsoid for the vector function* $\mathbf{h}(\Theta,\beta)$*, based on the standard estimator* $\mathbf{h}(\widehat{\Theta},\hat{\beta})$*, is the set*

$$\mathcal{E}_{1-\alpha}\left(\mathbf{h}(\Theta,\beta)\right)=\Big\{\mathbf{u}:\mathbf{u}\in\mathbb{R}^s\,,\\ (\mathbf{u}-\mathbf{h}(\widehat{\Theta},\hat{\beta}))'\mathbf{W}_S^{-1}(\mathbf{u}-\mathbf{h}(\widehat{\Theta},\hat{\beta}))\le\chi_s^2(1-\alpha)\Big\}\,,$$

where

$$\mathbf{W}_S=\left[\left(\frac{\partial\mathbf{h}(\Theta,\beta)}{\partial\Theta'},\frac{\partial\mathbf{h}(\Theta,\beta)}{\partial\beta'}\right)\mathrm{Var}\begin{pmatrix}\widehat{\Theta}\\ \hat{\beta}\end{pmatrix}\left(\frac{\partial\mathbf{h}(\Theta,\beta)}{\partial\Theta'},\frac{\partial\mathbf{h}(\Theta,\beta)}{\partial\beta'}\right)'\right].$$

Here the symbol $\chi_s^2(1-\alpha)$ *denotes* $(1-\alpha)$*-quantile of* χ^2*-distribution with* s *degrees of freedom.*

Proof Using the Taylor expansion, the linear approximation of the estimator $\mathbf{h}(\widehat{\Theta},\hat{\beta})$ is normally distributed and it holds $\mathbf{h}(\widehat{\Theta},\hat{\beta})\sim N_s[\mathbf{h}(\Theta,\beta);\mathbf{W}_S]$, the rest of the proof is an obvious consequence of Pearson lemma (cf. [2], p. 87). □

Remark 2.1. *Analogously we can determine the* $(1-\alpha)$*–confidence ellipsoid for the function* $\mathbf{h}(\widehat{\Theta},\widetilde{\widetilde{\beta}})$*, based on the* $\mathbf{H}$*–optimum estimator* $\mathbf{h}(\widehat{\Theta},\widetilde{\widetilde{\beta}}))$*. It is the set*

$$\tilde{\mathcal{E}}_{1-\alpha}\left(\mathbf{h}(\Theta,\beta)\right)=\Big\{\mathbf{u}:\mathbf{u}\in\mathbb{R}^s\,,\\ (\mathbf{u}-\mathbf{h}(\widehat{\Theta},\widetilde{\widetilde{\beta}}))'\mathbf{W}_H^{-1}(\mathbf{u}-\mathbf{h}(\widehat{\Theta},\widetilde{\widetilde{\beta}}))\le\chi_s^2(1-\alpha)\Big\}\,,$$

where

$$\mathbf{W}_H=\left[\left(\frac{\partial\mathbf{h}(\Theta,\beta)}{\partial\Theta'},\frac{\partial\mathbf{h}(\Theta,\beta)}{\partial\beta'}\right)\mathrm{Var}\begin{pmatrix}\widehat{\Theta}\\ \widetilde{\widetilde{\beta}}\end{pmatrix}\left(\frac{\partial\mathbf{h}(\Theta,\beta)}{\partial\Theta'},\frac{\partial\mathbf{h}(\Theta,\beta)}{\partial\beta'}\right)'\right].$$

Remark 2.2. *When confidence domains are calculated, functions* $\frac{\partial\mathbf{h}(\Theta,\beta)}{\partial\Theta'}$ *and* $\frac{\partial\mathbf{h}(\Theta,\beta)}{\partial\beta'}$ *are substituted by their estimators, i.e.,* $\frac{\partial\mathbf{h}(\widehat{\Theta},\hat{\beta})}{\partial\Theta'}$, $\frac{\partial\mathbf{h}(\widehat{\Theta}\widetilde{\widetilde{\beta}})}{\partial\Theta'}$, *etc.*

3 Numerical example

Example 3.1. The problem is to estimate plane coordinates of points P_1, P_2 and P_3 in a cartesian coordinates from the Figure 1. We have estimated values $\widehat{\Theta}_{2i-1},\widehat{\Theta}_{2i}$ of coordinates of the point $A_i=(\Theta_{2i-1},\Theta_{2i})$, $i=1,2$, and measured values $\mathbf{Y}_j$ and $\mathbf{Y}_k$ of lengths β_1, $j=1,2,3,4$, and angles β_k, $k=5,6,7$, respectively, at our disposal.

The aim of this example is to estimate coordiantes of the point P_3.

Let results from the first and the second stage of measurement be ($\widehat{\Theta}_1$, $\widehat{\Theta}_2$, $\widehat{\Theta}_3$, $\widehat{\Theta}_4$) = (500.00, 500.00, 879.60, 664.65) and (Y_1, Y_2, Y_3, Y_4, Y_5, Y_6, Y_7) = (134.04, 118.40, 116.17, 71.49, 164.50°, 156.50°, 165.00°).

The values $\widehat{\Theta}_i$, Y_i, $i = 1, 2, 3, 4$, are in meters, values Y_j, $j = 5, 6, 7$, in degrees.

The accuracy of measurements are given by the covariance matrices.

$P_1 = (\Theta_1 + \beta_1, \Theta_2)$

$P_2 = (\Theta_1 + \beta_1 - \beta_2 \cos\beta_5, \Theta_2 + \beta_2 \sin\beta_5)$

$P_3 = (\Theta_1 + \beta_1 - \beta_2 \cos\beta_5 + \beta_3 \cos(\beta_5 + \beta_6),$
$\Theta_2 + \beta_2 \sin\beta_5 - \beta_3 \sin(\beta_5 + \beta_6)) = (h_1(\Theta, \beta), h_2(\Theta, \beta))$

$P_4 = (\Theta_1 + \beta_1 - \beta_2 \cos\beta_5 + \beta_3 \cos(\beta_5 + \beta_6) - \beta_4 \cos(\beta_5 + \beta_6 + \beta_7),$
$\Theta_2 + \beta_2 \sin\beta_5 - \beta_3 \sin(\beta_5 + \beta_6) + \beta_4 \sin(\beta_5 + \beta_6 + \beta_7))$

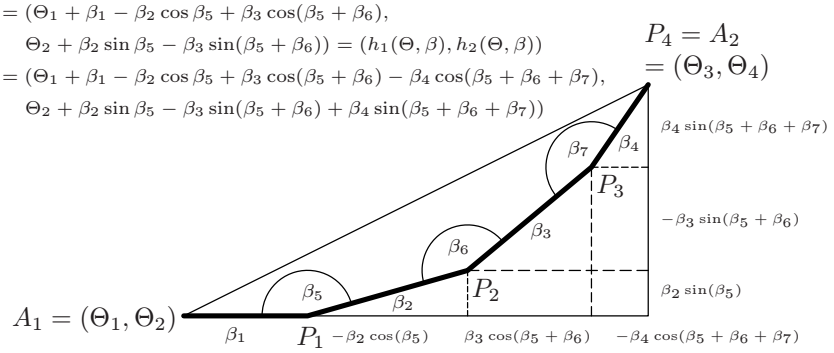


Figure 1: Polygonometric measurement.

In our case we will consider

$$\mathbf{\Sigma}_{1,1} = \begin{pmatrix} 0.0100, & -0.0049, & 0.0000, & 0.0000 \\ -0.0049, & 0.0064, & 0.0000, & 0.0000 \\ 0.0000, & 0.0000, & 0.0144, & 0.0025 \\ 0.0000, & 0.0000, & 0.0025, & 0.0016 \end{pmatrix}.$$

We suppose that in the second stage of measurement the angles were measured with dispersion of $\sigma_\omega^2 = (2'')^2$, and the distances were measured with dispersion of $\sigma_d^2 = 2^2$ cm^2.

One can observe in Figure 1, the condition $g(\beta, \Theta) = 0$ is implied for parameters Θ and β, where

$$\begin{aligned} g(\beta, \Theta) &= (\Theta_3 - \Theta_1)^2 + (\Theta_4 - \Theta_2)^2 - (\beta_1^2 - 2\beta_1\beta_2 \cos(\beta_4) + \beta_2^2 \\ &\quad + 2\beta_1\beta_3 \cos(\beta_4 + \beta_5) - 2\beta_2\beta_3 \cos(\beta_4)\cos(\beta_4 + \beta_5) + \\ &\quad + \beta_3^2 - 2\beta_2\beta_3 \sin(\beta_4)\sin(\beta_4 + \beta_5)), \\ x &= \beta_1 - \beta_2 \cos(\beta_5) + \beta_3 \cos(\beta_5 + \beta_6) - \beta_4 \cos(\beta_5 + \beta_6 + \beta_7), \\ y &= \beta_2 \sin(\beta_5) - \beta_3 \sin(\beta_5 + \beta_6) + \beta_4 \sin(\beta_5 + \beta_6 + \beta_7). \end{aligned}$$

The linear version of the condition $g(\beta, \Theta) = 0$, obtained by the using the Taylor expansion at the approximate point (β^0, Θ^0)=(Y_1,Y_2,Y_3,Y_4,Y_5,Y_6,Y_7, $\widehat{\Theta}_1$, $\widehat{\Theta}_2$, $\widehat{\Theta}_3$, $\widehat{\Theta}_4$) is in the form

$\mathbf{B}\delta\beta + \mathbf{C}\delta\Theta + a = 0$, where $\delta\beta = \beta - \beta^0$, $\delta\Theta = \Theta - \Theta^0$, $\mathbf{B} = \frac{\partial g(\beta^0, \Theta^0)}{\partial \beta'}$, $\mathbf{C} = \frac{\partial g(\beta^0, \Theta^0)}{\partial \Theta'}$, $a = g(\beta^0, \Theta^0)$.

In our linearized model, where $\mathbf{X}_1 = \mathbf{I}_{4\times 4}$, $\mathbf{D} = \mathbf{0}_{7\times 4}$ and $\mathbf{X}_2 = \mathbf{I}_{7\times 7}$, the standard estimator $\hat{\beta}$ (cf. Theorem 1.1, 1.2) and $\mathbf{H}$-optimum estimator $\widetilde{\widetilde{\beta}}$ (cf. Lemmas 1.1, Theorem 1.3), where $\mathbf{H}$ is in the form

$\mathbf{H} = (\frac{\partial h_1}{\beta'} + \frac{\partial h_2}{\beta'})(\frac{\partial h_1}{\beta'} + \frac{\partial h_2}{\beta'})'$, where $h_1 = \Theta_1 + \beta_1 - \beta_2 \cos\beta_5 + \beta_3 \cos(\beta_5 + \beta_6)$ and $h_2 = \Theta_2 + \beta_2 \sin\beta_5 - \beta_3 \sin(\beta_5 + \beta_6)$ (see figure 1), are

$\hat{\beta}$ = (134.02982 m, 118.38900 m, 116.15934 m, 71.48050 m, 163.99999°, 156.49999°, 164.99999°)′,

$\widetilde{\widetilde{\beta}}$ = (134.04002 m, 118.40002 m, 116.17001 m, 71.48998 m, 164.00231°, 156.49518°, 164.94551°)′.

By chosen matrix $\mathbf{H}$ minimizing data errors in the process estimation of the vector $\widetilde{\widetilde{\beta}}$ we got better estimator of the parameter β in comparison with the standard estimator $\hat{\beta}$. It follows from the fact that for the chosen matrix $\mathbf{H}$ is $\mathrm{Tr}(\mathbf{H}\mathrm{Var}(\widetilde{\widetilde{\beta}})) = 0.0011 < 0.0129 = \mathrm{Tr}(\mathbf{H}\mathrm{Var}(\hat{\beta}))$.

Hence, estimated coordinates of the point P_3 and their covariance matrices are

$$\hat{P}_3 = \begin{pmatrix} 837.464 \\ 606.519 \end{pmatrix}, \quad \mathrm{Var}(\hat{P}_3) = \begin{pmatrix} 0.010672 & -0.014312 \\ -0.014312 & 0.008770 \end{pmatrix},$$

$$\widetilde{\widetilde{P}}_3 = \begin{pmatrix} 837.491 \\ 606.528 \end{pmatrix}, \quad \mathrm{Var}(\widetilde{\widetilde{P}}_3) = \begin{pmatrix} 0.010254 & -0.004641 \\ -0.004641 & 0.006655 \end{pmatrix}.$$

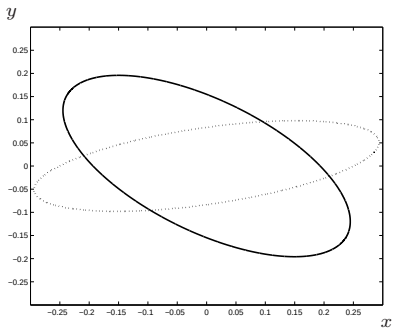


Figure 2: The 0.95–confidence domains for points A_1 and A_2; (solid line – point A_1, dot line – point A_2).

Now we will draw the 0.95–confidence domains for points A_1, A_2, P_1, P_2 and P_3. For better graphical comparison, the center of each domain is moved to the point $(0, 0)$.

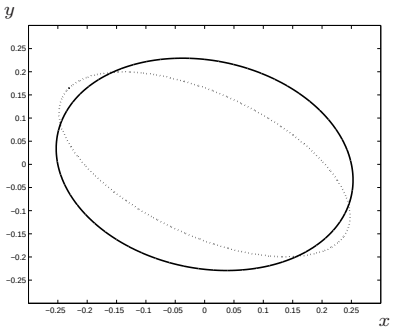


Figure 3: The 0.95–confidence domains for the point P_3; (solid line – standard estimator, dot line – **H**-optimum estimator for the point P_3).

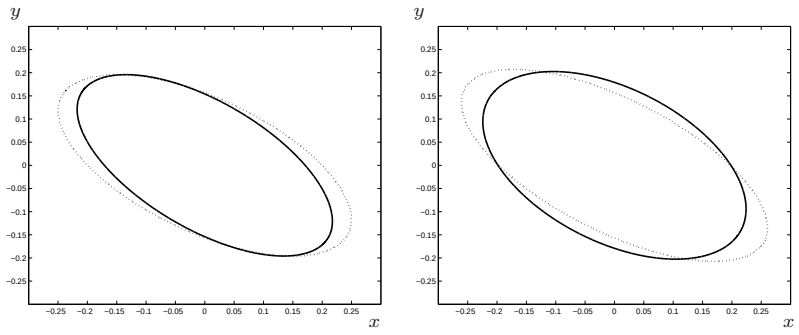


Figure 4: The 0.95–confidence domains for points P_1 and P_2; (solid line – standard estimator, dot line – **H**-optimum (for the point P_3) estimator).

References

[1] Kubáček L. (1993). *Two stage regression models with constraints.* Math. Slovaca **43**, 643 – 658.

[2] Kubáček L., Kubáčková L. (2000). *Statistics and metrology (in Czech).* Olomouc. Publishing House of Palacký University.

[3] Marek J. (2003). *Estimation in connecting measurements.* Acta Universitas Palackianae, Fac. rer. nat., Mathematica **42**, 69 – 86.

[4] Rao C.R., Mitra S.K. (1971). *Generalized inverse of matrices and its applications.* John Wiley & sons, New York–London–Sydney–Toronto.

Address: J. Marek, E. Fišerová, Department of Mathematical Analysis and Applied Mathematics, Faculty of Science, Palacký University, Tomkova 40, 779 00 Olomouc, Czech Republic

E-mail: `marek@inf.upol.cz, fiserova@inf.upol.cz`

OPTIMÁLNÍ SEGMENTACE DAT

Petr Novotný

Klíčová slova: Výpočetní statistika, po částech spojitá regrese.

Abstrakt: Snížení paměťové náročnosti při výpočtu po částech spojitého regresního modelu.

Motivace

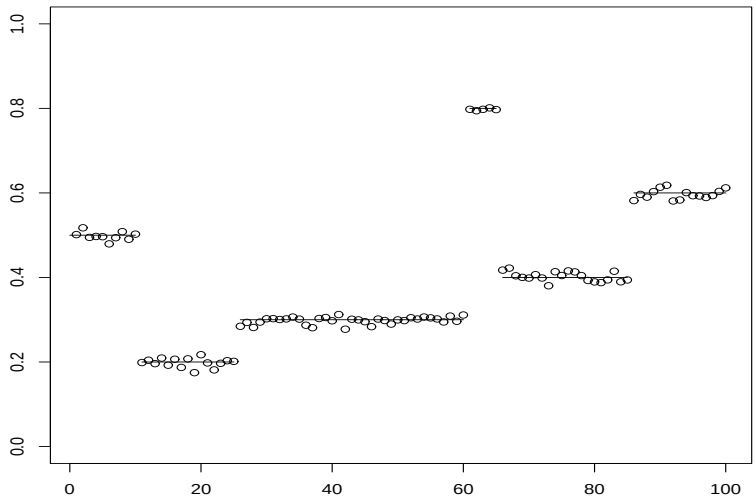
Úkolem je rozdělit sekvenci DNA na úseky podle obsažené informace (stavba těla, trávení apod.).

Biologické řešení

DNA se skládá ze 4 druhů bazí (adenosin (A), cytosin (C), guanin (G) a thymin (T)). Jeden z možných přístupů je rozdělit DNA na úseky podle poměru dvojic bazí.

Matematické řešení

Použijeme po částech konstantní regresní model a budeme minimalizovat reziduální součet čtverců. Například při hledání poměru A a C ku G a T budeme kódovat A a C jako 1 a zbylé jako 0. Počet bodů nespojitosti budeme navíc penalizovat vhodnou funkcí.



Americké výsledky

Po spolupráci s NSA (National Security Agency) a jejich superpočítači se americkým vědcům podařilo rozdělit DNA bakteriofágu λ. DNA tohoto bakteriofágu se skládá z 48 502 bazí. Já jsem schopen dosáhnout stejného výsledku s použitím současného PC během několika málo hodin.

Formalizace zadání

Máme dáno:

- Vektor pozorování

$$X_1, \ldots, X_N$$

- Ztrátovou funkci

$$\mathbf{R}(i,j) \equiv \mathbf{RSS}(X_i, \ldots, X_j) \quad 1 \leq i \leq j \leq N$$

Hledáme:

- Optimální dělení vektoru pozorování na právě K úseků.

  Tj. hledáme $J_2, \ldots, J_K : \quad 1 = J_1 < J_2 < \ldots < J_K \leq N$, která minimalizují:

$$\sum_{i=1}^{K-1} \mathbf{R}(J_i, J_{i+1} - 1) + \mathbf{R}(J_K, N)$$

  J_i jsou počáteční body úseků optimálního dělení.

Matice

- Horní trojúhelníková matice $\mathbf{R}$ typu $N \times N$:

$$\mathbf{R} = \begin{pmatrix} \mathbf{R}(1,1) & \ldots & \mathbf{R}(1,N) \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \mathbf{R}(N,N) \end{pmatrix}$$

  Tato obrovská matice obsahuje reziduální součty čtverců pro všechny možné úseky.

- Matice $\mathbf{Q}$ typu $K \times N$:

$$\mathbf{Q} = \begin{pmatrix} q(1,1) & \ldots & q(1,K) & \ldots & q(1,N) \\ \vdots & \ddots & \vdots & & \vdots \\ 0 & \ldots & q(K,K) & \ldots & q(K,N) \end{pmatrix}$$

  Prvek $q(i,j)$ vyjadřuje ztrátu odpovídající optimálnímu rozdělení podvektoru $X_1, \ldots, X_j$ na i úseků.

- Matice $\mathbf{P}$ typu $K \times N$:

$$\mathbf{P} = \begin{pmatrix} p(1,1) & \dots & p(1,K) & \dots & p(1,N) \\ \vdots & \ddots & \vdots & & \vdots \\ 0 & \dots & p(K,K) & \dots & p(K,N) \end{pmatrix}$$

  Prvek $p(i,j)$ obsahuje odkaz na předchozí dělení:

  $p(i,j) = k \Rightarrow q(i,j) = q(i-1,k) + \mathbf{R}(k+1,j)$.

Klasické řešení

Klasický algoritmus založený na dynamickém programování řeší naši úlohu ve dvou krocích.

1. Vypočteme si celou matici $\mathbf{R}$.
2. Ve druhém kroku se postupně vytváří matice Q a P.
   - Pro první řádek:
     $q(1,i) = \mathbf{R}(1,i)$
     $p(1,i) = 0$
   - Pro ostatní řádky:
     $q(i,j) = \min_{k<j}(q(i-1,k) + \mathbf{R}(k+1,j))$
     $p(i,j) = \operatorname{argmin}_{k<j}(q(i-1,k) + \mathbf{R}(k+1,j))$

Problémy klasického řešení

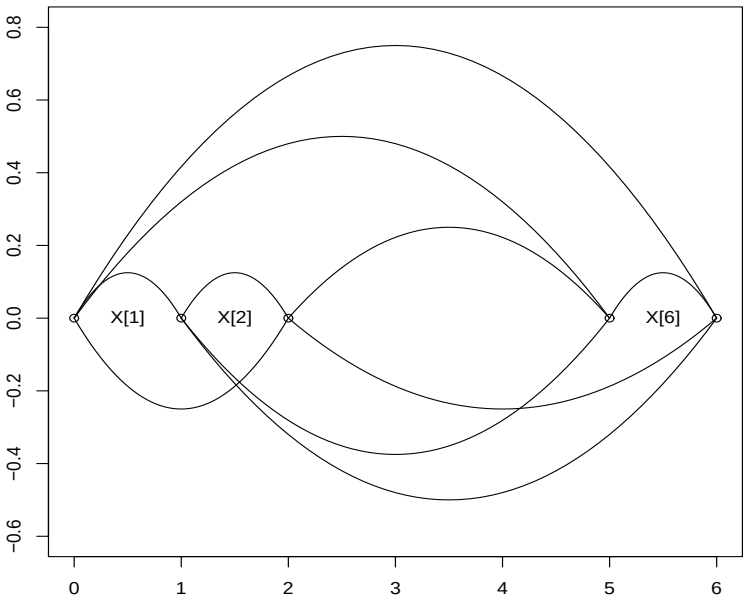
Největším nedostatkem klasického postupu je jeho paměťová náročnost. Matice $\mathbf{R}$ se nám brzy nevejde do operační paměti a její část se musí uložit na pevný disk a právě díky pomalému přístupu na pevný disk je výpočet neproveditelný.

Operační paměť o velikosti 512 MB je spotřebována již pro vektor délky $N = 9\,000$. Abychom zvýšili možné N na dvojnásobek potřebujeme čtyřikrát větší operační paměť.

Mým cílem bylo tento nedostatek odstranit.

Alternativní model situace

Situaci si můžeme představit jako orientovaný graf na množině $\{0, \dots, N\}$, kde z vrcholu i vede hrana do vrcholu j právě tehdy, když $i < j$. Naše data jsou mezery mezi vrcholy grafu. Cílem je najít nejkratší cestu z 0 do N, která má právě K hran. V tomto algoritmu si u každého vrcholu grafu pamatuji cenu nejkratší cesty z 0 do tohoto vrcholu, která má právě 1 až K hran.



Popis navrženého algoritmu

Výstavba matic Q a P

- V prvním kroku spočítám první řádek matice $\mathbf{R}$.

  $q(1,i) = \mathbf{R}(1,i)$
  $p(1,i) = 0$

- V i-tém kroku napočtu i-tý řádek matice $\mathbf{R}$.

- Pokud je $i \leq K$, přičtu ke všem hodnotám $\sum_{h=1}^{i-1} \mathbf{R}(h,h)$ a výsledek uložím do i-tého řádku matice Q od i-té souřadnice a i-tý řádek matice P inicializuji na $i-1$.

  Tyto hodnoty mi reprezentují dělení $1, 2, \ldots i$.

- Začnu přepočítávat hodnoty v dosud obsazených řádcích kromě prvního a bude se jednat o složky s vyšším indexem než i. Pokusím se zlepšit optimální dělení za předpokladu, že počáteční bod posledního intervalu je i.

Schematický zápis algoritmu:

Pro l od 2 do $\min(i, K)$ a pro $j \geq i$ dělej:

$$q(l, j) = \min(q(l, j), q(l-1, i-1) + \mathbf{R}(i, j))$$

Pokud došlo ke změně optimální hodnoty pak:

$$p(l, j) = i - 1$$

Tento způsob umožňuje použití paralelního přepočítávání jednotlivých řádků matic P a Q.

Výsledky simulací

Všechny simulace proběhly na počítači Pentium IV 2.8 GHz, 512 MB RAM. Vlastní algoritmus byl naprogramován v Matlabu, Release 13. Výsledky v tabulce jsou uvedeny v sekundách.

$K \setminus N$	10 000	20 000	50 000	100 000	200 000
5	0:00:18	0:01:09	0:07:19	0:30:57	2:03:20
10	0:00:37	0:02:36	0:16:08	1:06:19	4:27:15

Simulace výpočtu našeho motivačního příkladu ($N = 48\,502$ a $K = 40$) trvala 3 hodiny, 27 minut a 41 sekund. Optimální segmentace vektoru o délce 1 000 0000 na 5 segmentů by trvala asi 2 dny.

Zrychlení výpočtu reziduí

Při výpočtu reziduí používám zrychlený postup, který umožňuje vypočítat $\mathbf{R}(i, j)$ z hodnot $\sum_{k=i}^{j} X_k$ a $\sum_{k=i}^{j} X_k^2$.

$$\sum_{k=i}^{j}(X_k - \frac{1}{j-i+1}\sum_{l=i}^{j} X_l)^2 = \sum_{k=i}^{j} X_k^2 - \frac{1}{j-i+1}(\sum_{k=i}^{j} X_k)^2$$

Tímto způsobem snížím časovou náročnost o jeden řád. Postupně načítám první a druhé mocniny pozorování a z nich snadno spočítám odpovídající reziduální součet čtverců.

V našich datech jsou si navíc první a druhé mocniny rovny, protože stále sčítám buď 1 nebo 0, což velmi zrychlí výpočet reziduí.

Paměťová a časová náročnost

Zlepšený algoritmus sníží celkovou paměťovou náročnost výpočtu z $\mathrm{O}(N^2)$ na $((2 \times K + 1) \times N)$. Časová náročnost u tohoto algoritmu je $\mathrm{O}(N^2 \times K)$ oproti $\mathrm{O}(N^3 \times K)$ u klasického postupu.

Závěr

Navržený algoritmus umožňuje nejen podstatné zrychlení našeho výpočtu, ale zejména významné zvýšení maximální délky dělených dat. Obrovskou výhodou je možnost použití vektorového paralelního programování, která odsouvá počet dělení do pozadí našeho zájmu.

Další zlepšení při hledání optimálního dělení již zřejmě není možné, zbývá jen problematická heuristika.

Reference

[1] Braun J.V., Braun R.K., Müller H.-G. (2000). *Multiple changepoint fitting via quasilikelihood, with application to DNA seqvence segmentation.* Biometrica **87**, 2, 301 – 314.

Poděkování: Děkuji prof. Jaromíru Antochovi CSc. za všestrannou podporu a neocenitelné rady k obsahu i formě.

Adresa: P. Novotný, KPMS, Sokolovská 83,186 75 Praha 8

E-mail: `reter@centrum.cz`

THE TEST OF FULL SPECIFICATION OF THE NORMAL DISTRIBUTION

Marek Omelka

Keywords: Testing statistical hypothesis, Bahadur efficiency.

Abstract: For a random sample $X_1, \ldots, X_n$ from the normal distribution $\mathcal{N}(\mu, \sigma^2)$ we test the hypothesis $H\colon (\mu, \sigma^2)^\mathsf{T} = (0, 1)^\mathsf{T}$ against $K\colon (\mu, \sigma^2)^\mathsf{T} \neq (0, 1)^\mathsf{T}$. Next to three classical tests – Likelihood ratio test, Wald test and Rao score test, we consider also special tests which were suggested for this problem in the literature. Within these tests we will be especially interested in the test which locally maximizes the average power, Isaacson's approximation of the type D test and test based on Fisher's combination of independent statistics. We investigate both local and global properties of these tests. The Monte Carlo study confirms the importance of the condition of local unbiasedness of the test. We also show the Bahadur efficiency to be a good concept of the global performance of the test in this problem.

1 Introduction

Let $X_1, \ldots, X_n$ be a sample from the normal distribution $\mathcal{N}(\mu, \sigma^2)$, where both parameters are unknown and we want to test the null hypothesis that the parameters (μ, σ^2) equal to some prescribed values (μ_0, σ_0^2). Without loss of generality we can suppose $\mu_0 = 0$ and $\sigma_0^2 = 1$. Our problem is a little atypical, because in applications we are usually interested only in the location parameter and the scale parameter is nuisance. But we hope that this problem can be also of some practical interest. Although several tests have been proposed in the literature, no comparison of these tests leading to some practical recommendations is known to the author. And this is just the aim of this paper, which does not present any new theoretical results but may be of interest for people who would like to test the hypothesis of the full specification of the normal distribution in practice.

In Section 2 we explain some definitions useful for our problem and the Section 3 contains the results of our Monte Carlo study. And in the appendix we sketch the computation of the Bahadur slopes of the proposed tests.

2 Basic concepts and definitions

Consider a random sample $\mathbf{X} = (X_1, \ldots, X_n)$ from a distribution with density $f(x, \theta)$ which depends on the unknown vector parameter θ. Let H be a null hypothesis about this parameter and Φ be a test function defined on the sample space which gives the probability of rejecting H when the sample $\mathbf{X} = \mathbf{x}$ is observed. Denote $\beta_\Phi(\theta) = \mathbf{E}_\theta \Phi(\mathbf{X})$ the power function of this test.

A natural condition imposed on tests of the simple hypothesis $H\colon \theta = \theta_0$ against two-sided alternatives $K\colon \theta \neq \theta_0$ is that of the local unbiasedness:

Definition 2.1. *A level α test Φ is said to be locally unbiased, if there exists $\Delta > 0$ such that $\beta_\Phi(\theta) \geq \alpha$ for all θ with $0 < d(\theta) < \Delta$.*

Suppose that the power function is twice continuously differentiable, which is always true when the underlying density $f(x, \theta)$ belongs to the exponential family of distributions) and denote the first and second derivative of the power function by $\dot{\beta}_\Phi(\theta)$ and $\ddot{\beta}_\Phi(\theta)$ respectively. Then we can define two tests which are in some sense locally optimal. The first one was proposed by Isaacson [2] and is called **type D test**. This test maximizes the determinant of the matrix $\{\ddot{\boldsymbol{\beta}}_\Phi(\theta_0)\}$ subject to the conditions of size and unbiasedness. The disadvantage of this test is that it is very difficult to construct. To overcome this inconvenience, Gupta and Vermaire [1] came up with the test which is also locally unbiased but which maximizes the trace of the matrix $\{\ddot{\boldsymbol{\beta}}_\Phi(\theta_0)\}$. We shall refer to this test as the LMMPU (Locally Most Mean Powerful Unbiased) test.

Remark. We can easily see that $\dot{\beta}_\Phi(\theta_0) = 0$ is the necessary condition for the test Φ of the hypothesis $H\colon \theta = \theta_0$ against two-sided alternative to be locally unbiased.

3 Monte Carlo study

Let $X_1, \ldots, X_n$ be a sample from the normal distribution with the density $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$, where both parameters μ and σ are unknown. Set $\theta = (\mu, \sigma^2)^\mathsf{T}$ and consider testing $H\colon \theta = (0,1)^\mathsf{T}$ against $K\colon \theta \neq (0,1)^\mathsf{T}$. Let $\hat{\boldsymbol{\theta}} = (\bar{X}, S^2)^\mathsf{T}$ be the maximum likelihood estimate of the parameter θ, where $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and $S^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$. The Fisher score function is

$$\mathbf{l}(\mathbf{X}, \theta) = (\partial/\partial\theta) \log L(\mathbf{X}, \theta) = \left(\sum \frac{X_i - \mu}{\sigma^2},\ -\frac{n}{2\sigma^2} + \sum \frac{(X_i - \mu)^2}{2\sigma^4} \right)^\mathsf{T}$$

and the Fisher information matrix is

$$\mathbf{J}_n(\theta) = \mathbf{E}_\theta \mathbf{l}(\mathbf{X}, \theta)\mathbf{l}(\mathbf{X}, \theta)^\mathsf{T} = \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}. \tag{1}$$

We shall consider the following tests:

1. *Likelihood ratio (LR) test* with the test statistic
   $LR = 2\{\log L(\mathbf{X}, \hat{\boldsymbol{\theta}}) - \log L(\mathbf{X}, \theta_0)\} = n\bar{X}^2 + n\,(S^2 - 1 - \log(S^2))$.
2. *Wald (W) test* with the test statistic

   $$WT = (\hat{\theta} - \theta_0)^\mathsf{T} \mathbf{J}_n(\hat{\theta})(\hat{\theta} - \theta_0) = \frac{n\bar{X}^2}{S^2} + \frac{n\,(S^2 - 1)^2}{2\,S^4}.$$

   The matrix $\mathbf{J}_n(\hat{\theta})$ is sometimes replaced with $\mathbf{J}_n(\theta_0)$. But this would lead to a test which would be almost identical with the approximate D type test introduced later.

3. *Rao score (RS) test* with the test statistic
$RST = \mathbf{l}(\mathbf{X}, \theta_0)^{\mathsf{T}} \mathbf{J}_n(\theta_0)^{-1} \mathbf{l}(\mathbf{X}, \theta_0) = n\,\bar{X}^2 + \frac{n}{2}\left(\bar{X}^2 + S^2 - 1\right)^2$.

4. *Approximate type D (AD) test* – with the test statistic

$$AD = n\bar{X}^2 + \frac{n-1}{2}\left(\frac{nS^2}{n-1} - 1\right)^2.$$

This test was proposed Isaacson [2], as he was not able to construct the type D test.

5. *LMMPU test* ([1]) with the critical region $\left(\bar{X}^2 + S^2 - C\right)^2 + 4\bar{X}^2 \geq K^2$, where constants C, K are determined subject to the conditions of size and unbiasedness.

6. *Fisher test*: Let Φ stand for a distribution function of a standard normal variable and G_p for a distribution function of a variable with a χ^2-distribution with p degrees of freedom. The Fisher test is based on the statistic

$$Fisher = -2\,\log\left\{2\left[1 - \Phi(|\,\sqrt{n}\bar{X}_n\,|)\right]\right\} - 2\,\log\left\{1 - G_2[-2\log(H_n)]\right\},$$

where $H_n = 2\,G_{n-1}(nS^2)$ if $nS^2 \leq \tilde{G}_{n-1}$ or $H_n = 2\left[1 - G_{n-1}(nS^2)\right]$ otherwise, and $\tilde{G}_{n-1}$ stands for the median of the distribution G_{n-1}.

This construction is known as Fisher's method of combining independent test statistics. Under the null hypothesis the statistic $Fisher$ has χ^2-distribution with 4 degrees of freedom. The test is a one sample analogue of the test of Littel and Folks [3] who were dealing with the two sample problem. Analogously as in [3], it can be shown that our test is optimal in the sense of the Bahadur efficiency.

It is worth noting that all these tests try to combine the information from $\bar{X}$ and S^2. It is not so surprising as the pair $(\bar{X}, S^2)$ forms the minimal sufficient statistics for the normal distribution when both parameters are unknown.

We prescribe the size $\alpha = 0.05$ and fix the sample size $n = 20$. It is well known that under the null hypothesis the statistics LR, WT, RS and also AD have asymptotically the χ^2-distribution with 2 degrees of freedom. In practice we mostly approximate the critical values of these tests by the asymptotic ones. In [4] is shown that this approximation works fine with the exception of the W test. But to make the comparison of the tests fair, in the sequel the estimates of the true critical values are used ensuring that all the test have approximately the size 0.05.

At first, let us now look at local properties of the proposed tests. For convenience we will denote the first and second derivatives of power functions

Test	$\dot{\boldsymbol{\beta}}_2$	$\ddot{\boldsymbol{\beta}}_{11}$	$\ddot{\boldsymbol{\beta}}_{22}$	$\det\{\ddot{\boldsymbol{\beta}}_\Phi\}$	$\mathrm{tr}\{\ddot{\boldsymbol{\beta}}_\Phi\}$
LR	0	2.78	1.59	4.44	4.38
W	-0.23	0.87	1.31	1.15	2.19
RS	0.30	3.07	1.39	4.26	4.46
AD	0.27	2.87	**1.67**	**4.81**	4.55
LMMPU	0	**3.71**	0.97	3.61	**4.69**
Fisher	0	2.98	0.05	0.15	3.03

Table 1: The derivatives of the power functions of the tests at θ_0 sample size $n = 20$.

of tests at θ_0 as

$$\dot{\beta}_i = \left.\frac{\partial \beta_\Phi(\theta)}{\partial \theta_i}\right|_{\theta=\theta_0} \quad \text{and} \quad \ddot{\beta}_{ij} = \left.\frac{\partial^2 \beta_\Phi(\theta)}{\partial \theta_i \partial \theta_j}\right|_{\theta=\theta_0}.$$

These derivatives can be easily estimated by the means of Monte Carlo simulation ([4]) and are given in Table 1. The table does not include the values of derivatives $\dot{\beta}_1$, $\ddot{\beta}_{12}(=\ddot{\beta}_{21})$, since their values are zero for each of considered tests.

Firstly, we should note that the W, RS, and AD test are not locally unbiased. In agreement with the asymptotic results of Peers [5], the W test is more powerful when $\sigma^2 < 1$, and the RS test is more powerful when $\sigma^2 > 1$. This fact is also in a good agreement with the relative Bahadur efficieny of these two tests. We can also see that the LMMPU really maximizes the trace of the matrix $\{\ddot{\boldsymbol{\beta}}_\Phi\}$ and the AD test maximizes the determinant of this matrix although it is only an approximation of the type D test. When estimating the derivatives of the power functions for larger sample sizes another apparent fact is that the ratio of the second derivatives tends to one for any pair of the LR test, W test, RS test and AD test. But this is not true for the LMMPU tests and for the Fisher test, whose local performance seems to be completely different. The LMMPU test seems being extremely sensitive to small departures of μ from the null hypothesis and much less sensitive to small departures of σ^2 than these four tests. This is not so surprising because the LMMPU test gives more power in the direction where more information is available. And from the Fisher information matrix in (1) we can see that in the direction of the location parameter μ we have twice as much information as in the direction of the scale parameter σ^2. The sensitivity of the Fisher test to small departures of μ is even only average size, and this test seems to be quite insensitive to small departures of σ^2.

Global properties

After the local considerations, let us shortly consider also the global behavior of the considered tests. Some of the results for the sample size $n = 20$ can be found in Figures 1 and 2. In the first figure we can see the power functions of the tests. The x and y axes shows the true value of parameters.

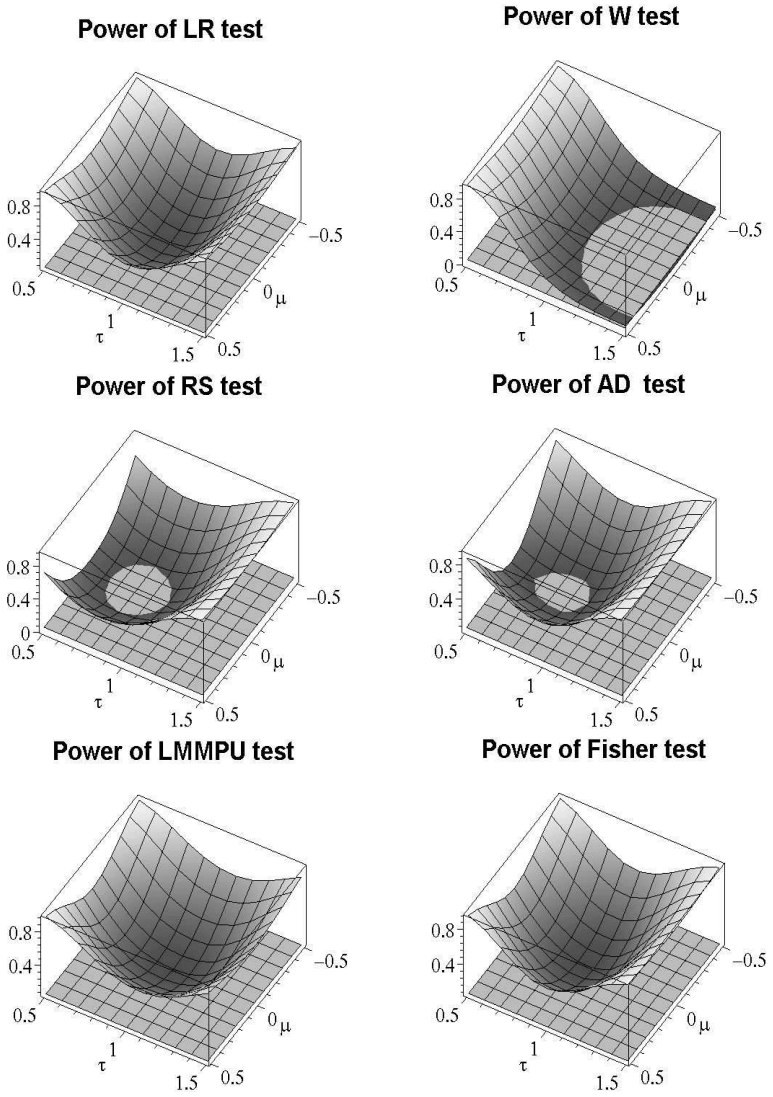


Figure 1: The plot of the power functions of the tests. The plane at the height 0.05 represents the prescribed size α of the test; $\tau = \log(\sigma)$.

For the scale parameter the logarithmic transformation is used. In all of the pictures of this figure there is a plane at the height of $z = 0.05$ which represents the prescribed size α of the test. So the area where this plane is above the surface of the power function marks the part of the parameter space where the power of a test is lower than α, which indicates that the test is doing very badly for the true values of the parameters in these region.

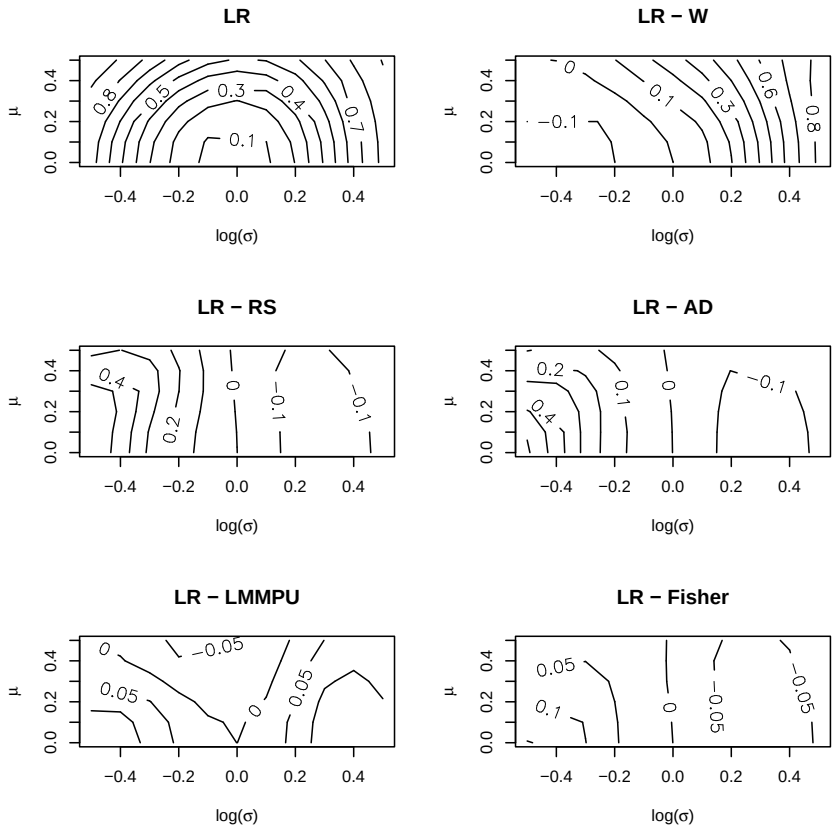


Figure 2: The contour plot of the difference of the power function of the LR test with respect to the power functions of the other tests.

Notice that especially the WT test is doing very badly for $\sigma^2 > 0$ but also the regions where the RS test and AD test are doing badly are not negligible. To be able to compare the power of the tests but to avoid large tables we draw Figure 2. Here we can see the contour plot of the difference of the power function of the LR test with respect to the power functions of the other tests. Because of the symmetry of the power functions in the parameter μ, only $\mu \geq 0$ are considered. We see immediately that the LR test has a very good performance. Although this test is not uniformly most powerful, the lack of power in the part of the parameter space is small in comparison with the excess of the power in the rest. This is especially true for the W, RS, and AD, which are not locally unbiased. The only tests which are comparable with the LR test are the locally unbiased tests. Moreover, it is interesting that the

Fisher test behaves very well, despite the small values of the derivative of the power function. As a conclusion, we recommend to choose the LMMPU test if the sensitivity to the changes in the location parameter is our main interest. However, it is difficult to compute the constants C, K of its critical region. On the other hand, the preference between the LR and Fisher tests might be a matter of taste. While the LR test does better for $\sigma^2 < 1$, the Fisher test is preferable in the opposite case. Table 1 also confirms a better local sensitivity of the LR test. However, the exact knowledge of the null distribution of the Fisher statistic strongly speaks in favour of this test, while the asymptotic critical value of the LR test, being used for small sample sizes, leads to the size exceeding 0.05 (see [4]). Therefore, the Fisher test may be a slightly more convenient. Nevertheless, our conclusions are in a good accordance with the theory of testing statistical hypotheses, because both these two most advisable tests are optimal in the Bahadur sense.

Appendix

In the following we will compute the Bahadur slopes of the considered tests. Suppose that the null hypothesis $H : \theta = \theta_0$ is rejected for large values of a test statistic T_n, and that

$$-\tfrac{2}{n}\log P_{\theta_0}(T_n \geq t) \to f(t), \qquad \text{for every t,} \tag{2}$$

$$T_n \to \phi(\theta), \quad \text{in probability} P_\theta, \tag{3}$$

for $n \to \infty$, where $f(t)$ is a function of t. Then the quantity $e(\theta) = f(\phi(\theta))$ is called the Bahadur slope of the test. Loosely speaking, the Bahadur slope measures how quickly the P-value of the test converges to zero under alternative for $n \to \infty$.

In our problem the $\theta = (\mu, \sigma^2)$ so P_θ is the probability measure of the normal distribution $\mathcal{N}(\mu, \sigma^2)$ and the T_n can be any of the test statistics suitably normalized. At first we compute the Bahadur slopes for the LR, W, RS and AD test. Let p_θ is the density of the distribution P_θ. Then it is well known (see e.g., [6]) that the upper bound for the Bahadur slope is $2\,K(\theta, \theta_0)$, where $K(\theta, \theta_0)$ is the Kullback-Leibler information number defined by

$$K(\theta, \theta_0) = \mathbf{E}_\theta \log \frac{p_\theta}{p_{\theta_0}} = \tfrac{1}{2}(-\log(\sigma^2) - 1 + \sigma^2 + \mu^2).$$

Now denote $\phi_{LR}(P_\theta) = p\lim_{n\to\infty} \frac{LR}{n} = \mu^2 + \sigma^2 - 1 - \log(\sigma^2)$. Similarly $\phi_W(P_\theta) = \frac{\mu^2}{\sigma^2} + \frac{(\sigma^2-1)^2}{2\sigma^4}$, $\phi_{RS}(P_\theta) = \mu^2 + \frac{1}{2}(\mu^2 + \sigma^2 - 1)^2$ and $\phi_{AD}(P_\theta) = \mu^2 + \frac{1}{2}(\sigma^2 - 1)^2$. To calculate the limit (2) we use Sanov's theorem (see [6] pp. 209 – 210) and for the LR test we get

$$f_{LR}(t) = 2 \inf_{\phi_{LR}(P_\theta)\geq t} \mathbf{E}_\theta \log \frac{p_\theta}{p_{\theta_0}} = \inf_{\phi_{LR}(P_\theta)\geq t} (-\log(\sigma^2) - 1 + \sigma^2 + \mu^2) = t.$$

Analogously we get

$$f_W(t) = \inf_{\phi_W(P_\theta)\geq t}(-\log(\sigma^2) - 1 + \sigma^2 + \mu^2) = \tfrac{1-\sqrt{2t}}{1-2t} - \log\left(\tfrac{1-\sqrt{2t}}{1-2t}\right) - 1$$
$$f_{RS}(t) = f_{AD}(t) = \sqrt{2t} - \log(\sqrt{2t}+1).$$

Now the Bahadur slopes of the tests are

$$e_{LR}(\theta) = f_{LR}(\phi_{LR}(P_\theta)) = \phi_{LR}(P_\theta) = \mu^2 + \sigma^2 - 1 - \log(\sigma^2)$$
$$e_{WT}(\theta) = f_{WT}(\phi_{WT}(P_\theta)) = \tfrac{1-\sqrt{2t}}{1-2t} - \log\left(\tfrac{1-\sqrt{2t}}{1-2t}\right) - 1\Big|_{t=\frac{\mu^2}{\sigma^2}+\frac{(\sigma^2-1)^2}{2\sigma^4}}$$
$$e_{RS}(\theta) = f_{RS}(\phi_{RS}(P_\theta)) = \sqrt{2t} - \log(\sqrt{2t}+1)|_{t=\mu^2+\frac{1}{2}(\mu^2+\sigma^2-1)^2}$$
$$e_{AD}(\theta) = f_{AD}(\phi_{AD}(P_\theta)) = \sqrt{2t} - \log(\sqrt{2t}+1)|_{t=\mu^2+\frac{1}{2}(\sigma^2-1)^2}.$$

The Bahadur slope of the Fisher test can be computed for example in this way. At first we obtain the individual Bahadur slopes of the test statistics $\bar{X}$ and S^2, which are μ^2 and $\sigma^2 - 1 - \log(\sigma^2)$, respectively. Then, as noted in [3], we can get the Bahadur slope of the Fisher test simply by adding these two slopes together.
The author unfortunately does not know how to calculate the Bahadur slope of the LMMPU test.

References

[1] Gupta A.S., Vermeire L. (1986). *Locally optimal tests for multiparameter hypotheses*. J. Amer. Statist. Assoc. **81**, 819–825.

[2] Isaacson S.L. (1951). *On the theory of unbiased tests of simple statistical hypothesis specifying the values of two or more parameters*. Ann. Math. Statist. **22**, 87–93.

[3] Littel R.C., Folks J.L. (1976). *A test of equality of two normal population means and variances*. J. Amer. Statist. Assoc. **71**, 968–971.

[4] Omelka M. (2004). *The behavior of locally most powerful tests*. To appear in Kybernetika.

[5] Peers H.W. (1971). *Likelihood ratio and associated test criteria*. Biometrika **58**, 577–587.

[6] van der Vaart A.W. (1998). *Asymptotic statistics*. Cambridge University Press.

Acknowledgement: The author would like to thank Prof. J. Jurečková for all the support he gets from her. The work was supported by the grant GAČR 201/03/0945 and by the Research plan MSM 113200008.

Address: M. Omelka, Department of Probability and Mathematical Statistics, Charles University in Prague, Sokolovská 83, CZ-186 75 Prague 8, Czech Republic

E-mail: `omelka@karlin.mff.cuni.cz`

TESTY A ODHADY PARETOVA INDEXU

Jan Picek

Klíčová slova: Paretův index, rozdělení extrémních hodnot, sféra přitažlivosti, Hillův odhad.

Abstrakt: Nechť $X_1,\ X_2, \ldots$ jsou nezávislé stejně rozdělené náhodné veličiny s distribuční funkcí F a nechť $M_n = \max(X_1, \ldots, X_n)$. Pro většinu obvyklých distribučních funkcí vhodně standardizovaná maxima M_n konvergují v distribuci k rozdělení extrémních hodnot G_γ. Podle hodnot shape parametru γ rozlišujeme tři základní třídy distribučních funkcí: $\gamma > 0$ – Fréchetova třída, $\gamma = 0$ Gumbelova a $\gamma < 0$ Weibullova. Z hlediska extrémních událostí je především zajímává třída Fréchetova, γ se v tomto kontextu často nazývá Paretovým indexem. V příspěvku se proto budeme zabývat semiparametrickými odhady γ především pro tuto třídu a testy o γ, zvláště se bude jednat o testy hypotézy $\gamma = 0$ proti alternativě $\gamma > 0$, tj. náhodný výběr je z rozdělení, který patří do Gumbelovy třídy proti alternativě, že rozdělení je z Fréchetovy třídy.

1 Úvod

Nechť $X_1,\ X_2, \ldots$ jsou nezávislé stejně rozdělené náhodné veličiny s distribuční funkcí F. Naše pozornost v tomto článku bude soustředěna na extremální události. Nechť tedy $M_n = \max(X_1, \ldots, X_n)$. Zřejmě distribuční funkce M_n je

$$P\left(M_n \leq x\right) = P\left(X_1 \leq x, \ldots, X_n \leq x\right) = F^n(x) \text{ s.j..}$$

Jednoduše je potom možné ukázat, že $M_n \to x_F$ s.j. pro $n \to \infty$, kde

$$x_F := \sup\{x \in I\!R : F(x) < 1\} \leq \infty.$$

Tato skutečnost nám neposkytne příliš mnoho informace. Pokud se inspirujeme centrální limitní větou, jistě je přirozené se zabývat standardizovanými maximy.

Předpokládejme, že můžeme najít posloupnost reálných čísel $a_n > 0$ a b_n tak, že posloupnost $(M_n - b_n)/a_n$ konverguje v distribuci, t.j.

$$P\left((M_n - b_n)/a_n \leq x\right) = F^n(a_n x + b_n) \to G(x),\ n \to \infty, \tag{1}$$

pro nějakou nedegenerovanou d.f. $G(x)$

Jestliže podmínka platí, říkáme, že F je ve sféře přitažlivosti G (domain of attraction) – $F \in \mathrm{MDA}(G)$. Přirozeně nás patrně napadnou otázky: jak vypadá G, jaké podmínky musí F splňovat, aby $F \in \mathrm{MDA}(G)$ a jak volit a_n a b_n . Odpověď na tyto základní otázky můžeme najít např. v [2].

Odpověď na první otázku známe už od roku 1928 – Fisherova-Tippettova věta: Jestliže $F \in \mathrm{MDA}(G)$ potom G je typu jedné z následujících tří distribučních funkcí:

Fréchet $\quad \Phi_{1/\gamma}(x) = \begin{cases} 0, & x \le 0 \\ \exp\left(-x^{-1/\gamma}\right), & x > 0 \end{cases} \quad \gamma > 0$

Weibull $\quad \Psi_{1/\gamma}(x) = \begin{cases} \exp\left\{-(-x)^{1/\gamma}\right\}, & x \le 0 \\ 1 & x > 0 \end{cases} \quad \gamma < 0$

Gumbel $\quad \Lambda(x) = \exp\left(-e^{-x}\right), \quad x \in \mathbb{R}.$

Po vhodné reparametrizaci můžeme tyto tři třídy charakterizovat jediným rozdělením – zobecněněným rozdělením extrémních hodnot (Generalized Extreme Value Distribution)

$$G(x) = G_\gamma(x) = \begin{cases} \exp\left(-(1+\gamma x)^{-1/\gamma}\right) & \gamma \neq 0 \\ \exp(-e^{-x}) & \gamma = 0 \end{cases},$$

kde $1 + \gamma x > 0$.

Hodnota „shape“ parametru $\gamma > 0$ odpovídá Fréchetově třídě, $\gamma = 0$ Gumbelově a $\gamma < 0$ Weibullově. Fisherova-Tippettova věta nám pak říká: jestliže vhodně standardizované maxima konvergují v distribuci k nedegenerované limitě, potom limitní rozdělení musí být rozdělení extrémních hodnot. Poznamenejme, že G je určena jednoznačně až na parametr polohy a měřítka.

Je možné ukázat, že v podstatě všechny běžně uvažované spojité rozdělení splňují podmínku (1).

Než se zaměříme na volbu a_n a b_n připomeňme několik pojmů z klasické teorie extrémních událostí.

Funkce $h(t)$ na $(0, \infty)$ je *pravidelně se měnící* funkce (regularly varying) v ∞ s indexem $\alpha \in \mathbb{R}$ ($h \in \mathcal{R}_\alpha$), jestliže

$$\lim_{x\to\infty} \frac{h(xt)}{h(x)} = t^\alpha, \ t > 0.$$

Funkce $L(t)$ na $(0, \infty)$ je *pomalu se měnící* funkce (slowly varying) v ∞ ($L \in \mathcal{R}_0$), jestliže

$$\lim_{x\to\infty} \frac{L(xt)}{L(x)} = 1, \ t > 0.$$

V oblasti extrémních hodnot se často pracuje s kvantilovou funkcí chvostu

$$U(t) = F^{-1}\left(1 - \frac{1}{t}\right) = \inf\{y : F(y) \ge 1 - 1/t\}, \ t > 0.$$

Věta 1.1. *a)* $F \in MDA(G_\gamma)$ *právě když*

$$\lim_{t\to\infty} \frac{U(tx) - U(t)}{a(t)} = \frac{x^\gamma - 1}{\gamma}$$

pro každé $x > 0$, *a je nějaká kladná funkce a* $\gamma \in \mathbb{R}$, $a_n = a(n), \ b_n = U(n)$.

b) $F \in \ MDA(G_\gamma),\ \gamma > 0$ *právě když*

$$\lim_{t\to\infty} \frac{U(tx)}{U(t)} = x^\gamma \tag{2}$$

pro každé $x > 0$ *s* $\gamma > 0$*, tj.* $U \in \mathcal{R}_\gamma$ *(*$a_n = U(n)$*).*

Důkaz a detaily např. v de Haan L. (1970).

Další a často používá charakterizace Fréchetovy třídy:

- $F \in \ \mathrm{MDA}(G_\gamma),\ \gamma > 0$ právě když $1 - F(x) \in \mathcal{R}_{-1/\gamma}$, tj. chvost rozdělení F je pravidelně se měnící funkce v ∞ s indexem $-1/\gamma$
- $$1 - F(x) = x^{-1/\gamma} L(x). \tag{3}$$

Statistickou inferenci v extremální statistice můžeme založit na základě limitního rozdělení, tj. na zobecněném rozdělením extrémních hodnot např. pomocí metody maximální věrohodnosti. Ukazuje se, že konvergence je však velmi pomalá, proto je nutné hledat alternativní přístupy. V následujícím textu ukážeme některé možné semiparametrické přístupy.

2 Testy

Případ $F \in \mathrm{MDA}(G_0)$ je zajímavý pro mnoho aplikací, které se zabývají extrémy. Důvodem je nejen jednodušší inference založená na Gumbelově sféře přitažlivosti, ale také široká paleta rozdělení s exponenciální chvosty. Jako zástupce jmenujme normální, lognormální a gamma rozdělení. Na druhé straně opravdu extrémní události jsou modelovány pomocí rozdělení z Fréchetovy třídy. Je tedy určitě v praxi užitečné rozhodnout do jaké třídy rozdělení našich dat patří. To znamená uvažovat následující test oboustranné hypotézy (respektive anologický jednostranný test)

$$F \in \mathrm{MDA}(G_0) \quad \text{proti alternativě} \quad F \in \mathrm{MDA}(G_\gamma)_{\gamma\neq 0}. \tag{4}$$

Asi nejpoužívanější test pro tuto situaci navrhli Hasofer A.M. and Wang Z. v roce 1992. Najdeme ho implementovaného v řadě softwarů pro statistiku extrémních událostí. Test jako většina semiparametrických postupů je založen na k největších pořádkových statistikách:

$$W_k = \frac{k\left(\overline{X}_k - X_{n-k+1:n}\right)^2}{(k-1)\sum_{i=1}^{k}\left(X_{n-i+1:n} - \overline{X}_k\right)^2}, \qquad \overline{X}_k := \frac{1}{k}\sum_{i=1}^{k} X_{n-i+1:n}. \tag{5}$$

Hasofer a Wang ukázali, že testová statistika W_k má asymptoticky normální rozdělení se střední hodnotou μ_k a rozptylem σ_k^2

$$\mu_k = \frac{1}{(k-1)}, \quad \sigma_k^2 = \frac{4(k-2)}{(k-1)^2}\frac{1}{(k+1)(k+2)}$$

Kritický obor pro oboustrannou alternativu je potom dán následovně

$$|W_k^*| > u_{1-\alpha/2},$$

kde $W_k^* := (W_k - \mu_k)/\sigma_k$ a u_ε je ε-kvantil normálního rozdělení.

Při praktickém provádění testu jistě narazíme na problém, jak zvolit vhodné k. Pokud budeme k zvyšovat, zvýšíme sílu testu, ale na druhé straně zvyšující se podíl k/n má neblahý vliv na chybu I.druhu. Volba se pak stává do jisté míry „alchymií", nicméně v literatuře existují doporučení, např. Boos navrhuje k/n=0.2 pro $50 \leq n \leq 500$ a $k/n = 0.1$ pro $500 < n \leq 5000$, Galambos radí volit $k = 2\sqrt{n}$.

Podobný typ testu navrhli C. Neves, J. Picek a M. I. Fraga Alves (2005). Jako testovou statistiku uvažují

$$T_{k,n}^* = \frac{X_{n:n} - X_{n-k:n}}{\frac{1}{k}\sum_{i=1}^{k}(X_{n-i+1:n} - X_{n-k:n})} - \log k \ . \tag{6}$$

Ukázali, že testová statistika $T_{k,n}^*$ za nulové hypotézy konverguje k Gumbelovu rozdělení $G(x) = \exp(-e^{-x})$ a že test je konzistentní. Nulová hypotéza je tedy zamítnuta na asymptotické hladině $\alpha \in (0,1)$ jestliže $T_{k,n}^* < g_{\alpha/2}$ nebo $T_{k,n}^* > g_{1-\alpha/2}$, kde g_ε označuje ε-kvantil Gumbelova rozdělení, tj. $g_\varepsilon = -\log(-\log\varepsilon)$.

Jako poslední přístup pro test (4) uveďme poměrně nedávný přístup J. Segerse a J.Teugelse (2001). Vychází z poměru uvažovaném Galtonem (1902):

$$G_n = \frac{X_{n:n} - X_{n-2:n}}{X_{n-1:n} - X_{n-2:n}}$$

Náhodný výběr o rozsahu n je rozdělen do m skupin $\sum_{i=1}^{m} n_i = n$. V každé je spočítán poměr

$$\xi_i = \frac{X_{n_i:n_i}^{(i)} - X_{n_i-2:n_i}^{(i)}}{X_{n_i-1:n_i}^{(i)} - X_{n_i-2:n_i}^{(i)}}, \qquad 1, \cdots, m$$

Podle Serflinga (1980), Segers a Teugels navrhují užít testovou statistiku

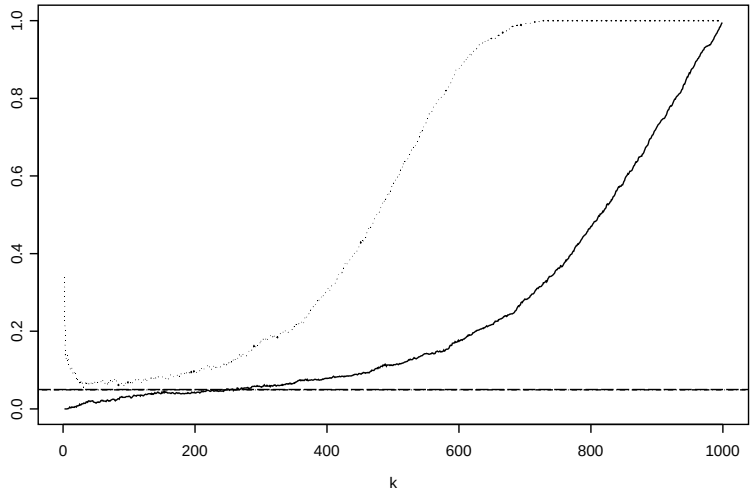
$$S_m = \frac{5}{m}\left(\sum_{i=1}^{m} T(\xi_i)\right)^2, \qquad T(x) := 1 - \frac{6x}{(1+x)^2}, \tag{7}$$

a ukazují, že za nulové hypotézy konverguje k χ_1^2 rozdělení pro $m \to \infty$.

Nulová hypotéza je tedy zamítnuta na asymptotické hladině α, je-li $S_m > \chi_1^2(1-\alpha)$, kde $\chi_1^2(\varepsilon)$ označuje ε-kvantil χ^2 rozdělení s 1 st. vol.

2.1 Numerická ilustrace

Zkusme ilustrovat chování výše uvedených testů na simulovaných datech a na jednom reálném příkladu. Nejprve jsme uvažovali platnost nulové hypotézy (4), tj. jako zástupce z Gumbelovy sféry přitažlivosti jsme zvolili Gumbelovo rozdělení $F(x) = \exp(-e^{-x})$. Z tohoto rozdělení jsme vygenerovali $1000 \times$ výběr o rozsahu 1000 a provedli výše uvedené testy. Na obr. 1 jsou zobrazeny výsledky ve formě relativního počtu zamítnutí nulové hypotézy na hladině $\alpha = 0.05$. Testy (5) a (6) byly provedeny pro $k = 2, \ldots, 999$ (počet použitých nejvyšších pořádkových statistik). Test (7) byl konstruován tak, že výběr byl rozdělen do 50 $(=m)$ bloků o rozsahu 20. Obr. 1 vlastně ilustruje odhad chyby prvního druhu. Je vidět, že odhad této chyby pro test (7) je prakticky 0.05, pokud přijmeme výše zmiňovaná doporučení pro volbu k, potom testy (5) a (6) mají odhad také blízký 0.05. Nicméně se zdá, že test (6) dovolí volit větší rozsah k aniž by to mělo výrazný vliv na chybu I. druhu. Testovali jsme i jiná rozdělení z Gumbelovy sféry přitažlivosti i pro jiné rozsahy, charakter křivek byl podobný s jedinou výjimkou a to exponenciálním rozdělením, pro které odhad chyby prvního druhu byl stabilní (blízko hodnoty 0.05) prakticky pro všechna možná k.



Obrázek 1: Relativní počet zamítnutí H_0 na hladině $\alpha = 0.05$ pro Gumbelovo rozdělení, $T^*_{k,n}$ (plná čára), W_k (tečkovaně), S_{50} (čerchovaně).

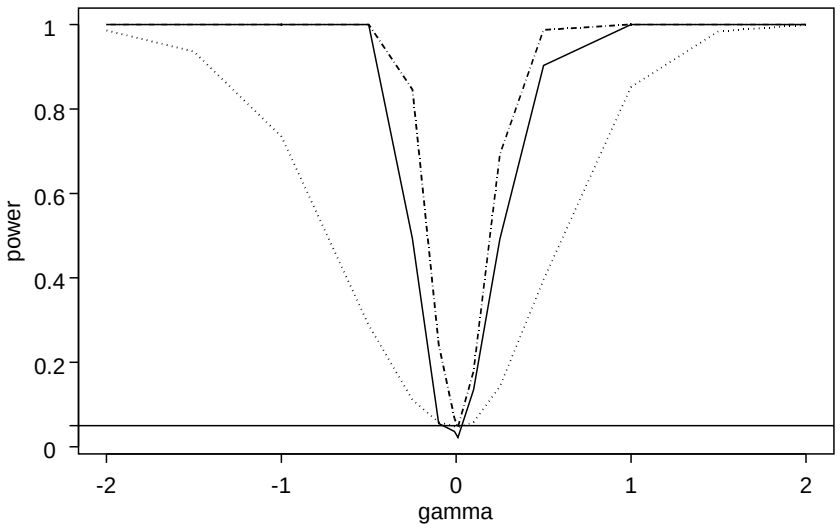
Jako další zástupce pro ilustraci bylo zvoleno zobecněné Paretovo rozdělení

$$F_\gamma(x) := 1 + \log G_\gamma(x) = 1 - (1 + \gamma x)^{-\frac{1}{\gamma}}$$

$$\text{pro} \begin{cases} x \geq 0 & \text{jestliže} \quad \gamma \geq 0 \\ 0 \leq x \leq -\frac{1}{\gamma} & \text{jestliže} \quad \gamma < 0 \end{cases}$$

Toto rozdělení závisí na parametru γ. Podle jeho hodnoty patří rozdělení do jedné z uvažovaných tříd. Opět byl 1000 krát generován výběr o rozsahu 1000 pro hodnoty $\gamma = -2.0$, -1.5, -1.0, -0.5, -0.25, -0.1, -0.01, 0.01, 0.1, 0.25, 0.5, 1.0, 1.5, 2.0. Pokud se opět zajímáme o relativní počet zamítnutí nulové hypotézy, pak v tomto kontextu dostáváme představu o síle testů.

Na obr. 2 vidíme srovnání pro všechny tři testy v závislosti na γ data. Testy (5) a (6) byly provedeny pro $k = 150$, test (7) s $m = 50$. Ve všech třech případech tak bylo použito 150 hodnot (i když ne nutně stejných).

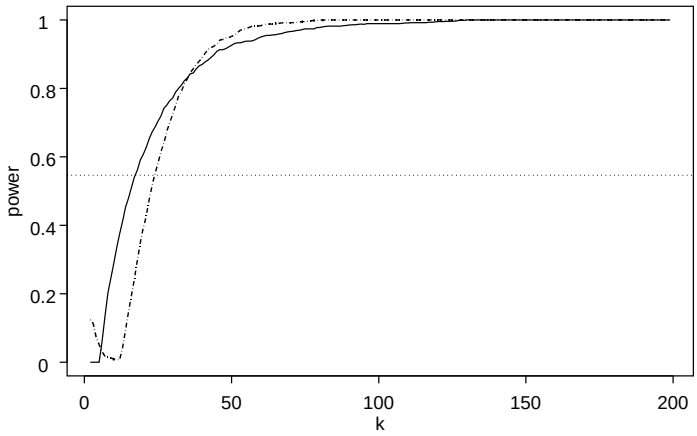


Obrázek 2: Síla testu: $T^*_{150,n}$ (plná), W_{150} (čerchovaná), S_{50} (tečkovaná) na hladině $\alpha = 0.05$ pro zobecněné Pareto ($\gamma = -2.0$, -1.5, -1.0, -0.5, -0.25, -0.1, -0.01, 0.01, 0.1, 0.25, 0.5, 1.0, 1.5, 2.0), rozsah $n = 1000$.

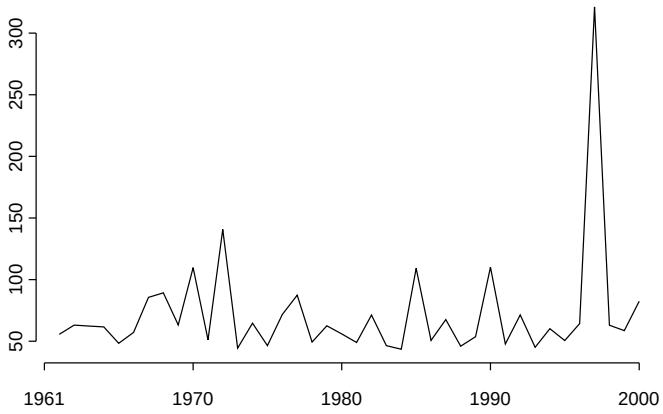
Vidíme, že z hlediska síly testu se nejlépe chová test (5), trochu hůře (6) a nejslabší je test (7). Ten byl nejslabší ve všech případech, které jsme zkoumali. Testy (5) a (6) se příliš nelišily a záviselo na konkrétní volbě k, rozdělení a rozsahu. Dokladem toho může být např. obr. 3, který zobrazuje závislost „síly testu“ na volbě k pro zobecněné Pareto rozdělení s $\gamma = 1.0$

Co se týče asymptotických vlastností a předpokladů všechny tři testy jsou rovnocenné, na druhou stranu vidíme, že pokud máme i poměrně velký rozsah dat, rozdíly najít můžeme. Nejslabším testem se zdá být do jisté míry (7). Test (5) je v praxi patrně nejpoužívanější, ale zdá se, že (6) je plně srovnatelný.

Podívejme se též na testy na reálných datech. V poslední době se vede diskuse, že počasí nabývá extrémního chování. Jedním z mnoha charakteristik tohoto chování počasí mohou být např. extrémní srážky. V České Republice jsou k dispozici data na řadě stanic od roku 1961. Extrémní srážky můžeme



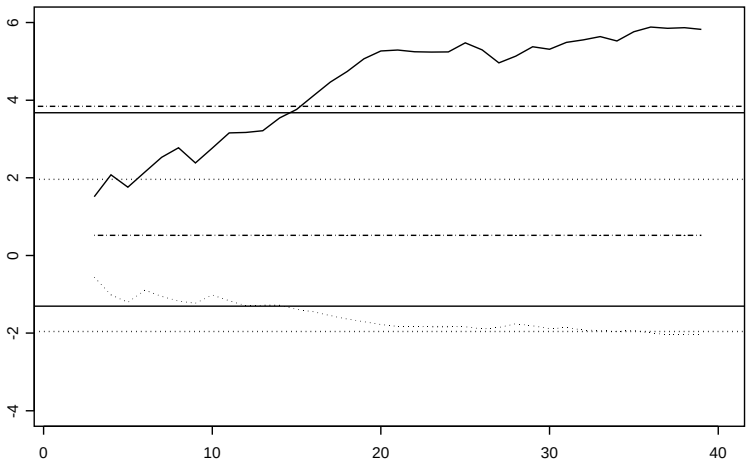
Obrázek 3: Síla testu: $T^*_{k,n}$ (plná čára), W_k (čerchovaná), S_{20} (tečkovaná) na hladině $\alpha = 0.05$ pro v závislosti na k, rozsah $n = 200$ pro zobecněné Pareto rozdělení s $\gamma = 1.0$ (vpravo).



Obrázek 4: Maximální třídenní úhrny srážek v letech 1961-2000 ve Valašském Meziříčí.

třeba charakterizovat maximálními třídenními úhrny srážek v daném roce (takovéto data měl autor k dispozici). Na obr. 4 vidíme tyto data pro stanici ve Valašském Meziříčí. Velmi dobře je vidět výjimečný rok 1997, který přinesl velké záplavy na Moravě.
Je otázkou pro další statistické úvahy, jaký základní model je pro tuto veličinu

(maximálními třídenními úhrny srážek v daném roce) vhodný, tj. Gumbelova nebo Fréchetova třída. Výsledky testů jsou graficky zobrazeny na obr. 5, kde vodorovné čáry odpovídají příslušným 97.5%-ním kvantilům pro oboustranný test. Vidíme, že zamítnutí nulové hypotézy je velmi problematické, zamítáme pouze pro větší hodnoty k a to hlavně testem (6), viděli jsme ze simulací, že větší k nedávají dobré výsledky co se týče platnosti nulové hypotézy. Hlavním problémem tu je však velmi malý počet pozorování ($n = 40$), který je v aplikacích týkajících se extrému nedostatečný, ale bohužel v praxi častý.



Obrázek 5: Srážky ve Valašském Meziříčí: Hodnoty $T^*_{k,40}$ (plná), W_k (tečkovaná), S_8 (čerchovaná). Vodorovné linky označují příslušné kvantily odpovídající $\alpha = 0.05$.

Další testy, které byly v poslední době konstruovány, uvažují hypotézy o hodnotách parametru γ (jak „těžké“ jsou těžké chvosty) pro $F \in \mathrm{MDA}(G_\gamma)$, $\gamma > 0$, viz [11], [16]. O těchto testech bylo referováno na Robustu 2000. Pokud se budeme zabývat úvahami o hodnotách γ, pak mnohem bohatší je literatura věnovaná odhadům. Proto následující část tohoto příspěvku věnujeme právě jim.

3 Odhady

Připomeňme, že vycházíme z náhodného výběru $X_1,\ X_2, \ldots$ z rozdělení s neznámou distribuční funkcí F. Pokud $F \in \mathrm{MDA}(G_\gamma),\ \gamma > 0$, pak patrně nejznámější odhadem γ je Hillův odhad z roku 1975 [8]:

$$H_n(k) = \frac{1}{k} \sum_{i=0}^{k-1} \log X_{(n-i:n)} - \log X_{(n-k:n)}. \tag{8}$$

Ukažme návrh jedné z možných cest jeho odvození. Vyjděme z charakterizace Fréchetovy třídy (2):

$$\lim_{t\to\infty} \frac{U(tx)}{U(t)} = x^\gamma, \text{ kde} \qquad U(t) = F^{-1}(1-1/t).$$

Po zlogaritmování dostaneme $\lim_{t\to\infty} \log U(t/x) - \log U(t) = -\gamma \log x$.

Výběrová verze kvantilové funkce chvostu U je $U_n(1/x) = F_n^{-1}(1-x) = X_{n(1-x),n}$, tj. $U_n(\frac{n}{k}) = X_{n-k,n}$ a $U_n(\frac{n}{kx}) = X_{n-kx,n}$. Tedy pro $0 < x < 1$ je $\log X_{n-kx,n} - \log X_{n-k,n} = -\gamma \log x$. Poté integrujme

$$\gamma = -\gamma \int_0^1 \log x \, dx = \lim_{t\to\infty} \int_0^1 \{\log U(t/x) - \log U(t)\} \, dx.$$

Dostaneme tak možný odhad γ

$$\begin{aligned} H_n(k) &= \int_0^1 (\log X_{n-kx,n} - \log X_{n-k,n}) \, dx \\ &= \frac{1}{k} \sum_{i=0}^{k-1} \log X_{(n-i:n)} - \log X_{(n-k:n)} \end{aligned}$$

Hillův odhad je konzistentní, tvrzení najdeme např. v [13].

Věta 3.1. *Je-li* $F \in MDA(G_\gamma)$, $\gamma > 0$, *potom* $H_n(k) \to \gamma$ *v pravděpodobnosti,* $k = k(n) \to \infty$, $k(n)/n \to 0 (n \to \infty)$.

Pokud nás zajímá asymptotické rozdělení odhadu, musíme klást další podmínky na distribuční funkci, abychom byli schopní ho odvodit. Nejčastěji se uvažuje následující podmínka (*regular variation of second order*): Nechť existuje $A(t)$ funkce konstantního znaménka a parametr ρ

$$\lim_{t\to\infty} \frac{\frac{U(tx)}{U(t)} - x^\gamma}{A(t)} = x^\gamma \cdot \frac{x^\rho - 1}{\rho} \tag{9}$$

pro všechna $x > 0$.

Věta 3.2. *Nechť podmínka (9) platí a nechť posloupnost* $k = k(n)$ *je taková, že* $k(n) \to \infty$ *a* $\sqrt{k}A(n/k) \to 0$, *potom*

$$\sqrt{k}(H_n(k) - \gamma)$$

je asymptoticky normální s nulovou střední hodnotou a rozptylem γ^2.

Pokud uvažujeme $F \in \mathrm{MDA}(G_\gamma)$, γ libovolné, pak lze „analogicky odvodit“ **Momentový odhad** [1]

$$M(k) = 1 + M(k)^{(1)} + \frac{1}{2}\left(\frac{\left(M(k)^{(1)}\right)^2}{M(k)^{(2)}} - 1\right)^{-1}, \tag{10}$$

kde

$$M(k)^{(j)} = \frac{1}{k}\sum_{i=1}^{k}\left(\log X_{(Nn-i+1:Nn)} - \log X_{(Nn-k:Nn)}\right)^j.$$

Alternativou momentového odhadu je **Pickandsův odhad** [17]

$$P(k) = \frac{1}{\log 2}\log\left(\frac{X_{Nn-k+1:Nn} - X_{Nn-2k+1:Nn}}{X_{Nn-2k+1:Nn} - X_{Nn-4k+1:Nn}}\right). \tag{11}$$

Výše uvedené odhady jsou patrně nejznámější, v literatuře existuje obrovské množství dalších odhadů: různá zobecnění Hillova odhadu, odhady založené na parametru druhého řádu ρ, viz (9) a mnoho a mnoho dalších alternativ. Uveďme alespoň jeden příklad, který vychází z (9) a uvažuje, že $\rho = -1$. Navrhli ho Gomes a Martin v roce 2002, viz [9].

$$GM(k) = \frac{1}{k}\sum_{i=1}^{k} U_i - \left(\frac{1}{k}\sum_{i=1}^{k} iU_i\right)\frac{\sum_{i=1}^{k}(2i-k-1)U_i}{\sum_{i=1}^{k} i(2i-k-1)U_i},$$

$$U_i = i\left[\log\frac{X_{Nn-i+1:Nn}}{X_{Nn-i:Nn}}\right], \tag{12}$$

Stejně jako u testů je problém volby k, lze řešit podobnými doporučeními nebo se uvažují postupy založené na bootstrapu - viz např. [10].

Pokud se podíváme do domácích luhů a hájů, tak i tady najdeme příspěvek ke konstrukci odhadů parametru γ za podmínky $F \in \mathrm{MDA}(G_\gamma)$, $\gamma > 0$. Tyto odhady nejsou založeny přímo na pořádkových statistikách na rozdíl od předcházejících. Vychází se opět z určité charakterizace Fréchetovy třídy:

$$\lim_{a\to\infty}\frac{-\log(1-F(a))}{m\log a} = 1. \tag{13}$$

Aplikací l'Hospitalova pravidla z von Mises podmínek (viz Embrechts a kol., Kap. 3), dostaneme, že $1 - F(x) = x^{-m}L(x)$, což je charakterizace uvedená v (3). Platí i opačná implikace. Principem spočívá v rozdělení výběru do skupin, v každé je spočtena nějaká jednoduchá statistika. Výsledný odhad je konstruován na základě empirické distribuční funkce sledované statistiky.

O prvním typu odhadu referovala na Robustu 2000 A. Fialová: Rozdělíme pozorování do N nepřekrývajících se výběrů rozsahu n a určíme zde průměry $(\bar{X}_n^{(1)}, \ldots, \bar{X}_n^{(N)})$. Dostaneme pak náhodný výběr z rozdělení s distribuční funkcí $F_{\bar{X}_n}(x) = \mathbb{P}(\bar{X}_n \leq x)$. Označíme $\widehat{F}_{\bar{X}_n}^{(N)}(x) = \frac{1}{N}\sum_{j=1}^{N} I[\bar{X}_n^{(j)} \leq x]$ empirickou distribuční funkcí založenou na $(\bar{X}_n^{(1)}, \ldots, \bar{X}_n^{(N)})$. Vyberme posloupnost $\{a_N\}_{N=1}^{\infty}$, $a_N \to \infty$ pro $N \to \infty$ ve tvaru $a_N = N^{\frac{1-\delta}{m_0}}$, s pevným $\delta \in (0,1)$.

Odhad parametru $m = 1/\gamma$ je potom

$$\widehat{m}_N = \tilde{m}_N(a_N) I[0 < \widehat{F}_{\bar{X}_n}^{(N)}(a_N) < 1] + m_0 I[\widehat{F}_{\bar{X}_n}^{(N)}(a_N) = 0 \text{ nebo } 1], \quad (14)$$

kde

$$\tilde{m}_N(a) = \frac{-\log\left(1 - \widehat{F}_{\bar{X}_n}^{(N)}(a)\right)}{\log a}, \quad a > 0.$$

Odhad (14) je konzistentní a jeho asymptotické rozdělení je normální, viz následující věty.

Věta 3.3. *Nechť* $\{X_1, X_2, \ldots\}$ *je posloupnost nezávislých stejně rozdělených náhodných veličin s distribuční funkcí* $F \in MDA(G_\gamma)$, $\gamma > 0$ *a hustotou* $f(x) = 0$ *pro* $x < 0$ *a* $0 < f(x) < \infty$ *for* $x \geq K_f \geq 0$. *Nechť* $\widehat{m}_N$ *je odhad* m. *Potom*

$$\widehat{m}_N \to m \quad \textit{s pravděpodobností } 1, \quad \textit{pro } N \to \infty.$$

Věta 3.4. *Za podmínek předcházející věty posloupnost*

$$N^{\frac{1}{2}} \log a_N \left(\frac{1 - F_{\bar{X}_n}(a_N)}{F_{\bar{X}_n}(a_N)}\right)^{\frac{1}{2}} \left(\widehat{m}_N - m + \frac{\log L^*(a_N)}{\log a_N}\right)$$

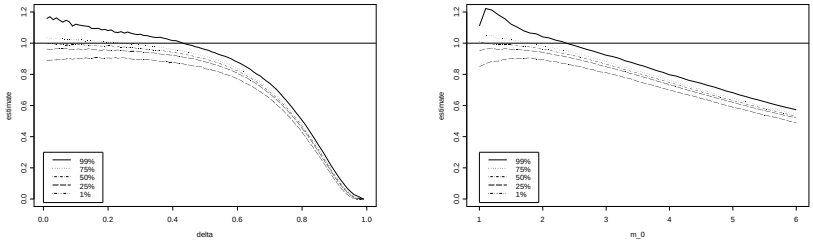
je asymptoticky normální $N \to \infty$.

Důkazy obou vět lze nalézt v [4]. Na rozdíl od Hillova (a i dalších výše zmíněných) odhadů je asymptotické rozdělení odvozeno za mnohem slabších předpokladů. Bohužel výsledek věty 3.4 obsahuje pomalu měnící se funkci L^*, kterou zpravidla neznáme, není proto možné jednoduše výsledku využít např. pro konstrukci intervalových odhadů. Pro tento odhad musíme zvolit δ, což je vlastně podobná úloha jako je určení vhodného k pro předcházející odhady, navíc je však nutné zvolit m_0, což vyžaduje nějakou počáteční informaci o tom, jak chvost rozdělení může být „těžký“. To poněkud omezuje užití odhadu pro praktické problémy.

Ilustrujme na simulovaných datech na chování odhadu právě v závislosti na volbě δ a m_0. Jako model dat použijeme Paretovo rozdělení, které je jedním z typicky používaných rozdělení pro popis extrémních událostí:

$$F(x) = 1 - \left(\frac{1}{1+x}\right)^{1/\gamma}, \quad x \geq 0 \quad (15)$$

Konkrétně byla simulace provedena pro Paretovo rozdělení s $\gamma = 1$, což je i hodnota, kterou chceme odhadnout. Výsledek můžeme vidět na obr. 6, z kterého vyplývá, že pokud je zhruba $\delta < 0.4$ je odhad poměrně stabilní a rozumný. Pro velké hodnoty δ odhad naprosto selhává. Zároveň je vidět, že čím horší máme apriorní informaci o správné hodnotě $\gamma = 1/m$, tím dostaneme horší výsledek.



Obrázek 6: Závislost odhadu v 1000 simulovaných výběrech Paretova rozdělení s $\gamma = 1$ na parametru δ pro dané $m_0 = 1.5$ (vlevo) závislost hodnot odhadu na parametru m_0 pro $\delta = 0.1$ (vpravo). Uvedeny jsou medián 1, 25, 75 a 99 percentily.

Jurečková, Picek (2004) navrhli odhad vycházející z postupů pro testování hypotézy o hodnotách parametru γ pro $F \in \mathrm{MDA}(G_\gamma)$, $\gamma > 0$. Krátká poznámka o nich byla v předcházející kapitole. Invertováním těchto testů (v duchu způsobu, který užil Hodges a Lehmann v roce 1963) dostaneme odhad

$$M_N = \frac{1}{2}(M_N^+ + M_N^-),$$

kde

$$M_N^- = \sup\{s:\ 1 - \hat{F}_N^*(a_{N,s})) < N^{-(1-\delta)}\},$$

$$M_N^+ = \inf\{s:\ 1 - \hat{F}_N^*(a_{N,s})) > N^{-(1-\delta)}\}.$$

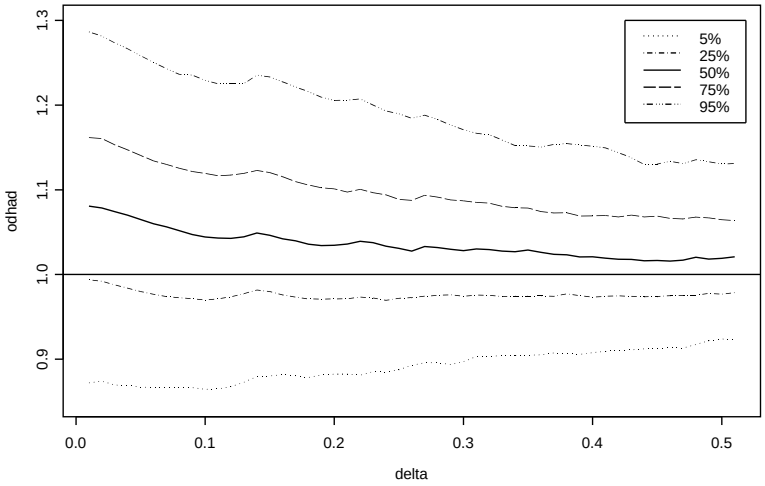
$X_{(n)}^{(1)}, \ldots, X_{(n)}^{(N)}$ jsou odpovídající výběrová maxima N skupin o rozsahu n, které vznikly rozdělením původního náhodného výběru. Jako $\hat{F}_N^*$ označujeme empirickou distribuční funkci odpovídající výběrovým maximům, $a_{N,m} = (nN^{1-\delta})^{1/m}$, kde $0 < \delta < \frac{1}{2}$ je konstanta.

Ilustrujme podobně jako u předcházejícího odhadu chování v závislosti na volbě δ. Jako model dat tentokráte použijeme Burrovo rozdělení, které je dalším typicky používaným rozdělením pro popis extrémních událostí:

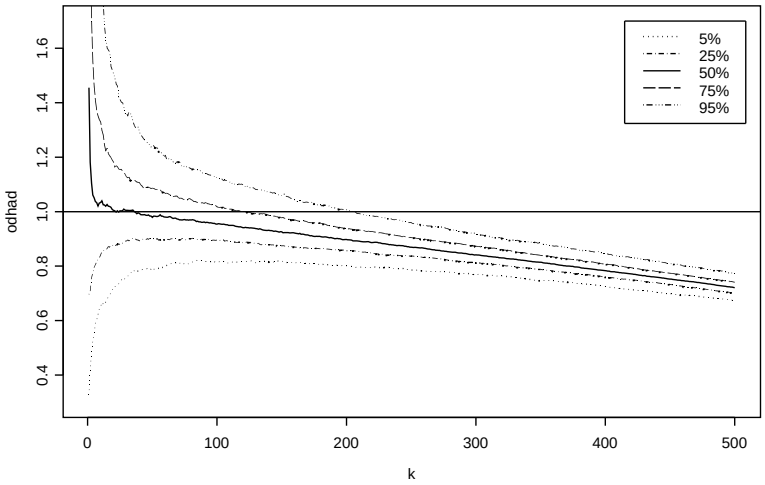
$$F(x) = 1 - \left(\frac{1}{1 + x^{1/\gamma}}\right)^{\alpha}, \quad x \geq 0 \tag{16}$$

Konkrétně byla simulace provedena pro Burrovo rozdělení s $\gamma = 1,\ \alpha = 1$, jednička je opět hodnota, kterou chceme odhadnout. Výsledek můžeme vidět

na obr. 7, z kterého vyplývá, že lepší výsledek dostaneme, pokud je δ blízké 0.5. Neplatí to obecně, pro jiná rozdělení to může dopadne úplně opačně. Na druhou stranu volba δ není tak problematická jako volba k u Hillova odhadu, viz obr. 8, kde na stejných datech je spočítán Hillův odhad.



Obrázek 7: Závislost odhadu v 1000 simulovaných výběrech Burrova rozdělení s $\gamma = 1,\ \alpha = 1$ na parametru δ. Uvedeny jsou medián 5, 25, 75 a 95 percentily.



Obrázek 8: Hillův odhad v 1000 simulovaných výběrech o rozsahu 1000 v závislosti na k pro Burrovo rozdělení s $\gamma = 1,\ \alpha = 1$. Uvedeny jsou medián 5, 25, 75 a 95 percentily.

Jurečková a Picek ukázali v [12], že odhad je silně konzistentní. Asymptotickou normalitu odvodil Omelka [15]. Odhad (16) potřebuje pouze volbu δ, což ho činí použitelnějším než odhad (14). I simulace dávají poměrně dobré výsledky - viz dále, přesto však musíme být v praktických aplikacích velmi opatrní. Oba odhady nejsou invariatní vzhledem ke změně měřítka na rozdíl od odhadu Hillova (8), Pickandsonova (11), momentového (10) i (12). Všechny zmíněné odhady nejsou invaritní vzhledem ke změně polohy. Při mechanickém použití odhadů to potom může vést k „zajímavým“ výsledkům. Byly proto uvažovány některé modifikace Hillova odhadu, viz např. [3]. Nějaké poznámky, jak se s naznačeným problémem vypořádat pro odhad (16) učinil Omelka [15].

3.1 Numerická ilustrace

V této části zkusíme porovnat zmiňované odhady na simulovaných datech. Jako výchozí model použijeme dříve zmiňovaná rozdělení: Paretovo, Burrovo a zobecněné Paretovo. U všech tří rozdělení zvolíme shape parametr ($\gamma = 1/m = 1$) tak, aby chvosty byly „stejně těžké“ a mohli tak sledovat vliv rozdělení. U zobecněného Paretova zvolíme ještě další dvě hodnoty γ: $1/3$ – lehčí a 2 – těžší chvost.

Z daného rozdělení jsme vygenerovali 1000 × výběr o rozsahu 1000 a provedli výše zmíněné odhady. Odhady (8), (11), (10) a (12) jsme spočítali pro $k = 2, \ldots, 998$. Pro odhady (14) a (16) jsme provedli rozdělení do 200 skupin ($= N$) po 5 hodnotách ($= n$) a spočítali odhad pro $\delta = 0.01, \ldots, 0.50$ s krokem 0.01, navíc pro (14) za m_0 jsme zvolili „skutečné“ $1/\gamma + 1$. Za účelem porovnání jsme pro každé k, respektive δ spočetli střední kvadratickou chybu (MSE) a vybrali takovou hodnotu k (δ), kdy je MSE minimální a spočítali nějaké výběrové charakteristiky z tisíce získaných hodnot odhadů. Výsledky najdeme v tabulce 1. Odhad (14) je v ní označen jako FJP a (16) jako JP. Tučně je zvýrazněna pro dané rozdělení minimální MSE mezi odhady.

Můžeme si všimnout, že pro opravdu těžké chvosty, tj. pro všechny případy kromě zobecněného Paretova rozdělení s $\gamma = 1/3$, dávají všechny odhady v průměru rozumné výsledky. Nejslabší se přesto zdá být odhad (14) a protože byly už některé výhrady diskutovány dříve, nelze ho doporučit pro praktické úlohy. Naopak odhad (16) je srovnatelný s klasickými, navíc pro lehčí chvosty dává často rozumnější výsledky než klasické odhady. Zdá se tedy, že s ním lze pracovat minimálně jako vhodnou alternativou.

Z tabulky je dále vidět a další simulace pro jiné případy a rozdělení to jen potvrzují, že index lehčích chvostů se odhaduje mnohem hůře. Překvapující je výsledek Pickandsova odhadu (alespoň pro autora tohoto příspěvku), protože tento odhad by měl fungovat pro odhad nejen ve Fréchetově třídě, ale i pro Gumbelovu a Weibullovu sféru přitažlivosti, tedy i pro „lehké“ chvosty.

rozdělení	metoda	$k,\ \delta$	MSE	průměr	medián	rozptyl
Pareto	Hill	$k = 998$	**0.0010**	1.0003	0.9984	0.0010
$\gamma = 1$	Moment	$k = 998$	0.0023	1.0053	1.0033	0.0022
	Pickands	$k = 985$	0.0221	1.0177	0.9967	0.0218
	Gomes	$k = 997$	0.0044	1.0016	0.9968	0.0044
	FJP	$\delta = 0.15$	0.0123	0.9542	0.9371	0.0102
	JP	$\delta = 0.49$	0.0147	1.0435	1.0427	0.0128
Burr	Hill	$k = 112$	0.0098	0.9517	0.9489	0.0075
$\alpha = 1$	Moment	$k = 257$	0.0101	0.9478	0.9383	0.0074
$\gamma = 1$	Pickands	$k = 985$	0.0221	1.0177	0.9967	0.0218
	Gomes	$k = 998$	**0.0012**	1.0007	0.9989	0.0012
	FJP	$\delta = 0.22$	0.0111	0.9574	0.9402	0.0093
	JP	$\delta = 0.49$	0.0047	1.0226	1.0192	0.0042
zobec.	Hill	$k = 310$	**0.0010**	0.4847	0.4841	0.0007
Pareto	Moment	$k = 367$	0.0010	0.4880	0.4863	0.0009
$\gamma = 2$	Pickands	$k = 993$	0.0020	0.5030	0.4997	0.0020
β=1	Gomes	$k = 482$	0.0025	0.5227	0.5210	0.0020
	FJP	$\delta = 0.01$	0.0084	0.4177	0.4123	0.0016
	JP	$\delta = 0.45$	0.0042	0.5519	0.5491	0.0015
zobec.	Hill	$k = 112$	0.0098	0.9517	0.9489	0.0075
Pareto	Moment	$k = 257$	0.0101	0.9478	0.9383	0.0074
$\gamma = 1$	Pickands	$k = 985$	0.0221	1.0177	0.9967	0.0218
$\beta = 1$	Gomes	$k = 998$	**0.0012**	1.0007	0.9989	0.0012
	FJP	$\delta = 0.22$	0.0111	0.9574	0.9402	0.0093
	JP	$\delta = 0.49$	0.0047	1.0226	1.0192	0.0042
zobec.	Hill	$k = 23$	0.5527	2.4329	2.3598	0.2314
Pareto	Moment	$k = 257$	0.5037	2.5140	2.4248	0.2678
$\gamma = 1/3$	Pickands	$k = 890$	16.1112	3.6237	3.0364	15.7379
$\beta = 1$	Gomes	$k = 102$	0.4795	2.4276	2.4020	0.1520
	FJP	$\delta = 0.01$	0.2869	2.5618	2.5565	0.0949
	JP	$\delta = 0.11$	**0.1166**	2.8500	2.8384	0.0942

Tabulka 1: Výběrové charakteristiky odhadů Paretova indexu pro minimální MSE při 1000 opakování generování dat rozsahu 1000 pro různá rozdělení.

Reference

[1] Dekkers A.L.M, Einmahl J.H.J., de Haan L. (1989). *A moment estimator for the index of an extreme value distribution.* Ann. Statist. **17**, 1833-1855.

[2] Embrechts P., Klüppelberg C., Mikosch T. (1997). *Modelling extremal events for insurance and finance.* Springer-Verlag, Berlin.

[3] Fraga Alves M.I. (2001). *A location invariant Hill-type estimator.* Extremes, **4** (2), 165 – 183.

[4] Fialová A., Jurečková J., Picek J. (2004). *Estimation of tail index based on sample mean.* Revstat, **2**, 75 – 99.

[5] de Haan L. (1970). *On regular variation and its application to the weak convergence of sample extremes.* Mathematical Centre Tract 32, Amsterdam.

[6] de Hann L., Stadtmüller U. (1996). *Generalized regular variation of second order.* J.Austral.Math.Soc. (A) **61**, 381 – 395.

[7] Hasofer A.M., Wang Z. (1992). *A test for extreme value domain of attraction.* JASA, **87**, 171 – 177.

[8] Hill B.M. (1975). *A simple general approach to inference about the tail of a distribution.* Ann. Statist. **3**, 1163 – 1174.

[9] Gomes M.I., Martins M.J. (2002). *Asymptotically unbiased estimators of the tail index based on the external estimation of the second order parameter.* Extremes **5** (1), 5 – 31.

[10] Gomez I., Oliviera O. (2001). *The bootstrap methodology in statistics of extremes-choice of the optimal sample fraction.* Extremes **4** (4), 331 – 358.

[11] Jurečková J., Picek J. (2001). *A class of tests on the tail index.* Extremes, **4**,(2), 165 – 183.

[12] Jurečková J., Picek J. (2004). *Estimates of the tail index based on nonparametric tests.* Theory and Applications of Recent Robust Methods, Birkhauser, Basel, 141 – 152.

[13] Mason D.M. (1982). *Laws of large numbers for sums of extreme values.* Ann. Probab. **10**, 754 – 764.

[14] Neves C., Picek J., Fraga Alves M. I. (2005). *The contribution of the maximum to the sum of excesses for testing max-domains of attractions.* J. Statist. Planning Infer., v tisku.

[15] Omelka M. (2005). *Asymptotic normality of the estimates of the tail index based on nonparametric tests.* Zasláno.

[16] Picek J., Jurečková J. (2001). *A class of tests on the tail index using the modified extreme regression quantiles.* Sborník konference ROBUST'00 (J.Antoch, G.Dohnal, eds.), JČMF Praha, 217 – 226.

[17] Pickands J. (1975). *Statistical inference using extreme order statistics.* Ann.Statist. **3**, 119 – 131.

[18] Segers J., Teugels J. (2001). *Testing the Gumbel hypothesis by Galton's ratio.* Extremes, **3:3**, 291 – 303.

Poděkování: Příspěvek vznikl za podpory Grantové agentury AV ČR – projekt B3042303 a výzkumného záměru MSM4674788501.

Adresa: J. Picek, Katedra aplikované matematiky, Technická univerzita v Liberci, Hálkova 6, 461 17 Liberec

E-mail: jan.picek@vslib.cz

MODIFIKACE WHITEOVA TESTU PRO NEJMENŠÍ VÁŽENÉ ČTVERCE

Pavel Plát

Klíčová slova: Robustní regrese, nejmenší vážené čtverce, Whiteův test, heteroskedasticita.

Abstrakt: Odhad regresních koeficientů v lineární regresním modelu metodou nejmenších vážených čtverců je odhadem $\sqrt{n}$-konzistentním a s asymptoticky normálním rozdělením. Znalost těchto vlastností nám umožňuje modifikovat myšlenku H. Whita a získat tak pro nejmenší vážené čtverce modifikaci Whiteova testu homoskedasticity disturbancí.

1 Lineární regresní model

Pro všechny $n \in N$ je lineární regresní model dán vztahem

$$Y_i = x_i^T \beta^0 + e_i, \qquad i = 1, 2, \ldots, n, \tag{1}$$

kde Y_i je vysvětlovaná proměnná, $x_i = (x_{i1}, \ldots, x_{ip})^T \in R^p$ jsou vysvětlující proměnné nebo též nosiče (uvažujeme model s pevnými, to jest deterministicky danými nosiči). $\beta^0 = (\beta_1^0, \ldots, \beta_p^0)^T \in R^p$ je "správný" vektor regresních koeficientů a e_i, $i = 1, 2, \ldots, n$ jsou disturbance, to znamená náhodné fluktuace Y_i od střední hodnoty $\mathsf{E}(Y_i)$. Poznamenejme, že formalismus, který jsme zavedli, je schopen zahrnout jak modely, kdy neuvažujeme intercept, tak modely s interceptem. Pokud totiž uvažujeme model s interceptem, stačí předpokládat, že první složka všech vektorů x_i, $i = 1, 2, \ldots, n$ je rovna 1. Pro všechny $\beta \in R^p$, $i = 1, 2, \ldots, n$ označme $r_i(\beta) = Y_i - x_i^T \beta$ i–té reziduum za předpokladu, že β je vektor regresních koeficientů. Konečně pro pořádkové statistiky druhých mocnin reziduí budeme používat označení $r^2_{(i)}(\beta)$, $i = 1, 2, \ldots, n$. To znamená, že pro všechny $\beta \in R^p$ platí

$$0 \le r^2_{(1)}(\beta) \le r^2_{(2)}(\beta) \le \cdots \le r^2_{(n)}(\beta).$$

2 Definice $\hat{\beta}^{(LWS,n,w)}$

Nejprve definujme váhovou funkci.

Definice 2.1. *Nechť funkce* $w : [0,1] \longrightarrow [0,1]$ *je nerostoucí a spojitá na* $[0,1]$, $w(0) = 1$ *a* $w(1) = 0$. *Dále nechť ve všech bodech intervalu* $(0,1)$ *existují obě jednostranné derivace funkce* w, *jsou stejně omezené a nechť v bodě* 0 *resp.* 1 *existuje konečná derivace zleva resp. zprava. Potom funkci* w *nazveme váhovou funkcí.*

Nyní již přistupme k definici odhadu metodou nejmenších vážených čtverců.

Definice 2.2. *Nechť $\mathcal{K} \subset R^p$ je kompaktní množina a platí $\beta^0 \in \mathcal{K}^o$. Dále nechť w je váhová funkce. Potom odhad regresních koeficientů daný vztahem*

$$\hat{\beta}^{(LWS,n,w)} = \underset{\beta \in \mathcal{K}}{\arg\min} \sum_{i=1}^{n} w\left(\frac{i-1}{n}\right) r^2_{(i)}(\beta) \tag{2}$$

nazveme odhad metodou nejmenších vážených čtverců.

Označíme-li

$$\tilde{w}_\ell^{(n,1)} = w\left(\frac{\ell-1}{n}\right) - w\left(\frac{\ell-1}{n}\right), \qquad \ell = 1, 2, \ldots, n,$$

můžeme definiční vztah (2) jednoduchou úpravou převést na tvar

$$\hat{\beta}^{(LWS,n,w)} = \underset{\beta \in \mathcal{K}}{\arg\min} \sum_{\ell=1}^{n} \tilde{w}_\ell^{(n)} \sum_{i=1}^{n} r_i^2(\beta) I\left\{r_i^2(\beta) \le r^2_{(\ell)}(\beta)\right\}.$$

Podobně označíme-li pro libovolné $s \in N$

$$\tilde{w}_\ell^{(n,s)} = w^s\left(\frac{\ell-1}{n}\right) - w^s\left(\frac{\ell}{n}\right), \qquad \ell = 1, 2, \ldots, n, \tag{3}$$

platí následující vztah

$$w^s\left(\frac{\ell-1}{n}\right) = \sum_{\ell=1}^{n} \tilde{w}_\ell^{(n,s)} I\left\{r_i^2(\beta) \le r^2_{(\ell)}(\beta)\right\}. \tag{4}$$

3 Základní předpoklady

Dříve než se budeme věnovat asymptotickým vlastnostem $\hat{\beta}^{(LWS,n,w)}$, shrňme předpoklady o náhodných disturbancích a nosičích, za kterých můžeme tyto vlastnosti dokázat. Nejprve označme $F(z)$ distribuční funkci náhodné veličiny e_1 a $f(z)$ její hustotu. Dále označme $Q_n = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T$.

Předpoklady $\mathcal{A}$

$\{e_i\}_{i=1}^{\infty}$ je posloupnost nezávislých, stejně rozdělených náhodných veličin. Pro rozdělení pravděpodobnosti náhodné veličiny e_1 platí:

- Její distribuční funkce $F(z)$ je absolutně spojitá.
- Hustota pravděpodobnosti $f(z)$ je symetrická, omezená a ostře klesající na R^+.
- Na celém R existuje $f'(z)$ a je v absolutní hodnotě omezená.
- $\mathsf{E}\left(e_1^4\right) = \kappa_4 \in R^+$.

$\{x_i\}_{i=1}^{\infty}$ je posloupnost pevných (nenáhodných) vektorů z R^p a platí:

- $\sum_{i=1}^{n} \|x_i\|^4 = \mathcal{O}(n)$.
- $\lim_{n\to\infty} Q_n = Q$, kde Q je regulární matice z $R^{p,p}$.

Na první pohled se může zdát, že požadavky na rozdělení pravděpodobnosti náhodných disturbancí jsou příliš omezující. Musíme si nicméně uvědomit, že u "klasických" nejmenších čtverců požadujeme normalitu disturbancí, tedy dokonce ještě silnější podmínku. A není-li tato podmínka splněna (a stačí pouze malá odchylka od normality - viz [1]), jsou nejmenší čtverce optimální pouze ve třídě lineárních odhadů. To jinými slovy znamená, že můžeme (a často to není obtížné) nalézt nelineární odhad, který je lepší než nejmenší čtverce.

4 Asymptotické vlastnosti $\hat{\beta}^{(LWS,n,w)}$

Označme G distribuční funkci náhodné veličiny e_1^2 a pro všechny $\alpha \in [0,1]$ označme u_α^2 horní α–kvantil rozdělení $G(z)$, tzn. $P(e_1^2 > u_\alpha^2) = 1-G(u_\alpha^2) = \alpha$. Dále definujme $\varsigma_z^2 = \int_{-u_z}^{u_z} z^2 \mathrm{d}F(z)$.

Věta 4.1. *Nechť platí Předpoklady $\mathcal{A}$. Dále nechť w je nějaká váhová funkce. Potom*

$$\sqrt{n}\left(\hat{\beta}^{(LWS,n,w)} - \beta^0\right) = \mathcal{O}_p(1)$$

a $\hat{\beta}^{(LWS,n,w)}$ má asymptoticky normální rozdělení s vektorem středních hodnot rovným β^0 a kovarianční maticí

$$V\left(\hat{\beta}^{(LWS,n,w)}, F\right) = -\frac{\int_0^1 \varsigma_{1-z}^2 \mathrm{d}w^2(z)}{\left(\int_0^1 \left(z - 2u_{1-z}f\left(u_{1-z}\right)\right)\mathrm{d}w(z)\right)^2} \cdot Q^{-1},$$

tzn.

$$\mathcal{L}\left(\sqrt{n}\left(\hat{\beta}^{(LWS,n,w)} - \beta^0\right)\right) \to \mathcal{N}\left(\ 0\ ,\ V\left(\hat{\beta}^{(LWS,n,w)}, F\right)\ \right) \text{ pro } n \to \infty.$$

Důkaz. [4]. □

Poznamenejme, že pro lineární regresní model s náhodnými nosiči můžeme obdobné výsledky nalézt v [3].

5 Modifikace Whiteova testu

Důkazy v této kapitole vyžadují o trochu přísnější požadavky na vysvětlující proměnné. Doplňme tedy předpoklady z oddílu 3.

Předpoklady $\mathcal{B}$

- Jsou splněny Předpoklady $\mathcal{A}$
- $\sup_{i,j\in\{1,2,\ldots,n\}} |x_{ij}| = \mathcal{O}(1)$.

Se zajímavou myšlenkou, jak testovat homoskedasticitu disturbancí v lineárním regresním modulu při použití klasických nejmenších čtverců přišel H. White (viz. [7]). Jeho nápad spočívá v tom, že porovnáme dva odhady matice $\sigma^2 Q$, konkrétně odhady $\frac{1}{n}\sum_{i=1}^n r_i^2(\hat\beta^{(LS,n)})Q_n$ a $\frac{1}{n}\sum_{i=1}^n r_i^2(\hat\beta^{(LS,n)})x_i^T x_i$. Podobný "trik" můžeme použít i v případě nejmenších vážených čtverců. V tomto případě půjde o dva různé odhady matice $\sigma_w^2 \cdot Q$, kde jsme označili

$$\sigma_w^2 = -\int_0^1 \varsigma_{1-z}^2 \mathrm{d}w^2(z).$$

Dále označme

$$\hat\sigma_{w,n}^2 = \frac{1}{n}\sum_{i=1}^n w^2\left(\frac{i-1}{n}\right) r_{(i)}^2\left(\hat\beta^{(LWS,n,w)}\right)$$

a

$$\hat V_1 = \hat\sigma_{w,n}^2 \cdot Q_n, \qquad \hat V_2 = \frac{1}{n}\sum_{i=1}^n w^2\left(\frac{k_i-1}{n}\right) r_i^2\left(\hat\beta^{(LWS,n,w)}\right)\cdot x_i x_i^T.$$

Pro odhady $\hat\sigma_{w,n}^2$, $\hat V_1$ a $\hat V_2$ platí následující tvrzení.

Lemma 5.1. *Nechť platí Předpoklady $\mathcal{B}$. Dále nechť w je nějaká váhová funkce. Potom*

$$\hat\sigma_{w,n}^2 \xrightarrow{P} \sigma_w^2 \qquad \textit{pro } n\to\infty.$$

Dále pak

$$\hat V_1 \xrightarrow{P} \sigma_w^2\cdot Q \qquad a \qquad \hat V_2 \xrightarrow{P} \sigma_w^2\cdot Q, \qquad \textit{pro } n\to\infty.$$

Důkaz. Použijeme-li vztah (4) snadno dostáváme

$$\hat\sigma_{w,n}^2 =$$

$$\frac{1}{n}\sum_{i=1}^n r_i^2\left(\hat\beta^{(LWS,n,w)}\right)\sum_{\ell=1}^n \tilde w_\ell^{(n,2)} I\left\{r_i^2\left(\hat\beta^{(LWS,n,w)}\right) \le r_{(\ell)}^2\left(\hat\beta^{(LWS,n,w)}\right)\right\}$$

Z $\sqrt{n}$-konvergenci $\hat\beta^{(LWS,n,w)}$ a Lemmat 3 a 10 z [4] dostáváme (podrobnější postup při úpravě obdobných výrazů viz. [4])

$$\hat\sigma_{w,n}^2 = \frac{1}{n}\sum_{i=1}^n e_i^2 \sum_{\ell=1}^n \tilde w_\ell^{(n,2)} I\left\{e_i^2 \le u_{1-\frac{\ell}{n}}^2\right\} + o_p(1). \tag{5}$$

Náhodné veličiny $e_i^2 \sum_{\ell=1}^n \tilde{w}_\ell^{(n,2)} I\left\{e_i^2 \le u^2_{1-\frac{\ell}{n}}\right\}$, $i = 1, 2, \ldots . n$ jsou nezávislé (viz. Lemma D.1), stejně rozdělené a platí

$$\mathsf{E}\left(e_i^2 \sum_{\ell=1}^n \tilde{w}_\ell^{(n,2)} I\left\{e_i^2 \le u^2_{1-\frac{\ell}{n}}\right\}\right) = \sum_{\ell=1}^n \tilde{w}_\ell^{(n,2)} \mathsf{E}\left(e_i^2 I\left\{e_i^2 \le u^2_{1-\frac{\ell}{n}}\right\}\right) =$$

$$= \sum_{\ell=1}^n \tilde{w}_\ell^{(n,2)} \varsigma^2_{(1-\frac{\ell}{n})} = -\int_0^1 \varsigma^2_{1-z} \mathrm{d}w^2(z) + o(1) \tag{6}$$

a

$$\mathsf{Var}\left(e_i^2 \sum_{\ell=1}^n \tilde{w}_\ell^{(n,2)} I\left\{e_i^2 \le u^2_{1-\frac{\ell}{n}}\right\}\right) \le \mathsf{E}\left(e_i^4 \sum_{\ell=1}^n \tilde{w}_\ell^{(n,4)} I\left\{e_i^2 \le u^2_{1-\frac{\ell}{n}}\right\}\right) \le$$

$$\le \mathsf{E}\left(e_i^4\right) = \kappa_4. \tag{7}$$

Ze vztahů (5) – (5) pak již vyplývá, že $\hat{\sigma}^2_{w,n}$ konverguje v pravděpodobnosti k σ^2_w.

Konvergence $\hat{V}_1$ v pravděpodobnosti k $\sigma^2_w \cdot Q$ je nyní přímým důsledkem dokázané konvergence $\hat{\sigma}^2_{w,n}$ a předpokladu $\lim\limits_{n\to\infty} Q_n = Q$.

Konvergence $\hat{V}_1$ v pravděpodobnosti k $\sigma^2_w \cdot Q$ se pak dokáže analogickým postupem jako konvergence $\hat{\sigma}^2_{w,n}$. □

Označme nyní

$$\psi_{is} = x_{ik} x_{il} \qquad s = 1, \ldots, p(p+1)/2;\ k = 1, \ldots, p;\ l = 1, \ldots, p$$

a

$$\psi_i \in R^{p(p+1)/2}$$

vektor se složkami ψ_{is}. Dále pak označme

$$\hat{B}_n = \frac{1}{n} \sum_{i=1}^n \left[w^2\left(\frac{k_i - 1}{n}\right) r_i^2\left(\hat{\beta}^{(LWS,n,w)}\right) - \hat{\sigma}^2_{w,n}\right]^2 \psi_i^T \psi_i$$

a

$$\hat{D}_n = \frac{1}{n} \sum_{i=1}^n \left[w^2\left(\frac{k_i - 1}{n}\right) r_i^2\left(\hat{\beta}^{(LWS,n,w)}\right) - \hat{\sigma}^2_{w,n}\right] \psi_i.$$

kde čísla k_i jsou pro všechna $i = 1, 2, \ldots . n$ definována vztahem

$$r_i^2\left(\hat{\beta}^{(LWS,n,w)}\right) = r^2_{(k_i)}\left(\hat{\beta}^{(LWS,n,w)}\right).$$

Věta 5.1. *Nechť platí Předpoklady* $\mathcal{B}$. *Dále nechť* w *je nějaká váhová funkce a* $\mathsf{E}\left(e_1^8\right) < +\infty$. *Potom*

$$\mathcal{L}\left(n \hat{D}_n^T \cdot \hat{B}_n^{-1} \cdot \hat{D}_n\right) \to \chi^2_{p(p+1)/2},$$

Důkaz. Označme

$$\hat{D}_n^0 = \frac{1}{n}\sum_{i=1}^{n}\left[e_i^2\sum_{\ell=1}^{n}\tilde{w}_\ell^{(n,2)}I\left\{e_i^2 \le u^2_{1-\frac{\ell}{n}}\right\}\right.$$

$$\left.-\frac{1}{n}\sum_{j=1}^{n}e_j^2\sum_{\ell=1}^{n}\tilde{w}_\ell^{(n,2)}I\left\{e_j^2 \le u^2_{1-\frac{\ell}{n}}\right\}\right]\psi_i$$

a

$$\hat{B}_n^0 = \frac{1}{n}\sum_{i=1}^{n}\left[e_i^2\sum_{\ell=1}^{n}\tilde{w}_\ell^{(n,2)}I\left\{e_i^2 \le u^2_{1-\frac{\ell}{n}}\right\}\right.$$

$$\left.-\frac{1}{n}\sum_{j=1}^{n}e_j^2\sum_{\ell=1}^{n}\tilde{w}_\ell^{(n,2)}I\left\{e_j^2 \le u^2_{1-\frac{\ell}{n}}\right\}\right]^2\psi_i^T\psi_i.$$

Z $\sqrt{n}$-konvergenci $\hat{\beta}^{(LWS,n,w)}$ a Lemmat 3 a 10 z [4] dostáváme (podrobnější postup při úpravě obdobných výrazů viz. [4])

$$\sqrt{n}\hat{D}_n = \sqrt{n}\hat{D}_n^0 + o_p(1). \tag{8}$$

Ukážeme, že $\sqrt{n}\hat{D}_n^0$ má asymptoticky $\frac{p(p+1)}{2}$ rozměrné normální rozdělení s nulovou střední hodnotou a kovarianční maticí $\mathsf{E}\left(\hat{B}_n^0\right)$. Stačí když ukážeme (viz. Lemmatu D.2), že pro libovolné $\eta \in R^{\frac{p(p+1)}{2}}$ má náhodná veličina $\sqrt{n}\eta^T\hat{D}_n^0$ asymptoticky normální rozdělení s nulovou střední hodnotou a rozptylem $\eta^T\hat{B}_n^0\eta^T$. Využijeme-li opět, že náhodné veličiny $e_i^2\sum_{\ell=1}^{n}\tilde{w}_\ell^{(n,2)}$ $I\left\{e_i^2 \le u^2_{1-\frac{\ell}{n}}\right\}$, $i = 1, 2, \ldots, n$ jsou nezávislé (Lemma D.1) a stejně rozdělené dostáváme

$$\mathsf{E}\left(\sqrt{n}\eta^T\hat{D}_n^0\right) = 0$$

a

$$\mathsf{Var}\left(\sqrt{n}\eta^T\hat{D}_n^0\right) = \mathsf{E}\left(\sqrt{n}\eta^T\hat{D}_n^0\right)^2 = \mathsf{E}\left(\eta^T\hat{B}_n^0\eta^T\right).$$

Uvědomíme-li si, že jednak pro všechny $\omega \in \Omega$ je $\sum_{\ell=1}^{n}\tilde{w}_\ell^{(n,4)}I\left\{e_i^2 \le u^2_{1-\frac{\ell}{n}}\right\}$ $\le$ 1, na druhou stranu ale díky vlastnostem váhové funkce w a skutečnosti, že $f(z) > 0$ pro všechny $z \in R$ také existuje $\gamma > 0$ tak, že pro všechny $n \in N$ je $\mathsf{E}\left(e_1^4\sum_{\ell=1}^{n}\tilde{w}_\ell^{(n,4)}I\left\{e_i^2 \le u^2_{1-\frac{\ell}{n}}\right\}\right) > \gamma$, můžeme pro náhodné veličiny

$$\left[e_i^2\sum_{\ell=1}^{n}\tilde{w}_\ell^{(n,2)}I\left\{e_i^2 \le u^2_{1-\frac{\ell}{n}}\right\} - \frac{1}{n}\sum_{j=1}^{n}e_j^2\sum_{\ell=1}^{n}\tilde{w}_\ell^{(n,2)}I\left\{e_j^2 \le u^2_{1-\frac{\ell}{n}}\right\}\right]\psi_i$$

ověřit platnost Feller-Linderbergovy podmínky a podle Centrální limitní věty (viz. např. [6]) tedy platí

$$\mathcal{L}\left(\frac{\sqrt{n}\eta^T\hat{D}_n^0}{\sqrt{\mathsf{E}\left(\eta^T\hat{B}_n^0\eta^T\right)}}\right) \longrightarrow \mathcal{N}(0,1) \tag{9}$$

Konečně opět z $\sqrt{n}$-konvergenci $\hat{\beta}^{(LWS,n,w)}$ a Lemmat 3 a 10 z [4] dostáváme (podrobnější postup při úpravě obdobných výrazů viz. [4])

$$\hat{B}_n = \hat{B}_n^0 + o_p(1) \tag{10}$$

a použijeme-li obdobný postup jako v důkazu Lemmatu 1 dostváme

$$\hat{B}_n^0 \xrightarrow{P} \mathsf{E}\left(\hat{B}_n^0\right) \qquad \textit{pro } n \to \infty. \tag{11}$$

Ze vztahů (8) – (11) již plyne tvrzení věty. □

6 Shrnutí

V oddíle 4 jsme viděli, že za určitých předpokladů (Předpoklady $\mathcal{A}$) je odhad metodou nejmenších vážených čtverců (LWS) $\sqrt{n}$-konzistentní, asymptoticky normální a můžeme odvodit jeho asymptotickou reprezentaci. Jestliže dále připomeneme, že jde o odhad s potenciálně vysokým bodem selháním (více viz. např. [3] nebo [4]), je patrné, že LWS jsou metodou, jejíž využití může být při regresní analýze dat přínosné. Chceme-li ji však korektně používat, nemůžeme se vyhnout ověřování základních předpokladů. V oddíle 5 jsme pro LWS odvodili modifikaci jednoho z testů heteroskedasticity disturbancí používaného pro "klasické" nejmenší čtverce (LS), totiž Whiteova testu. Podobně jako v případě "klasického" Whiteova testu můžeme očekávat, že námi odvozená modifikace nebude citlivá pouze na porušení homoskedasticity, ale i dalších předpokladů. Dále je třeba také poznamenat, že u LWS, jakožto robustního odhadu očekáváme určitou "odolnost" vůči porušení základních předpokladů. Právě v případě požadavku na homoskedasticitu distubancí se ukazuje, že LWS se s určitou mírou heteroskedasticity dokáží poměrně dobře vyrovnat. To je důsledek faktu, že v případě LWS minimalizujeme součet vážených reziduí a právě toto vážení nám může pomoci vliv heteroskedasticity disturbancí potlačit. Námi navržená modifikace Whiteova testu pak vlastně více než heteroskedasticitu "původních" disturbancí testuje heteroskedasticitu "zvážených" disturbancí. Na jednu stranu musíme být tedy opatrní při interpretaci výsledku tohoto testu vzhledem k "původním" disturbancím, na druhou stranu takto postavený test nám může podat relevantní informaci, pokud jde o vhodnost použití LWS.

Dodatek

Lemma D.1. *Nechť ζ_1 and ζ_2 jsou nezávislé náhodné veličiny a $u > 0$. Potom $\zeta_1 \cdot I\{|\zeta_1| < u\}$ a $\zeta_2 I\{|\zeta_2| < u\}$ jsou opět nezávislé náhodné veličiny.*

Důkaz Provedeme přímý výpočet. Nechť a_1 and a_2 jsou reálná čísla. Potom

$$\begin{aligned} &P(\zeta_1 \cdot I\{|\zeta_1| < u\} \le a_1, \zeta_2 \cdot I\{|\zeta_2| < u\} \le a_2) = \\ &= P(-u \le \zeta_1 \le \min\{a_1, u\}, -u \le \zeta_2 \le \min\{a_2, u\}) \\ &= P(-u \le \zeta_1 \cdot \le \min\{a_1, u\}) \cdot P(-u \le \zeta_2 \cdot \le \min\{a_2, u\}) = \\ &= P(\zeta_1 \cdot I\{|\zeta_1| < u\} \le a_1) \cdot P(\zeta_2 \cdot I\{|\zeta_2| < u\} \le a_2). \end{aligned}$$

□

Lemma D.2. *(Štěpán (1987), page 166, II.7.26) Náhodný vektor $(Y_1, \ldots, Y_n)$ má $n-$rozměrné normální rozdělení s kovarianční maticí Γ právě tehdy, když náhodná veličina $Y^T t$ má pro všechny $t \in R^n$ normální rozdělení s rozptylem $t^T \Gamma t$.*

Reference

[1] Fisher R.A. (1922): *On the mathematical foundations of theoretical statistics.* Philos. Trans. Roy. Soc. London Ser. A **222**, 309 – 368.

[2] Chatterjee S., Price B. (1977). *Regression analysis by example.* J. Wiley & Sons, New York.

[3] Mašíček L. (2003). *Diagnostika a senzitivita robustních modelů.* PhD disertace, MFF UK.

[4] Plát P. (2003). *Odhad metodou nejmenších vážených čtverců.* Diplomová práce, FJFI, ČVUT.

[5] Rousseeuw P.J., Leroy A.M. (1987). *Robust regression and outlier detection.* J.Wiley *&* Sons, New York.

[6] Štěpán J. (1987). *Teorie pravděpodobnosti (Probability Theory).* Prague: Academia.

[7] White H. (1980). *A heteroskesticity-consistent covariance matrix estimator and a direct test for heteroskedasticity.* Econometrica **48**, 817 – 838.

Poděkování: Výzkum podporován grantem GA ČR č. 402/03/0084.

Adresa: P. Plát, Katedra matematiky, FJFI, ČVUT

E-mail: `plat@kmlinux.fjfi.cvut.cz`

METODA BOOTSTRAP

Zuzana Prášková

Klíčová slova: Bootstrap, wild bootstrap, blokový bootstrap.

Abstrakt: V tomto článku jsou shrnuty základní vlastnosti metody bootstrap. Je popsán klasický přístup založený na nezávislých stejně rozdělených náhodných veličinách a také modifikace této metody pro časové řady.

1 Základní myšlenka metody bootstrap

Metoda bootstrap patří mezi tzv. intenzivní počítačové metody pro statistickou analýzu dat. První článek o bootstrapu [4] vyvolal velký ohlas a brzy po jeho zveřejnění byla publikována řada dalších teoretických i simulačních studií, které měly za cíl zkoumat použití, účinnost a spolehlivost této metody v nejrůznějších aplikacích.

V tomto článku uvedeme základní vlastnosti metody bootstrap a zmíníme se i o současných trendech.

Uvažujme nezávislé stejně rozdělené (*iid*) náhodné veličiny $X_1, \ldots, X_n$, jejichž distribuční funkce F není blíže specifikována. Nechť $\theta = \theta(F)$ je nějaká charakteristika rozdělení; je to pro nás neznámý parametr, který má být odhadnut na základě realizace náhodného výběru.

Nechť $T_n = T_n(X_1, \ldots, X_n)$ je statistika pro odhad parametru θ, nechť $R_n = R_n(X_1, \ldots, X_n)$ je její vhodně standardizovaná verze, např. $R_n = \sqrt{n}(T_n - \theta)$, nebo nějaká její funkce. Nechť

$$H_n(x) = P[R_n(X_1, \ldots, X_n, F) \leq x]$$

značí distribuční funkci statistiky R_n.

Explicitní odvození rozdělení H_n i výpočet číselných charakteristik mohou být v jednotlivých případech značně obtížné, či dokonce analyticky neproveditelné a to i tehdy, když je distribuční funkce F známá. V takovém případě (při známé distribuční funkci) lze postupovat metodou Monte Carlo, generovat dlouhou sérii nezávislých náhodných výběrů z rozdělení s danou distribuční funkcí (tj. mnohokrát uměle opakovat experiment), pro každé opakování spočítat hodnotu příslušné charakteristiky a její skutečné rozdělení aproximovat empirickým rozdělením získaným z řady takto uměle získaných hodnot.

Je-li skutečná distribuční funkce F neznámá, což je mnohem častější případ, je možné aproximovat H_n asymptotickým rozdělením, které lze odvodit na základě limitních vět teorie pravděpodobnosti. Přesnost takové aproximace však je ovlivněna a omezena počtem pozorování, která jsou skutečně k dispozici.

Metoda *bootstrap* nabízí řešení, které kombinuje tzv. *substituční princip* a *metodu Monte Carlo.*

Vysvětleme nejdříve substituční princip. Nechť $F_n(x)$ je nějaký odhad distribuční funkce. Nejčastěji se uvažuje empirická distribuční funkce založená na náhodném výběru $X_1, \ldots, X_n$, tj.

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I[X_i \leq x],$$

kde $I[A]$ značí indikátor množiny A. Při daných hodnotách $X_1, \ldots, X_n$ je F_n *známá* funkce.

Nechť $X_1^*, \ldots, X_n^*$ je nezávislý náhodný výběr z F_n, tj. při daných pozorováních $X_1, \ldots, X_n$ jsou $X_1^*, \ldots, X_n^*$ (podmíněně) nezávislé, stejně rozdělené náhodné veličiny, z nichž každá nabývá hodnot $X_1, \ldots, X_n$ s pravděpodobností $\frac{1}{n}$. Soubor $X_1^*, \ldots, X_n^*$ se nazývá *bootstrapový výběr.* V dalších úvahách původní výběr nahradíme bootstrapovým výběrem a neznámou distribuční funkci F známou distribuční funkcí F_n. Dostaneme parametr $\theta^* = \theta(F_n)$ a statistiky $T_n^* = T_n(X_1^*, \ldots, X_n^*)$ a $R_n^* = R_n(X_1^*, \ldots, X_n^*, F_n)$.

Nyní můžeme definovat charakteristiky jako

$$E^* T_n^* = \int T_n(x_1, \ldots, x_n) d(F_n(x_1) \ldots F_n(x_n)),$$

$$\mathrm{var}^* T_n^* = \int [T_n(x_1, \ldots, x_n) - E^* T_n^*]^2 d(F_n(x_1), \ldots F_n(x_n))$$

a distribuční funkci

$$\begin{aligned} H_n^*(x) &= P^*(R_n(X_1^*, \ldots, X_n^*, F_n) \leq x) \\ &= P(R_n(X_1^*, \ldots, X_n^*, F_n) \leq x | X_1, \ldots, X_n); \end{aligned}$$

jsou to tzv. *teoretické charakteristiky* a *teoretická distribuční funkce* získané metodou bootstrap.

Řekneme, že H_n^* je konsistentní odhad H_n, jestliže

$$\rho(H_n^*, H_n) \to 0 \text{ při } n \to \infty$$

v pravděpodobnosti (slabá konzistence) nebo skoro jistě (silná konzistence), kde ρ je nějaká metrika na prostoru distribučních funkcí. Nejčastěji používané metriky v tomto případě jsou *supremální metrika*

$$\rho_\infty(G, H) = \sup_{x \in \mathbb{R}} |G(x) - H(x)|$$

nebo tzv.*Mallowsova vzdálenost,* která je pro distribuční funkce G a H z rodiny distribučních funkcí s konečnými $r-$tými momenty definovaná předpisem

$$\widetilde{\rho}_r(G, H) = \inf_{\mathcal{T}_{X,Y}} (E|X - Y|^r)^{1/r},$$

kde $\mathcal{T}_{X,Y}$ je množina všech možných sdružených rozdělení vektorů (X, Y), jejichž marginální rozdělení jsou G a H. O vlastnostech Mallowsovy vzdálenosti a souvislosti s konvergencí v distribuci náhodných veličin viz např. [2].

Konzistence bootstrapových charakteristik se definuje přirozeným způsobem. Řekneme např., že $\operatorname{var}{}^*T_n^*$ je konzistentní odhad rozptylu $\operatorname{var} T_n$, jestliže

$$\operatorname{var}{}^*T_n^* / \operatorname{var} T_n \to 1 \text{ při } n \to \infty$$

buď v pravděpodobnosti nebo skoro jistě.

Pro praktické použití jsou teoretické bootstrapové charakteristiky vhodné jen v případě, že jsou explicitními funkcemi pozorování $X_1, \ldots, X_n$. Přesné stanovení bootstrapového rozdělení by vyžadovalo provedení všech n^n možných výběrů s vracením z populace pozorovaných hodnot $X_1, \ldots, X_n$. To je však uskutečnitelné jen pro výběry o malém rozsahu. I kdybychom se omezili jen na vzájemně různé výběry, máme takových výběrů stále ještě $\binom{2n-1}{n}$, což již pro $n = 10$ je hodnota 92 378.

Nejčastěji se proto na bootstrapový výběr $X_1^*, \ldots X_n^*$ a známou distribuční funkci F_n aplikuje metoda Monte Carlo, kdy se mnohokrát (B-krát) generuje nezávislý náhodný výběr z rozdělení F_n, při každém opakování se spočtou hodnoty T_n^*, R_n^* a z nich se stanoví aritmetický průměr. Dostaneme tak *bootstrapové odhady* původního rozdělení a původních charakteristik.

Např. bootstrapový odhad rozptylu T_n dostaneme tak, že opakujeme nezávislý náhodný výběr z rozdělení F_n celkem $B-$krát a spočteme vždy hodnotu statistiky T_n^*. Dostáváme tak hodnoty $T_{n,1}^*, \ldots, T_{n,B}^*$, ze kterých spočteme

$$\widehat{\operatorname{var}{}^*}T_n^* = \frac{1}{B}\sum_{b=1}^{B}\left(T_{n,b}^* - \frac{1}{B}\sum_{k=1}^{B}T_{n,k}^*\right)^2.$$

Podobně odhadneme distribuční funkci statistiky R_n jako

$$\widehat{H_n^*}(x) = \frac{1}{B}\sum_{b=1}^{B} I\{R_n(X_{1,b}^*, \ldots, X_{n,b}^*, F_n) \le x\},$$

kde $\{X_{1,b}^*, \ldots, X_{n,b}^*\}, b = 1, \ldots, B$, jsou nezávislé výběry z F_n. Pro některé účely se lépe hodí histogram pořízený z hodnot R_n^*.

1.0.1 Příklad. Nechť $X_1, \ldots, X_n$ je náhodný výběr z rozdělení se střední hodnotou μ a rozptylem σ^2. Přirozeným odhadem parametru $\theta = e^\mu$ je statistika $T_n = e^{\overline{X}_n}$, kde $\overline{X}_n$ je výběrový průměr. Zabývejme se odhadem směrodatné odchylky $s_n = \sqrt{\operatorname{var} T_n}$.

V případě, že $X_i \sim \mathcal{N}(\mu, \sigma^2)$, má statistika T_n logaritmicko-normální rozdělení s parametry μ a $\frac{\sigma^2}{n}$, tedy

$$s_n = \left[e^{2\mu+\frac{\sigma^2}{n}}\left(e^{\frac{\sigma^2}{n}} - 1\right)\right]^{\frac{1}{2}}. \tag{1}$$

n	s_n	$\widetilde{s}_n$	s_{boot}
50	9,760	9,798	10,492
200	4,407	4,384	4,484
500	2,731	2,736	2,748

Tabulka 1: Porovnání odhadů směrodatné odchylky $T_n = e^{\bar{X}_n}$; s_n, $\widetilde{s}_n$, resp. s_{boot} značí skutečnou, simulovanou, resp. bootstrapovou směrodatnou odchylku.

V tabulce 1 jsou porovnány odhady skutečné hodnoty s_n ze vzorce (1) pro různé rozsahy náhodného výběru z normálního rozdělení $\mathcal{N}(3,9)$ jednak metodou Monte Carlo, tj. opakováním náhodného výběru $X_1,\dots,X_n$, jednak metodou bootstrap. Hodnota $\widetilde{s}_n$ je hodnota spočtená metodou Monte Carlo z 10 000 opakování náhodného výběru $\mathcal{N}(3,9)$ o rozsahu n. Hodnota s_{boot} značí průměrnou hodnotu bootstrapového odhadu založeného na $B = 500$ realizacích bootstrapového výběru o rozsahu n spočtenou pro 10 000 simulačních experimentů.

Dále se zabývejme odhadem distribuční funkce statistiky

$$R_n = \sqrt{n}(T_n - \theta) = \sqrt{n}(e^{\overline{X}_n} - e^{\mu}). \tag{2}$$

Předpokládáme-li stále, že $X_i \sim \mathcal{N}(\mu,\sigma^2)$, zjistíme snadno, že R_n má distribuční funkci

$$H_n(x) = \Phi\Big(\big(\ln\big(\frac{x}{\sqrt{n}} + e^{\mu}\big) - \mu\big)\frac{\sqrt{n}}{\sigma}\Big) \tag{3}$$

a hustotu

$$h_n(x) = \frac{\sqrt{n}}{\sigma(x + e^{\mu}\sqrt{n})}\varphi\Big(\big(\ln\big(\frac{x}{\sqrt{n}} + e^{\mu}\big) - \mu\big)\frac{\sqrt{n}}{\sigma}\Big), \tag{4}$$

kde Φ a φ značí distribuční funkci a hustotu $\mathcal{N}(0,1)$.

Pokud bychom rozdělení náhodných veličin $X_1,\dots,X_n$ neznali, mohli bychom se pokusit nalézt asymptotické rozdělení. Z Taylorova rozvoje dostaneme, že

$$R_n = \sqrt{n}(e^{\overline{X}_n} - e^{\mu}) = \sqrt{n}(\overline{X}_n - \mu)e^{\mu} + o_p(1),$$

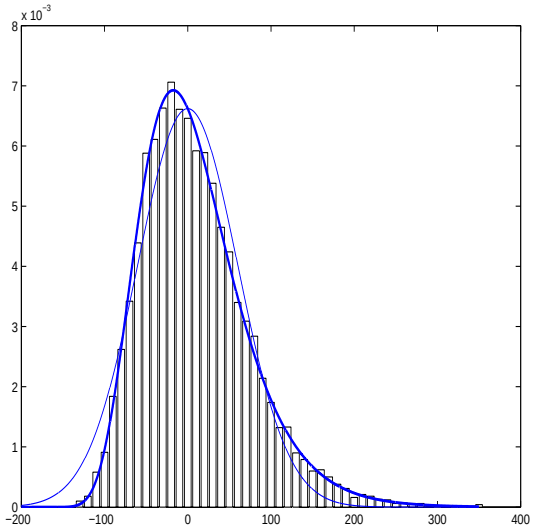
tedy R_n má asymptoticky normální rozdělení s nulovou střední hodnotou a rozptylem $e^{2\mu}\sigma^2$.

Další z možností je použít bootstrap. Nechť $X_1^*,\dots,X_n^*$ je bootstrapový výběr pořízený z pozorování $X_1,\dots,X_n$. Potom $X_1^*\dots,X_n^*$ jsou *iid,* pro které platí

$$E^*X_1^* = \mu^* = \frac{1}{n}\sum_{j=1}^{n} X_j = \overline{X_n}, \quad \mathrm{var}^*X_1^* = \sigma^{2\,*} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2,$$

a bootstrapová statistika je

$$R_n^* = \sqrt{n}(T_n^* - \theta^*) = \sqrt{n}(e^{\overline{X_n^*}} - e^{\mu^*}) = \sqrt{n}(e^{\overline{X_n^*}} - e^{\overline{X_n}}). \tag{5}$$



Obrázek 1: Hustota statistiky $\sqrt{n}(e^{\overline{X}_n} - e^{\mu})$: skutečná (tučná čára), asymptotická (slabá čára) a bootstrapová (histogram).

Na obrázku 1 je znázorněna hustota skutečného rozdělení statistiky (2) spočtená podle (4) pro náhodný výběr o rozsahu $n = 100$ z normálního rozdělení $\mathcal{N}(3, 9)$ (silnou čarou), hustota asymptotického normálního rozdělení (slabou čarou) a histogram odpovídající bootstrapové statistiky (5) pro $B = 10\,000$ bootstrapových výběrů.

Úvahy, které jsme dosud provedli, se dají zobecnit na vícerozměrný případ. Jsou-li $\mathbf{X}_1, \ldots, \mathbf{X}_n$ nezávislé stejně rozdělené náhodné vektory s distribuční funkcí F, lze spočítat empirickou distribuční funkci a definovat bootstrapový výběr $\mathbf{X}_1^*, \ldots, \mathbf{X}_n^*$ jako nezávislý náhodný výběr z rozdělení s touto empirickou distribuční funkcí. Bootstrapové odhady lze potom definovat zcela analogicky jako v jednorozměrném případě.

2 Vlastnosti bootstrapových aproximací

2.1 Přesnost aproximace rozdělení

Teoretické výsledky zkoumající přesnost bootstrapové aproximace rozdělení jsou založeny na centrálních limitních větách a jejich zpřesněních pomocí Berryovy- Esséenovy nerovnosti a Edgeworthova rozvoje pro normované, případně studentizované statistiky. Základní výsledky lze nalézt v pracech [16], [2], [1], [7], jejich shrnutí např. v [15].

Uveďme nejprve základní výsledky týkající se výběrového průměru. Uvažujme distribuční funkce výběrových statistik a jejich bootstrapové verze

$$H_n(x) = P(\sqrt{n}(\overline{X}_n - \mu) \le x), \quad H_n^*(x) = P^*(\sqrt{n}(\overline{X}_n^* - \mu^*) \le x),$$

$$\widetilde{H}_n(x) = P\Big(\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \le x\Big), \quad \widetilde{H}_n^*(x) = P^*\Big(\sqrt{n}\frac{\overline{X}_n^* - \mu^*}{\sigma^*} \le x\Big),$$

kde μ, σ^2 jsou střední hodnota a rozptyl a $\mu^* = \overline{X}_n$ a $\sigma^{2\,*} = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2$ jsou příslušné bootstrapové protějšky. Potom lze zformulovat následující velmi důležité tvrzení.

2.1.1 Věta Nechť $X_1, \dots, X_n$ jsou nezávislé stejně rozdělené náhodné veličiny s distribuční funkcí F.

(i) Jestliže $EX_1^2 < \infty$, potom

$$\rho_\infty(H_n^*, H_n) \xrightarrow[n\to\infty]{\text{s. j.}} 0.$$

(ii) Jestliže $EX_1^4 < \infty$, potom

$$\overline{\lim_{n\to\infty}} \frac{\sqrt{n}\rho_\infty(H_n^*, H_n)}{\sqrt{\log\log n}} = \frac{\sqrt{\mathrm{var}(X_1 - \mu)^2}}{2\sigma^2\sqrt{2\pi e}} \quad \text{s. j.}$$

(iii) Jestliže $E|X_1|^3 < \infty$ a F je řešetovitá, tj. existují konstanty c, h takové, že $\sum_{k=-\infty}^{\infty} P(X_1 = c + kh) = 1$, potom

$$\overline{\lim_{n\to\infty}} \sqrt{n}\rho_\infty(\widetilde{H}_n^*, \widetilde{H}_n) = \frac{h}{\sqrt{2\pi}\sigma} \quad \text{s. j.}$$

(iv) Jestliže $E|X_1|^3 < \infty$ a F není řešetovitá, potom

$$\sqrt{n}\rho_\infty(\widetilde{H}_n^*, \widetilde{H_n}) \xrightarrow[n\to\infty]{\text{s. j.}} 0.$$

Důkaz. Viz [16].

Tvrzení (i) je důkaz konzistence pro nestandardizované statistiky, tvrzení (ii) udává rychlost této konvergence. Tvrzení (iii) a (iv) udávají rychlost konvergence bootstrapové aproximace rozdělení standardizovaného výběrového průměru. Porovnejme ji s rychlostí aproximace normálním rozdělením. Ta je pro řešetovitá i neřešetovitá rozdělení podle Berryovy-Esséenovy nerovnosti řádu $O(n^{-\frac{1}{2}})$ a nedá se zlepšit. Pro neřešetovitá rozdělení je tudíž podle (iv) bootstrapová aproximace lepší.

Dále uveďme podobné výsledky pro hladké funkce výběrového průměru. Lze totiž ukázat, že mnoho výběrových statistik je možno přepsat jako funkci výběrového průměru nějakých náhodných vektorů.

Mějme nezávislé stejně rozdělené p-rozměrné náhodné vektory $\mathbf{X}_1, \ldots, \mathbf{X}_n$ s distribuční funkcí F, se střední hodnotou $\boldsymbol{\mu}$ a varianční maticí $\boldsymbol{\Sigma}$. Dále uvažujme funkci g z $\mathbb{R}^p$ do $\mathbb{R}$, $\nabla g(\mathbf{x}) := l(\mathbf{x})$ nechť je gradient g spočtený v bodě $\mathbf{x}$. Označme

$$s^2 = l^T(\boldsymbol{\mu})\boldsymbol{\Sigma}\, l(\boldsymbol{\mu}), \qquad s^{2*} = l^T(\boldsymbol{\mu}^*)\boldsymbol{\Sigma}^* l(\boldsymbol{\mu}^*),$$

$$S_n^2 = l^T(\overline{\boldsymbol{X}}_n)\boldsymbol{\Sigma}_n l(\overline{\boldsymbol{X}}_n), \quad S_n^{2*} = l^T(\overline{\boldsymbol{X}}_n{}^*)\boldsymbol{\Sigma}_n^* l(\overline{\boldsymbol{X}}_n{}^*),$$

kde

$$\boldsymbol{\mu}^* = \overline{\boldsymbol{X}}_n, \quad \boldsymbol{\Sigma}^* = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i - \overline{\boldsymbol{X}}_n)(\boldsymbol{X}_i - \overline{\boldsymbol{X}}_n)^T = \boldsymbol{\Sigma}_n.$$

Nyní uvažujme distribuční funkce statistik

$$H_n(x) = P(\sqrt{n}(g(\overline{\boldsymbol{X}}_n) - g(\boldsymbol{\mu})) \le x),$$
$$H_n^*(x) = P^*(\sqrt{n}(g(\overline{\boldsymbol{X}}_n^*) - g(\boldsymbol{\mu}^*)) \le x),$$

$$\widetilde{H}_n(x) = P\Big(\sqrt{n}\frac{g(\overline{\boldsymbol{X}}_n) - g(\boldsymbol{\mu})}{s} \le x\Big), \quad \widetilde{H}_n^*(x) = P^*\Big(\sqrt{n}\tfrac{g(\overline{\boldsymbol{X}}_n^*) - g(\boldsymbol{\mu}^*)}{s^*} \le x\Big),$$
$$\widehat{H}_n(x) = P\Big(\sqrt{n}\frac{g(\overline{\boldsymbol{X}}_n) - g(\boldsymbol{\mu})}{S_n} \le x\Big), \quad \widehat{H}_n^*(x) = P^*\Big(\sqrt{n}\tfrac{g(\overline{\boldsymbol{X}}_n^*) - g(\boldsymbol{\mu}^*)}{S_n^*} \le x\Big).$$

2.1.2 Věta Nechť $\mathbf{X}_1, \ldots, \mathbf{X}_n$ jsou *iid* p-rozměrné vektory se střední hodnotou $\boldsymbol{\mu}$ a varianční maticí $\boldsymbol{\Sigma}$. Nechť g je funkce z $\mathbb{R}^p$ do $\mathbb{R}$.

(i) Nechť $\mathrm{E}\|\mathbf{X}_1\|^2 < \infty$, nechť g je spojitě diferencovatelná v $\boldsymbol{\mu}$ a $l(\boldsymbol{\mu}) \neq \mathbf{0}$. Potom
$$\rho_\infty(H_n^*, H_n) \xrightarrow[n\to\infty]{\text{s. j.}} 0.$$

(ii) Nechť $\mathrm{E}\|\mathbf{X}_1\|^3 < \infty$, nechť pro charakteristickou funkci $\phi(\boldsymbol{t})$ náhodného vektoru $\boldsymbol{X}_1$ platí $|\phi(\boldsymbol{t})| < 1$ pro $\boldsymbol{t} \neq \mathbf{0}$. Nechť g je třikrát spojitě diferencovatelná na okolí $\boldsymbol{\mu}$ a $l(\boldsymbol{\mu}) \neq 0$. Potom pro distribuční funkce standardizovaných statistik platí
$$\sqrt{n}\rho_\infty(\widetilde{H}_n^*, \widetilde{H}_n) \xrightarrow[n\to\infty]{\text{s. j.}} 0.$$

Pro distribuční funkce studentizovaných statistik platí
$$\sqrt{n}\rho_\infty(\widehat{H}_n^*, \widehat{H}_n) \xrightarrow[n\to\infty]{\text{s. j.}} 0.$$

Důkaz. Viz [1].

2.2 Redukce vychýlení odhadu bootstrapem

Nechť $X_1, \ldots, X_n$ je náhodný výběr z rozdělení s konečnou střední hodnotou μ a rozptylem σ^2. Nechť g je spojitá funkce, taková, že $E\,|g(\overline{X}_n)| < \infty$; uvažujme parametr $\theta = g(\mu)$. Víme, že výběrový průměr $\overline{X}_n$ je nestranný a konzistentní odhad parametru μ, tj. $E\,\overline{X}_n = \mu$ a $\overline{X}_n \to \mu$ skoro jistě. Potom $g(\overline{X}_n) \to g(\mu)$ skoro jistě, tj. $g(\overline{X}_n)$ je konzistentní odhad $g(\mu)$, ale obecně je vychýlený, neboť $Eg(\overline{X}_n) \neq g(\mu)$, pokud g není lineární.

Studujme velikost vychýlení $b_n = Eg(\overline{X}_n) - g(\mu)$. Nadále předpokládejme, že g je dostatečně hladká funkce a X_i mají konečné momenty takového řádu, že platí Taylorův rozvoj

$$g(\overline{X}_n) - g(\mu) = (\overline{X}_n - \mu)g'(\mu) + \frac{1}{2}(\overline{X}_n - \mu)^2 g''(\mu) + R_n, \tag{6}$$

kde $ER_n = O(n^{-2})$ (viz např. [5, kap. 5.4]). Označíme-li

$$B_n = \frac{1}{2}\frac{\sigma^2}{n} g''(\mu),$$

potom

$$b_n = E(g(\overline{X}_n) - g(\mu)) = \frac{1}{2}\frac{\sigma^2}{n} g''(\mu) + O(n^{-2}) = B_n + O(n^{-2}) = O(n^{-1}).$$

Nyní uvažujme bootstrap. Je-li $X_1^*, \ldots, X_n^*$ bootstrapový výběr, potom bootstrapová verze (6) je

$$g(\overline{X}_n^*) - g(\overline{X}_n) = (\overline{X}_n^* - \overline{X}_n)g'(\overline{X}_n) + \frac{1}{2}(\overline{X}_n^* - \overline{X}_n)^2 g''(\overline{X}_n) + R_n^*,$$

kde díky silnému zákonu velkých čísel platí $E^* R_n^* = O(n^{-2})$ skoro jistě. Pro bootstrapové vychýlení $b_n^* = E^* g(\overline{X}_n^*) - g(\overline{X}_n)$ tak máme

$$E^*(g(\overline{X}_n^*) - g(\overline{X}_n)) = \frac{1}{2}\frac{\sigma^{*2}}{n} g''(\overline{X}_n) + O(n^{-2}) \quad \text{s. j.}$$

Z dalšího Taylorova rozvoje dostaneme $g''(\overline{X}_n) = g''(\mu) + \frac{1}{2}(\overline{X}_n - \mu)g'''(\mu) + \widetilde{R}_n$ a odtud spočteme

$$Eb_n^* = E\Big(\frac{1}{2n^2}\sum_{j=1}^{n}(X_i - \overline{X}_n)^2 g''(\mu)\Big) + O(n^{-2}) = B_n + O(n^{-2}).$$

Uvažujme nyní místo odhadu $g(\overline{X}_n)$ jeho opravu $g(\overline{X}_n) - b_n^*$. Vychýlení tohoto opraveného odhadu je

$$\begin{aligned} E[g(\overline{X}_n) - b_n^*] - g(\mu) &= b_n - Eb_n^* \\ &= B_n + O(n^{-2}) - B_n + O(n^{-2}) = O(n^{-2}), \end{aligned}$$

což je ve srovnání s b_n řádově lepší výsledek.

odhad	B_n	redukce (%)	$rmse$
$\widehat{\theta}_{20}$	5,186	25,821	19,675
$\widehat{\theta}^c_{20}$	-1,011	5,034	14,712
$\widehat{\theta}_{50}$	1,863	9,275	9,941
$\widehat{\theta}^c_{50}$	-0,145	0,721	8,893
$\widehat{\theta}_{100}$	0,940	4,682	6,514
$\widehat{\theta}^c_{100}$	-0,041	0,207	6,195

Tabulka 2: Redukce vychýlení odhadu $T_n = e^{\bar{X}_n}$ metodou bootstrap.

V tabulce 2 jsou uvedeny výsledky simulační studie, která srovnává vychýlení $b_n = E\widehat{\theta}_n - \theta$, kde $\theta = e^{\mu}, \widehat{\theta}_n = e^{\overline{X}_n}$, a vychýlení opraveného odhadu $\widehat{\theta}^c_n = \widehat{\theta}_n - b^*_n$, kde $b^*_n = E^*\widehat{\theta}^*_n - \widehat{\theta}_n$, v závislosti na rozsahu náhodného výběru. Náhodné výběry byly generovány z normálního rozdělení $\mathcal{N}(3, 9)$, skutečná hodnota $\theta = 20,086$. Pro každý výběr o rozsahu n bylo generováno 500 bootstrapových výběrů. Ve sloupci B_n je uveden vždy rozdíl odhadnuté a skutečné hodnoty, ve sloupci *redukce* je podíl $\frac{|B_n|}{\theta}$, ve sloupci *rmse* odmocnina ze střední kvadratické chyby pro 10 000 simulačních experimentů (převzato z [18]).

2.3 Intervaly spolehlivosti

2.3.1 Studentizované intervaly spolehlivosti. Označme jako $\widehat{\theta}_n$ odhad parametru θ a uvažujme studentizovanou statistiku

$$R_n = \frac{\widehat{\theta}_n - \theta}{S_n}$$

a její bootstrapový protějšek

$$R^*_n = \frac{\widehat{\theta}^*_n - \theta_n}{S^*_n}.$$

Je-li H_n distribuční funkce R_n a γ_p je příslušný $p-$kvantil, potom interval spolehlivosti pro θ s koeficientem $1 - \alpha$ je

$$\big(\widehat{\theta}_n - \gamma_{1-\frac{\alpha}{2}} S_n,\ \widehat{\theta}_n - \gamma_{\frac{\alpha}{2}} S_n\big).$$

Bootstrapový interval spolehlivosti je

$$\big(\widehat{\theta}_n - \gamma^*_{1-\frac{\alpha}{2}}\, S_n,\ \widehat{\theta}_n - \gamma^*_{\frac{\alpha}{2}}\, S_n\big),$$

kde γ^*_p je $p-$kvantil distribuční funkce H^*_n statistiky R^*_n spočtený jako $\gamma^*_p = R^*_{([B\,p])}$, tj. výběrový kvantil spočtený z uspořádaného výběru $R^*_{(1)}, \ldots, R^*_{(B)}$.

skutečný	(5,4634; 30,4683)
asymptotický	(9,7289; 33,3495)
hybridní	(5,3061; 30,6092)
percentilový	(12,4691; 37,7722)

Tabulka 3: 95%−ní intervaly spolehlivosti parametru e^{μ} pro výběr o rozsahu 100 z $\mathcal{N}(3,9)$, použito 10 000 bootstrapových výběrů.

Lze ukázat [7, kap. 3 a 5], že takto sestrojené intervaly pokrývají neznámý parametr s přesností, která je lepší než klasické intervaly využívající asymptotickou normalitu statistiky R_n. Nevýhodou tohoto postupu je velký počet výpočetních operací. Bootstrapový odhad S_n^* totiž většinou hledáme metodou Monte Carlo; k tomu potřebujeme B_1 bootstrapových výběrů. Dalších B_2 výběrů potřebujeme pro odhad kvantilů, tedy celkem $B_1 \cdot B_2$ výběrů. Je-li $B_1 = 200$ a $B_2 = 1000$, budeme potřebovat 100 000 bootstrapových výběrů.

2.3.2 Percentilové intervaly. Tato metoda počítá intervalový odhad parametru θ pomocí kvantilů distribuční funkce nestandardizované statistiky $\widehat{\theta}_n^*$, tj. z distribuční funkce $G_n^*(x) = P^*(\widehat{\theta}_n^* \le x)$. Dostaneme intervalový odhad

$$\Big(G_n^{*-1}(\frac{\alpha}{2}),\ G_n^{*-1}\big(1-\frac{\alpha}{2}\big)\Big).$$

Tato metoda se hodí v případě, že neznámý parametr je kvantil distribuční funkce uvažovaného náhodného výběru.

2.3.3 Hybridní intervaly. Je-li $H_n(x) = P(\sqrt{n}(\widehat{\theta}_n - \theta) \le x)$, je interval spolehlivosti pro θ s koeficientem $1-\alpha$

$$\Big(\widehat{\theta}_n - c_{1-\frac{\alpha}{2}}\frac{1}{\sqrt{n}},\ \widehat{\theta}_n - c_{\frac{\alpha}{2}}\frac{1}{\sqrt{n}}\Big),$$

kde c_p je $p-$ kvantil distribuční funkce H_n.

Je-li $H_n(x) \approx H_n^*(x) = P^*(\sqrt{n}(\widehat{\theta}_n^* - \widehat{\theta}_n) \le x)$, můžeme místo neznámých kvantilů c_p uvažovat odpovídající kvantily c_p^* distribuční funkce H_n^* a potom dostaneme interval spolehlivosti

$$\Big(\widehat{\theta}_n - c^*_{1-\frac{\alpha}{2}}\frac{1}{\sqrt{n}},\ \widehat{\theta}_n - c^*_{\frac{\alpha}{2}}\frac{1}{\sqrt{n}}\Big).$$

V tabulce 3 jsou pro ilustraci uvedeny intervaly spolehlivosti pro parametr $\theta = \mathrm{e}^{\mu}$ z příkladu 1.0.1. Skutečné intervaly jsou založeny na kvantilech distribuční funkce statistiky $R_n = \sqrt{n}(e^{\overline{X}_n} - e^{\mu})$.

Více podrobností o konstrukci intervalových odhadů lze nalézt např. v knize [15].

2.4 Některé výpočetní aspekty

2.4.1 Volba počtu bootstrapových výběrů. Současné počítače jsou dostatečně rychlé, abychom ve většině případů mohli zvolit počet opakování bootstrapových výběrů B mnohem větší než rozsah n. Větší počet opakování však často nepřináší další podstatné zlepšení našich výsledků. Chyba aproximace metodou Monte Carlo, kterou používáme pro pořízení bootstrapových statistik, by však měla být zanedbatelná vzhledem k chybě teoretického bootstrapového odhadu.

Zabývejme se dvěma základními okruhy problémů, totiž odhady rozptylu a odhady distribuční funkce (podrobněji viz [15, kap. 5.4]). Označme

$$\widehat{\text{var}}^{\,*} T_n^* = \frac{1}{B}\sum_{b=1}^{B}\left(T_{n,b}^* - \frac{1}{B}\sum_{k=1}^{B} T_{n,k}^*\right)^2 := v_{boot}^B. \tag{7}$$

Lze ukázat, že pro bootstrapový koeficient variace pro $s_{boot}^B = \sqrt{v_{boot}^B}$ přibližně platí

$$c = CV^*(s_{boot}^B) = \frac{\sqrt{Var^* s_{boot}^B}}{E^* s_{boot}^B} \approx \frac{1}{\sqrt{2B}},$$

když se zanedbá bootstrapový koeficient šikmosti. Potom pro požadovanou míru variability c je $B = \frac{1}{2}c^{-2}$, např. pro $c = 5\%$ je $B = 200$. Ukazuje se, že pro varianční odhady typu (7) se přesnost příliš nezlepší velkým počtem opakování, tj. zvětšováním B. Obecně se pro odhady momentů doporučuje volit B mezi 200-600, podle jiných doporučení však stačí jen 50-200.

Distribuční funkci H_n statistiky R_n odhadujeme jako

$$\widehat{H_n^*}(x) = \frac{1}{B}\sum_{b=1}^{B} I\{R_n(X_{1,b}^*, \ldots, X_{n,b}^*, F_n) \le x\} := H_{boot}^B(x). \tag{8}$$

Podle [15] je asymptotická chyba metody Monte Carlo

$$\rho_\infty(H_{boot}^B, H_n^*) = \epsilon_n + \sqrt{B^{-1}\log\log B},$$

kde $\epsilon_n = \rho_\infty(H_n, H_n^*)$ je chyba bootstrapové aproximace. Má-li být chyba Monte Carlo zanedbatelná vzhledem k ϵ_n, měli bychom volit B tak, aby $B^{-1}\log\log B = o(\epsilon_n^2)$. Pokud $\epsilon_n = O_P(n^{-1})$, je možno volit $B = n^2 \log\log n$. Pro $n = 30$ tak dostaneme přibližně $B \approx 1\,100$ opakování. Obecně pro odhady kvantilů a distribuční funkce se doporučuje $B = 1000$ jako minimální hodnota.

2.4.2 Redukce počtu operací. Existuje celá řada postupů jak zefektivnit a urychlit bootstrapové výpočty. Zmiňme např. *rovnovážný bootstrap,* podle kterého lze B bootstrapových výběrů pořídit následujícím postupem.

Nejprve se každé pozorování zkopíruje právě $B-$krát; dostaneme posloupnost délky nB. Nyní se provede náhodná permutace na prvky této posloupnosti. Prvky s pořadím $1,\ldots,n$ této zpermutované posloupnosti tvoří bootstrapový výběr $X^*_{1,1},\ldots,X^*_{n,1}$, prvky s pořadím $n+1,\ldots,2n$ bootstrapový výběr $X^*_{1,2},\ldots,X^*_{n,2}$, atd, prvky s pořadím $(n-1)B+1,\ldots,nB$ bootstrapový výběr $X^*_{1,B},\ldots,X^*_{n,B}$.

Další užívané techniky Monte Carlo pro bootstrap, např. *importance resampling*, jsou popsány v [7] a [15].

2.4.3 Odlehlá pozorování Odlehlá pozorování mohou ovlivnit bootstrapové výpočty. Je-li mezi hodnotami $X_1,\ldots,X_n$ jedno odlehlé pozorování, potom pravděpodobnost, že nebude obsaženo v bootstrapovém výběru, je $(1-\frac{1}{n})^n$, což pro velká n bude přibližně e^{-1}, tj. asi 37%. Ukazuje se ale, že histogram hodnot bootstrapových průměrů $\overline{X}^*_n$ není příliš citlivý na přítomnost odlehlých pozorování. Naopak histogram statistiky $\overline{X}^*_n - \overline{X}^*_n(k)$, kde

$$\overline{X}^*_n(k) = \frac{1}{n-2k}\sum_{i=k+1}^{n-k} X^*_{(i)}$$

je bootstrapová verze k-useknutého výběrového průměru, je velmi citlivý na odlehlá pozorování a přítomnost takových pozorování se projeví v tom, že histogram statistiky $\overline{X}^*_n(k)$ nebude unimodální. Statistiku $\overline{X}^*_n - \overline{X}^*_n(k)$ lze užít k detekci odlehlých pozorování. Teoretická zdůvodnění lze nalézt např. v článku [17].

2.5 Meze použitelnosti metody bootstrap

V předchozích odstavcích jsme se pokusili vysvětlit některé přednosti metody bootstrap, zejména schopnost redukovat vychýlení odhadů a zpřesňovat rozdělení statistik a tedy i přesnost pokrytí neznámých parametrů konfidenčními intervaly pro pivotální, resp. studentizované statistiky. Od r. 1979, kdy vyšel základní článek o bootstrapu [4], byla teoreticky zkoumána použitelnost a konzistence metody bootstrap v mnoha nejrůznějších statistických úlohách. Obecné teoretické výsledky, za kterých je bootstrap konzistentní, jsou uvedeny např. v knize [11].

Zmiňme se aspoň o několika příkladech, které ukazují, že bootstrap nelze používat automaticky.

- Pro rozdělení s nekonečným rozptylem neodhaduje bootstrap konzistentně rozdělení hladkých funkcí výběrového průměru.
- Bootstrap neodhaduje konzistentně rozdělení odhadů parametrů, které leží na hranici parametrické množiny.
- Bootstrap neodhaduje konzistentně rozdělení extremálních odhadů.

Řadu dalších příkladů lze nalézt např. v [15]. Pro většinu nich lze dokázat, že metoda je konzistentní pro bootstrapové výběry rozsahu m, pro které $\frac{m}{n}\to 0$ při $m,n\to\infty$.

- Jestliže X_n nejsou nezávislé, bootstrapové statistiky nedávají konzistentní odhady parametrů a rozdělení. V tomto případě je třeba metodu bootstrap modifikovat tak, aby bootstrapový výběr odrážel závislostní strukturu.

3 Bootstrap pro závislá pozorování

Existuje celá řada modifikací metody bootstrap pro závislá pozorování, které se liší tím, jakým způsobem využívají informaci o procesu, kterým jsou generována data.

3.1 Bootstrap závislý na modelu (parametrický)

Předpokládejme, že proces, kterým jsou generována data, je specifikován až na nějaké parametry. Nejčastěji používanou bootstrapovou technikou je tzv. *reziduální bootstrap*. Tato varianta se hodí v případě, že data jsou identifikována jako parametrický model typu AR nebo ARMA, případně jako dynamický regresní model. Předpokládejme např., že náhodné veličiny $X_1, \ldots, X_n$ se řídí autoregresním modelem AR(p)

$$X_t = \beta_1 X_{t-1} + \cdots + \beta_p X_{t-p} + Y_t,\ t = 1, \ldots, n, \tag{9}$$

kde Y_t jsou *iid* s nulovou střední hodnotou a konečným kladným rozptylem σ^2, $X_0, \ldots, X_{1-p}$ jsou počáteční pozorování a jsou splněny předpoklady pro stacionaritu posloupnosti $\{X_t\}$. Bootstrap v klasické podobě zde použít nelze, neboť $X_1, \ldots, X_n$ nejsou nezávislé. Nezávislé chyby Y_t většinou neznáme, ale můžeme je odhadnout. Použijeme-li konzistentní odhady pro neznámé parametry $\beta_1, \ldots, \beta_p$, potom odhadnutá rezidua se chovají přibližně jako nezávislé náhodné veličiny. Generování bootstrapového výběru potom probíhá podle následujícího algoritmu, který lze zobecnit i na ARMA modely.

- Nejprve se odhadnou rezidua $\widehat{Y}_t = X_t - \widehat{\beta}_1 X_{t-1} - \cdots - \widehat{\beta}_p X_{t-p}$, kde $\widehat{\beta}_1, \ldots, \widehat{\beta}_p$ jsou konzistentní odhady parametrů $\beta_1, \ldots, \beta_p$.
- Spočte se empirická distribuční funkce F_n centrovaných reziduí $\widehat{Y}_t - \overline{Y}_n$, kde $\overline{Y}_n$ je aritmetický průměr z hodnot $\widehat{Y}_1, \ldots, \widehat{Y}_n$. Centrování je nezbytné, pokud nemáme model s interceptem.
- Generují se náhodné veličiny $Y_1^*, \ldots Y_n^*$, které jsou (podmíněně při daných $X_1, \ldots, X_n$) *iid* a mají distribuční funkci F_n.
- Bootstrapový výběr, který kopíruje strukturu původních dat je generován předpisem

$$X_t^* = \widehat{\beta}_1 X_{t-1}^* + \cdots + \widehat{\beta}_p^* X_{t-p}^* + Y_t^*, \quad t = 1, \ldots, n,$$

kde se položí $X_{1-p}^* = 0, \ldots, X_0^* = 0$ nebo $X_{1-p}^* = X_{1-p}, \ldots, X_0^* = X_0$.

Bose [3] dokázal konzistenci této metody a její zpřesnění proti normální aproximaci pro rozdělení bootstrapových odhadů pro parametry $\beta_1, \ldots, \beta_p$ metodou nejmenších čtverců, Prášková [12] rozšířila tento výsledek na hladké

funkce průměrů jistých vektorů a z nich plynoucí studentizované odhady. Kreiss a Franke [9] dokázali konzistenci metody bootstrap pro třídu M-odhadů v modelech ARMA. Přehled dalších výsledků, zabývajících se použitím varianty reziduální bootstrap v modelech AR a ARMA lze nalézt např. v [10].

Další používanou metodou je tzv. *wild bootstrap*. Tato modifikace metody bootstrap byla původně odvozena pro regresní modely s heterogenními chybami, např. [19], [11]. Lze ji však aplikovat i na časové řady v situacích, kdy je identifikován parametrický model (např. ARMA), ve kterém ale nelze šum modelovat jako posloupnost nezávislých stejně rozdělených náhodných veličin. Sem patří i v současné době velmi populární modely s podmíněnou heteroskedasticitou (modely typu ARCH, GARCH), nebo autoregresní modely s náhodnými parametry, které mají podobnou strukturu jako modely ARCH, viz např. [13], [14], [6].

Uvažujme opět autoregresní model AR(p) definovaný v (9), ve kterém Y_t nejsou nezávislé a stejně rozdělené náhodné veličiny. Mějme pozorování $X_1, \dots, X_n$ a uvažujme odhady autoregresních parametrů metodou nejmenších čtverců, které jsou výpočetně velmi jednoduché. Vzhledem k obecné nestacionaritě však lze obtížně určit jejich asymptotické rozdělení. Odhad rozdělení metodou reziduální bootstrap není v tomto případě konzistentní (např. [13]).

V případě, že Y_t jsou nezávislé nebo slabě závislé, ale heterogenní, lze použít techniku wild bootstrap, který zachovává heteroskedasticitu. Možné jsou dvě varianty této metody.

První z nich generuje bootstrapový výběr na základě regrese. Spočtou se rezidua $r_t = X_t - \widehat{\beta}_1 X_{t-1} - \cdots - \widehat{\beta}_p X_{t-p}$, kde $\widehat{\beta}_1, \dots, \widehat{\beta}_p$ jsou odhady metodou nejmenších čtverců. Dále se generuje nový proces chyb

$$Y_t^w = r_t K_t, \quad t = 1, \dots, n,$$

kde K_t jsou *iid* s nulovou střední hodnotou a jednotkovým rozptylem, nezávislé na $X_0, \dots, X_n$. Potom se generuje bootstrapový výběr podle schématu

$$X_t^w = \widehat{\beta}_1 X_{t-1} + \cdots + \widehat{\beta}_p X_{t-p} + Y_t^w, \quad t = 1, \dots, n.$$

Bootstrapové hodnoty se tedy řídí regresním modelem s konstantními regresory $X_{t-1}, \dots, X_{t-p},\ t = 1, \dots, n$.

Ve druhé variantě se generuje proces chyb $Y_t^w = r_t K_t,\ t = 1, \dots, n$, stejně jako v regresní variantě, ale bootstrapový výběr se generuje podle autoregresního schématu

$$X_{1-p}^{*w} = 0, \dots, X_0^{*w} = 0,$$
$$X_t^{*w} = \widehat{\beta}_1 X_{t-1}^{*w} + \cdots + \widehat{\beta}_p X_{t-p}^{*w} + Y_t^w,\ t = 1, \dots, n,$$

takže přesně kopíruje strukturu závislosti v původním modelu. Zde opět $\widehat{\beta}_1, \dots, \widehat{\beta}_p$ jsou odhady parametrů $\beta_1, \dots, \beta_p$ v původním modelu AR(p).

Lze ukázat, že obě tyto varianty konzistentně odhadují rozdělení odhadů $\widehat{\beta}_1, \dots, \widehat{\beta}_p$ získaných metodou nejmenších čtverců za různých podmínek na hetroskedasticitu chyb Y_t, viz [13], [14], [6].

3.2 Bootstrap nezávislý na modelu (blokový)

Nechť je k dispozici n pozorování časové řady $X_1, \ldots, X_n$; předpokládáme, že jde o stacionární posloupnost, ale o závislostní struktuře nemáme žádné informace.

Bootstrapový výběr se v takovém případě dá sestrojit následovně. Vektor $(X_1, \ldots, X_n)$ se nejdříve rozdělí na bloky délky l. Pro $n = k \cdot l$ tak dostaneme $N = k$ nepřekrývajících se bloků

$$\mathbf{Y}_1 = (X_1, \ldots, X_l), \mathbf{Y}_2 = (X_{l+1}, \ldots, X_{2l}), \ldots, \mathbf{Y}_k = (X_{(k-1)l+1}, \ldots, X_n).$$

Jinou možností je vytvořit $N = n - l + 1$ klouzavých bloků

$$\mathbf{Y}_1 = (X_1, \ldots, X_l), \mathbf{Y}_2 = (X_2, \ldots, X_{l+1}), \ldots, \mathbf{Y}_{n-l+1} = (X_{n-l+1}, \ldots, X_n).$$

Dále se provádí nezávislý náhodný výběr s vracením z populace vektorů $\mathbf{Y}_1, \ldots, \mathbf{Y}_N$. Dostáváme tak postupně vektory $\mathbf{Y}_1^*, \mathbf{Y}_2^* \ldots$. Za bootstrapový výběr potom považujeme vektor náhodných veličin

$$(X_1^*, \ldots, X_n^*) = (\mathbf{Y}_1^*, \mathbf{Y}_2^*, \ldots, \mathbf{Y}_k^*).$$

Existují i další varianty, jak vytvářet bloky. Důležité je, že v bootstrapovém výběru je vždy l po sobě jdoucích pozorování, která mají stejnou strukturu jako původní data.

Z teoretického hlediska, pro dosažení konzistentních výsledků, musí být délka bloku dostatečně dlouhá, $l \to \infty$ pro $n \to \infty$. Existují teoretická odvození pro stanovení optimální délky bloku (např. [10]). V praxi však nelze podle těchto teoretických výsledků vždy postupovat, neboť teoretické stanovení délky bloku závisí na hodnotách autokovarianční funkce, kterou neznáme. Většinou se tedy délka bloku stanoví tak, že se řada pozorování nejdříve odhadne nějakým parametrickým modelem, ve kterém je teoretická autokovarianční funkce známá, a její odhad se potom dosadí do vzorců pro výpočet délky bloku. Existují i jiné algoritmy, založené na konkrétních pozorováních.

Více o blokovém bootstrapu se lze dočíst v Lahiri [10] nebo např. v práci [8]. Tam je možno nalézt také další prameny, které zde pro nedostatek místa neuvádíme.

Reference

[1] Babu G.J., Singh K. (1984). *On one term Edgeworth correction by Efron's bootstrap.* Sankhyā A **46**, 219 – 232.

[2] Bickel P.J., Freedman D.A. (1981). *Some asymptotic theory for the bootstrap.* Ann. Statist. **9**, 1196 – 1217.

[3] Bose A. (1988). *Edgeworth correction by bootstrap in autoregression.* Ann. Statist. **16**, 1709 – 1722.

[4] Efron B. (1979). *Bootstrap methods: another look at the jackknife.* Ann. Statist. **7**, 1 – 26.

[5] Fuller W.A. (1976). *Introduction to statistical time series*. Wiley, New York.

[6] Gonçalves S., Kilian L. (2004). *Bootstrapping autoregression with conditional heteroskedasticity of unknown form*. J. Econometrics **123**, 89 – 120.

[7] Hall P. (1992). *The bootstrap and the Edgeworth expansion*. Springer-Verlag, New York.

[8] Härdle W., Horowitz J., Kreiss J.-P., (2003). *Bootstrap methods for time series*. International Statistical Review **71**, 435 – 459.

[9] Kreiss J.P., Franke J. (1992). *Bootstrapping stationary autoregressive – moving average models*. J. Time Ser. Anal. **13**, 297 – 317.

[10] Lahiri (2003). *Resampling methods for dependent data*. Springer-Verlag, New York.

[11] Mammen E. (1992). *When does bootstrap work? Asymptotic results and simulations*. Springer-Verlag, Heidelberg.

[12] Prášková Z. (1995). *A contribution to bootstrapping autoregressive processes*. Kybernetika **31**, 359 – 373.

[13] Prášková Z. (2002). *Bootstrap in nonstationary autoregression*. Kybernetika **38**, 389 – 404.

[14] Prášková Z. (2003). *Wild bootstrap in RCA(1) model*. Kybernetika **39**, 1 – 12.

[15] Shao J., Tu D. (1995). *The jackknife and bootstrap*. Springer-Verlag, New York.

[16] Singh K. (1981). *On the asymptotic accuracy of Efron's bootstrap*. Ann. Statist. **9**, 1187 – 1195.

[17] Singh K., Xie M. (2003). *Bootlier-plot - bootstrap based outlier detection plot*. Sankhyā **65**, 532 – 559.

[18] Šindlář J. (2003). *Počítačové postupy a výběry z konečné populace*. Diplomová práce MFF UK, Praha.

[19] Wu C.F.J. (1986). *Jackknife, bootstrap and other resampling methods in regression analysis (with discussions)*. Ann. Statist. **14**, 1261 – 1350.

Poděkování: Práce vznikla za podpory výzkumného záměru MŠMT číslo MSM 113200008 a grantu GAČR č. 201/03/0945.

Adresa: Z. Prášková, Univerzita Karlova, Fakulta matematicko-fyzikální, Katedra pravděpodobnosti a matematické statistiky, Sokolovská 83, 186 75 Praha 8

E-mail: praskova@karlin.mff.cuni.cz

DETEKCE LINEÁRNÍHO TRENDU V ROZPTYLU NORMÁLNÍHO ROZDĚLENÍ

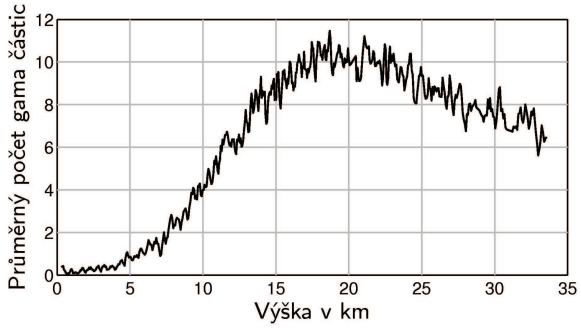
Luboš Prchal

Klíčová slova: Detekce změny v rozptylu, regrese v L_1 a L_2 normě, radioaktivní záření.

Abstrakt: Tento příspěvek je věnován detekci a odhadu dvou neznámých bodů změny, mezi nimiž se rozptyl nezávislých normálně rozdělených náhodných veličin lineárně mění z jedné konstantní úrovně na druhou. V první části je navržena vhodná testová statistika, v druhé části se pak věnujeme porovnání L_1 a L_2 odhadů bodů změny a parametrů rozptylu. Postup je ilustrován na reálné analýze variability vertikálních profilů radioaktivního záření.

1 Úvod

Studium problematiky detekce a odhadu měnícího se rozptylu normálně rozdělených náhodných veličin bylo motivováno praktickou potřebou analyzovat variabilitu měření vertikálních profilů radioaktivního záření. Tato data statistické veřejnosti představil na konferenci Robust'98 Hlubinka [2]. Připomeňme, že data se měří pomocí meteorologických balónů vypouštěných ze stanici v Praze-Libuši a stoupajících do výšky kolem 35-ti km, přičemž výsledkem měření jsou dvojice $(x_i, y_i),\ i = 1, \ldots, n$, představující průměrnou intenzitu záření y_i v nadmořské výšce x_i. Typický vertikální profil gama radiace je znázorněn na obrázku 1.



Obrázek 1: Typický průběh průměrného počtu gama částic v závislosti na nadmořské výšce.

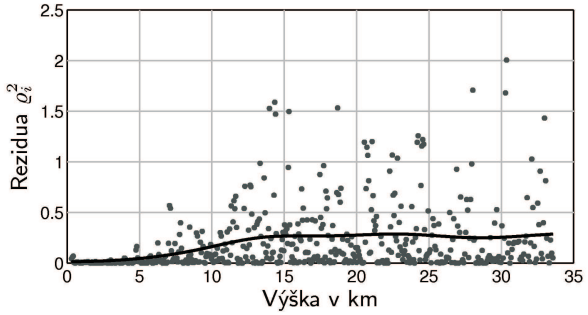
Hlubinka [2] podrobně diskutuje jak parametrický tak neparametrický přístup k modelování „trendu" pomocí regresního modelu $Y = m(X)+\eta$, kde X a Y jsou náhodné veličiny odpovídající nadmořské výšce, resp. intenzitě radiace, $m(\cdot)$ představuje průměrnou radiaci a η náhodnou složku měření.

Nedostatky navrhovaného parametrického modelu založeného na derivaci tzv. *Richardsovy růstové křivky* jsou pak odstraněny jeho rozšířením podrobně popsaným v práci [3].

V tomto příspěvku se zaměříme na evidentně se měnící variabilitu měření radioaktivního záření. Odhad rozptylu radiace v závislosti na výšce $\sigma_i^2 = \mathsf{var}\,[Y_i|\, X = x_i]$ založíme na čtvercích reziduí $\varrho_i^2 = \big(Y_i - \widehat{m}(X_i)\big)^2$. Jejich průběh proložený jádrovým odhadem

$$g_K(x) = \frac{\sum_{i=1}^n y_i K\big(\frac{x-x_i}{h}\big)}{\sum_{i=1}^n K\big(\frac{x-x_i}{h}\big)} \tag{1}$$

s normálním jádrem a vyhlazovacím parametrem h zvoleným pomocí křížového ověřování (*cross validation*) je znázorněn na obrázku 2. Poznamenejme, že budeme-li dále mluvit o *jádrové regresi*, pak budeme mít na mysli neparametrický jádrový odhad (1).



Obrázek 2: Typický průběh čtverců reziduí ϱ_i^2 proložený jádrovou regresí s normálním jádrem a vyhlazovacím parametrem $h = 2,74$.

Pro parametrický popis chování reziduí ϱ_i^2 se jako vhodný jeví lineární model ve tvaru

$$\boldsymbol{\varrho} = \boldsymbol{D}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

kde $\boldsymbol{\varrho} = (\varrho_1^2, \ldots, \varrho_n^2)'$, $\boldsymbol{\beta} = (\sigma^2, \delta^2)'$ je vektor neznámých reálných parametrů, $\sigma, \delta > 0$, a regresní matice $\boldsymbol{D} = \big(\mathbf{1} \quad \boldsymbol{d}\big)_{n\times 2}$ je dána vektorem samých jedniček $\mathbf{1}_{n\times 1}$ a vektorem $\boldsymbol{d} = (d_1, \ldots, d_n)'$ definovaným předpisem

$$\begin{aligned} d_i &= 0, && i = 1, \ldots, s, \\ &= \frac{x_i - x_s}{x_t - x_s}, && i = s+1, \ldots, t, \\ &= 1, && i = t+1, \ldots, n. \end{aligned}$$

Uvědomme si, že $\boldsymbol{\beta}$ obecně závisí na neznámých bodech změny x_s, x_t, a dodejme, že s pomocí uvedeného regresního modelu odhadneme podmíněný rozptyl σ_i^2 jako $\widehat{\sigma}_i^2 = \boldsymbol{d}_i'\widehat{\boldsymbol{\beta}}$, kde $\boldsymbol{d}_i'$ představuje $i-$tý řádek matice $\boldsymbol{D}$ a $\widehat{\boldsymbol{\beta}}$ je „vhodný“ odhad neznámých parametrů.

2 Testování modelu

Podívejme se nyní, jak otestovat hypotézu o konstantním rozptylu proti alternativě, že existují dva body změny, v nichž se charakter rozptylu mění v duchu výše popsaného regresního modelu.

Než se dostaneme k samotnému testování, dodejme, že v této části budeme předpokládat normální rozdělení a nezávislost jednotlivých měření Y_i. Dále předpokládejme, že „sousední měření mají stejný rozptyl“, přesněji, že $\sigma^2_{2j-1} = \sigma^2_{2j},\ j = 1, \dots, \widetilde{n}$, kde $\widetilde{n} = \lfloor n/2 \rfloor$ je spodní celá část $n/2$. Uvědomme si, že z tohoto předpokladu přirozeně vyplývá, že body změny s a t jsou sudá čísla.

Pracujeme tedy s náhodnými veličinami

$$\begin{aligned} Y_{2j-1} &\sim \mathrm{N}\big(m(x_{2j-1}), \sigma^2_{2j-1}\big), \qquad j = 1, \dots, \widetilde{n}, \\ Y_{2j} &\sim \mathrm{N}\big(m(x_{2j}), \sigma^2_{2j-1}\big), \qquad j = 1, \dots, \widetilde{n}, \end{aligned}$$

a chceme testovat hypotézu o konstantnosti jejich rozptylu, tj.

$$\mathrm{H}_1\colon\ \sigma_i^2 = \sigma^2, \qquad \forall x_i \in \{x_1, x_2, \dots, x_n\},$$

proti alternativě

$$\begin{aligned} \mathrm{A}_1\colon \exists\, x_s, x_t \in \{x_1, x_2, \dots, x_n\}, &\qquad x_s < x_t, \\ \sigma_i^2 = \sigma^2, &\qquad x_i \in \{x_1, \dots, x_s\}, \\ = \sigma^2 + \frac{x_i - x_s}{x_t - x_s}\delta^2, &\qquad x_i \in \{x_{s+1}, \dots, x_t\}, \\ = \sigma^2 + \delta^2, &\qquad x_i \in \{x_{t+1}, \dots, x_n\}. \end{aligned}$$

Uvažme, že díky normálnímu rozdělení veličin Y_i mají náhodné veličiny

$$V_j = \frac{\big(Y_{2j-1} - m(x_{2j-1})\big)^2 + \big(Y_{2j} - m(x_{2j})\big)^2}{2}, \qquad j = 1, \dots, \widetilde{n},$$

exponenciální rozdělení s parametry $\vartheta_j = \sigma^2_{2j-1}$. Nahraďme proto nelineární regresní model $m(x)$ jeho odhadem $\widehat{m}(x)$ a neznámý podmíněný rozptyl σ^2_{2j-1} jeho parametrickým odhadem založeným na lineárním modelu $\widehat{\sigma}^2_{2j-1} = \boldsymbol{\beta}'\boldsymbol{d}_{2j-1}, j = 1, \dots, \widetilde{n}$. Definujme dále náhodné veličiny

$$W_j = \frac{\big(Y_{2j-1} - \widehat{m}(x_{2j-1})\big)^2 + \big(Y_{2j} - \widehat{m}(x_{2j})\big)^2}{2} = \frac{\varrho^2_{2j-1} + \varrho^2_{2j}}{2}, j = 1, \dots, \widetilde{n},$$

a předpokládejme, že i veličiny W_j mají díky normálnímu rozdělení Y_i a odhadu $m(x)$ metodou nejmenších čtverců exponenciální rozdělení, tentokrát s parametry $\theta_j = \widehat{\sigma}^2_{2j-1} = \boldsymbol{\beta}'\boldsymbol{d}_{2j-1}$.

Pomocí právě popsané transformace jsme nejen přešli od normálně rozdělených Y_i k výběru $W_j \sim \mathrm{Exp}(\theta_j),\ j = 1, \dots, \widetilde{n}$, ale současně naši hypotézu H_1, resp. alternativu A_1, můžeme ekvivalentně přepsat jako hypotézu o konstantní hodnotě parametru exponenciálního rozdělení

$$\mathrm{H}_2\colon\ \theta_j = \sigma^2, \qquad j = 1, \dots, \widetilde{n},$$

proti alternativě

$$\begin{aligned} \mathrm{A}_2\colon \exists\, \widetilde{s}, \widetilde{t} \in \{1, 2, \dots, \widetilde{n}\}, & \qquad 1 \le \widetilde{s} < \widetilde{t} \le \widetilde{n}, \\ \theta_j = \sigma^2, & \qquad j = 1, \dots, \widetilde{s}, \\ = \sigma^2 + \frac{x_{2j-1} - x_{2\widetilde{s}}}{x_{2\widetilde{t}} - x_{2\widetilde{s}}}\delta^2, & \qquad j = \widetilde{s}+1, \dots, \widetilde{t}, \\ = \sigma^2 + \delta^2, & \qquad j = \widetilde{t}+1, \dots, \widetilde{n}, \end{aligned}$$

kde $\widetilde{s} = \lfloor s/2 \rfloor$ a $\widetilde{t} = \lfloor t/2 \rfloor$.

Odvození testové statistiky T pro úlohu testování hypotézy H_2 proti alternativě A_2 vychází z [4] a [1] a je podrobně popsáno v [3]. Na tomto místě jen uveďme, že ji můžeme vyjádřit vztahem

$$T = \frac{\sum_{j=2}^{\widetilde{n}} \gamma_j W_j}{\Gamma \cdot \sum_{j=1}^{\widetilde{n}} W_j},$$

přičemž normovací konstanta Γ a konstanty γ_j mají v tomto konkrétním případě tvar

$$\Gamma = \frac{1}{\widetilde{n}} \sum_{j=1}^{\widetilde{n}} \gamma_j \quad \text{a} \quad \gamma_j = \frac{(j-1)(j-2)}{2} + \sum_{\widetilde{t}=j}^{\widetilde{n}} \sum_{\widetilde{s}=1}^{j-1} \frac{x_{2j-1} - x_{2\widetilde{s}}}{x_{2\widetilde{t}} - x_{2\widetilde{s}}}.$$

Lze ukázat, že testová statistika T má za platnosti hypotézy H_2 a při $\widetilde{n} \to \infty$ asymptoticky normální rozdělení, a tudíž

$$U = \frac{T - \mathsf{E}\, T}{\sqrt{\mathsf{var}\, T}}$$

má asymptoticky normované normální rozdělení $\mathrm{N}(0, 1)$, přičemž $\mathsf{E}\, T = 1$ a rozptyl $\mathsf{var}\, T$ lze vyjádřit vztahem

$$\mathsf{var}\, T = \frac{\widetilde{n}}{\widetilde{n}+1} \left[1 + \frac{\sum_{j=2}^{\widetilde{n}} \gamma_j^2}{\left(\sum_{j=2}^{\widetilde{n}} \gamma_j \right)^2} \right] - 1 = \frac{\widetilde{n}}{\widetilde{n}+1} \frac{\sum_{j=2}^{\widetilde{n}} \gamma_j^2}{\left(\sum_{j=2}^{\widetilde{n}} \gamma_j \right)^2} - \frac{1}{\widetilde{n}+1}.$$

Ze simulací ilustrujících rychlost konvergence rozdělení statistiky U k normálnímu rozdělení vyplývá, že asymptotických vlastností lze využít již při $n = 50$, podrobněji viz [3]. V tom případě hypotézu H_2, resp. H_1, zamítáme na hladině α ve prospěch alternativy A_2, resp. A_1, jestliže $U > u(1-\alpha)$, kde $u(\alpha)$ je $100\alpha\%$-ní kvantil normovaného normálního rozdělení. Připomeňme, že testujeme proti jednostranné alternativě „zvětšení“ variability o $\delta^2 > 0$, a proto uvažujeme pouze „horní“ kvantil $u(1-\alpha)$.

Při analýze variability radiace máme k dispozici výběry s rozsahy $n \approx 550$, můžeme tedy bez obav užít asymptotického rozhodovacího pravidla, přičemž dle očekávání hypotézu H_1 na hladině $\alpha = 0,05$ jednoznačně zamítáme pro všechna pozorování beta i gama částic.

3 Odhad modelu

V předcházejících odstavcích jsme ukázali jak otestovat „adekvátnost“ uvažovaného regresního modelu $\boldsymbol{\varrho} = \boldsymbol{D\beta} + \boldsymbol{\varepsilon}$ a zbývá nám tedy odhadnout jeho neznámé parametry; body změny s a t a složky rozptylu σ^2 a δ^2. Odhad parametrů provedeme ve dvou krocích. Nejprve pro pevné hodnoty s a t, $1 \leq s < t \leq n$, odhadneme parametry $\boldsymbol{\beta}(s,t)$ jako

$$\widehat{\boldsymbol{\beta}}(s,t) = \underset{\boldsymbol{\beta}\in\mathbb{R}^2}{\arg\min} \sum_{i=1}^{n} \Psi\big(\varrho_i^2 - \boldsymbol{d}_i'\boldsymbol{\beta}\big) = \underset{\boldsymbol{\beta}\in\mathbb{R}^2}{\arg\min}\, \mathrm{RS}\big(\boldsymbol{\beta}(s,t)\big),$$

kde Ψ je vhodně zvolená funkce. Ve druhém kroku odhadneme body změny s a t tak, abychom minimalizovali ztrátovou funkci $\mathrm{RS}\big(\boldsymbol{\beta}(s,t)\big)$, tedy

$$\big\{\widehat{s}, \widehat{t}\,\big\} = \underset{\substack{s=1,\ldots,n-1\\ t=s+1,\ldots,n}}{\arg\min}\, \mathrm{RS}\big(\widehat{\boldsymbol{\beta}}(s,t)\big).$$

Tím také dostaneme výsledný odhad $\boldsymbol{\beta}$ pomocí $\widehat{\boldsymbol{\beta}}\,(\widehat{s}, \widehat{t}\,)$.

Parametry lineárního modelu většinou odhadujeme metodou *nejmenších čtverců* (dále jen L_2 *regrese*). V našem případě však nemáme splněn jeden ze základních předpokladů klasického lineárního modelu, a sice homoskedasticitu náhodné složky ε. „Neblahý vliv“ heteroskedasticity náhodné složky na odhad parametrů $\boldsymbol{\beta}$ lze omezit užitím metody *vážených nejmenších čtverců* (WLS) s diagonální maticí vah $\boldsymbol{W}_{n\times n}$ tvořenou prvky $w_{ii} = 1/\widehat{\tau}_i^2$, kde $\widehat{\tau}_i^2$ je odhad rozptylu $\mathsf{var}\,\varepsilon_i = \tau_i^2$. Jako vhodný, „nezávislý“ na metodě nejmenších čtverců se nabízí odhad pomocí již zmíněné jádrové regrese (1) ve tvaru $\widehat{\tau}_i^2 = \big(\varrho_i^2 - \widehat{g}_K(x_i)\big)^2$. S využitím informace o variabilitě náhodné složky ε dostáváme odhad neznámých parametrů $\boldsymbol{\beta}$ v podobě

$$\widehat{\boldsymbol{\beta}}_{WLS} = (\boldsymbol{D}'\boldsymbol{W}\boldsymbol{D})^{-1}\boldsymbol{D}'\boldsymbol{W}\boldsymbol{\varrho}. \tag{2}$$

Jako robustní alternativu k metodám nejmenších čtverců uveďme regresi v L_1 normě (dále jen L_1 *regrese*) odpovídající minimalizační úloze

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^2} \sum_{i=1}^{n} |\varrho_i^2 - \boldsymbol{d}_i'\boldsymbol{\beta}|. \tag{3}$$

Jedním z přístupů k řešení L_1 regrese je numerická metoda *iteračně vážených nejmenších čtverců* (IWLS). Minimalizace úlohy (3) pomocí metody IWLS odpovídá řešení soustavy

$$\boldsymbol{D}'\boldsymbol{W}(\boldsymbol{\beta})\boldsymbol{\varrho} = \boldsymbol{D}'\boldsymbol{W}(\boldsymbol{\beta})\boldsymbol{D\beta},$$

vzhledem k neznámým parametrům $\boldsymbol{\beta} \in \mathbb{R}^2$, přičemž matice vah $\boldsymbol{W}(\boldsymbol{\beta})_{n\times n}$ je diagonální s prvky $w_{ii}(\boldsymbol{\beta})$ definovanými předpisem

$$\begin{aligned} w_{ii}(\boldsymbol{\beta}) &= \frac{\mathrm{sgn}\big(\varrho_i^2 - \boldsymbol{d}_i'\boldsymbol{\beta}\big)}{\varrho_i^2 - \boldsymbol{d}_i'\boldsymbol{\beta}}, \qquad && \varrho_i^2 - \boldsymbol{d}_i'\boldsymbol{\beta} \neq 0, \\ &= 0, && \varrho_i^2 - \boldsymbol{d}_i'\boldsymbol{\beta} = 0. \end{aligned} \tag{4}$$

Jelikož váhy na rozdíl od metody WLS tentokrát závisejí na neznámých parametrech $\boldsymbol{\beta}$, není možné pro odhad $\boldsymbol{\beta}$ užít přímo vztah (2), nýbrž je třeba přikročit k numerickému řešení. Metoda IWLS vychází z počátečního odhadu $\boldsymbol{\beta}_{L_1}^{(0)}$, který v jednotlivých iteracích postupně „vylepšuje“ předpisem

$$\boldsymbol{\beta}_{L_1}^{(l+1)} = \left(\boldsymbol{D}'\boldsymbol{W}(\boldsymbol{\beta}_{L_1}^{(l)})\boldsymbol{D}\right)^{-1} \boldsymbol{D}'\boldsymbol{W}(\boldsymbol{\beta}_{L_1}^{(l)})\boldsymbol{\varrho} \tag{5}$$

až do splnění vhodného zastavovacího pravidla. Dodejme, že $\boldsymbol{\beta}_{L_1}^{(l)}$ značí L_1 odhad parametrů $\boldsymbol{\beta}$ po l iteracích dle vztahu (5) a $\boldsymbol{W}\left(\boldsymbol{\beta}_{L_1}^{(l)}\right)$ jsou známé váhy dány předpisem (4) pro již spočtenou hodnotu $\boldsymbol{\beta}_{L_1}^{(l)}$. V jednotlivých iteracích tedy známe matici vah $\boldsymbol{W}(\boldsymbol{\beta}_{L_1}^{(l)})$, a proto následující odhad parametrů $\boldsymbol{\beta}_{L_1}^{(l+1)}$ získáme stejně jako u metody WLS vztahem (2).

Druhým přístupem vedoucím k nalezení optimálního řešení problému L_1 regrese, pokud takové řešení existuje, je přeformulovat regresní úlohu (3) jako standardní minimalizační úlohu lineárního programování ve tvaru

$$\min_{\boldsymbol{\epsilon}^+,\,\boldsymbol{\epsilon}^-} \sum_{i=1}^{n} \left(\epsilon_i^+ + \epsilon_i^-\right)$$

za podmínek

$$\boldsymbol{d}_i'\boldsymbol{\beta} + \epsilon_i^+ - \epsilon_i^- = \varrho_i^2, \quad \epsilon_i^+, \epsilon_i^- \geq 0, \quad i = 1, \ldots, n, \quad \sigma^2, \delta^2 \geq 0.$$

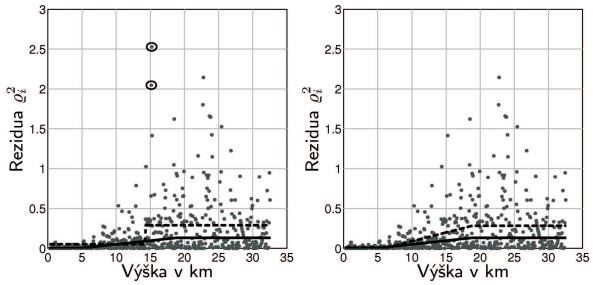
K jejímu vyřešení pak lze užít standardních nástrojů obsažených v matematických a statistických programech, např. funkci `linprog` implementovanou v `Optimization Toolbox` programového vybavení `Matlab, Release13`. Dodejme, že zde prezentované výsledky jsou získány metodu IWLS s tím, že výrazně rychlejší numerické řešení se jen nepatrně liší od přesného řešení úlohy lineárního programování. Podrobněji je volba metody diskutována v práci [3].

4 Porovnání metod

Podívejme se nyní na získané odhady prezentovanými metodami jednak z pohledu odhadu složek variability σ^2 a δ^2, jednak z pohledu bodů, ve kterých dochází k jejich změnám.

O odhadech bodů změny se obecně dá říci, že použití L_2 regrese vede k odhadům pouze jednoho bodu změny, tedy na model se skokovou změnou v chování rozptylu. Tento typický rys L_2 regrese je způsoben značnou citlivostí L_2 odhadů na velké, „odlehlé“ hodnoty reziduí. Naproti tomu, užitím L_1 ztrátové funkce získáme „očekáváný“ odhad s postupným lineárním růstem variability $\sigma^2(x)$.

Rozdíl v chování L_2 a L_1 odhadů ilustrujme na pozorování gama částic z 10. října 1995. Průběh čtverců reziduí ϱ_i^2 proložený L_2 i L_1 odhadem variability je znázorněn na obrázku 3 (levý graf). Na tomtéž obrázku 3 jsou kolečkem vyznačena dvě rezidua ve výšce asi 15 km, která způsobují skok



Obrázek 3: Čtverce reziduí proložené odhadnutým rozptylem metodou WLS (čárkovaně) a L_1 regresí (plná čára). Levý graf znázorňuje odhady získané ze všech reziduí, pravý graf odhady po vynechání dvou zakroužkovaných „odlehlých“ reziduí.

v L_2 odhadu. Na pravém grafu obrázku 3, který odpovídá stejnému pozorování, vynecháme-li dvě vyznačená rezidua, vidíme, že nově odhadnuté body změny metodou WLS se „přiblížily“ nezměněnému L_1 odhadu.

Ačkoli nemáme a priori žádnou informaci o chování variability měření, zdá se rozumné přiklonit se k robustnějším L_1 odhadům, vesměs podporujícím myšlenku lineární změny mezi dvěma konstantními hladinami rozptylu. Odhady bodů změny x_s a x_t pomocí L_1 regrese více odpovídají také naší původní představě o průběhu rozptylu založené na neparametrické jádrové regresi (1).

Připomeňme, že parametry σ^2 a δ^2 si můžeme představit jako rozptyl měření ve výškách do x_s, resp. nad x_t. Uvážíme-li dále, že jsme při testování předpokládali normální rozdělení dat, pak odhad neznámých parametrů σ^2 a δ^2 regresí v L_1 normě se zdá být nevhodný. Pro odhad samotných složek rozptylu σ^2 a δ^2 bychom spíše měli volit metodu nejmenších čtverců, resp. její váženou variantu WLS.

Ve světle předcházejících úvah se jako optimální metoda pro odhad chování variability měření radiace jeví kombinace L_1 a L_2 přístupu. Počítejme proto nejprve L_1 odhad bodů změny obvyklým dvoukrokovým postupem, tj. v prvním kroku odhadněme pro pevné body změny s a t, $1 \leq s < t \leq n$, složky rozptylu vztahem

$$\widehat{\boldsymbol{\beta}}_{L_1}(s,t) = \operatorname*{arg\,min}_{\boldsymbol{\beta}\in\mathbb{R}^2} \mathrm{RS}_{L_1}\big(\boldsymbol{\beta}(s,t)\big) = \operatorname*{arg\,min}_{\boldsymbol{\beta}\in\mathbb{R}^2} \sum_{i=1}^{n} |\varrho_i^2 - \boldsymbol{d}_i'\boldsymbol{\beta}|,$$

a na základě ztrátové funkce $\mathrm{RS}_{L_1}\big(\boldsymbol{\beta}(s,t)\big)$ pak v druhém kroku odhadněme body změny jako

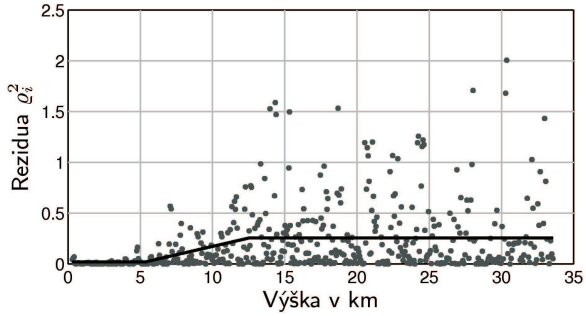
$$\{s^*, t^*\} = \operatorname*{arg\,min}_{\substack{s=1,\ldots,n-1\\ t=s+1,\ldots,n}} \mathrm{RS}_{L_1}\big(\widehat{\boldsymbol{\beta}}_{L_1}(s,t)\big).$$

Složky rozptylu následně odhadněme metodou vážených nejmenších čtverců

$$\boldsymbol{\beta}^* = \widehat{\boldsymbol{\beta}}_{WLS}(s^*, t^*) = (\boldsymbol{D}'_{s^*t^*}\boldsymbol{W}\boldsymbol{D}_{s^*t^*})^{-1}\boldsymbol{D}'_{s^*t^*}\boldsymbol{W}\boldsymbol{\varrho},$$

kde $\boldsymbol{D}_{s^*t^*}$ je regresní matice odpovídající pevným bodů změny $s = s^*$ a $t = t^*$. Matice vah $\boldsymbol{W}$ je diagonální s prvky $w_{ii} = 1/\widehat{\tau}_i^2, i = 1, \ldots, n$, přičemž $\widehat{\tau}_i^2$ je odhad rozptylu ϱ_i^2 získaný pomocí jádrové regrese $\widehat{\tau}_i^2 = \left(\varrho_i^2 - \widehat{g}_K(x_i)\right)^2$.

Představená kombinace L_1 a L_2 metod si zachovává výhody robustního odhadu bodů změny, přičemž oproti samotné L_1 metodě „věrněji“ odhaduje složky rozptylu. Odhad průběhu rozptylu měření získaný touto kombinovanou metodou je znázorněn na obrázku 4.



Obrázek 4: Optimální odhad variability měření kombinovanou metodou.

Reference

[1] Gupta A.K., Ramanayake A. (2001). *Change points with linear trend for the exponential distribution.* J. Statist. Plann. Inference **93**, 181 – 195.

[2] Hlubinka D. (1998). *Metody pro prokládání křivek s použitím na reálných datech.* In ROBUST'98 (Antoch J. a Dohnal G., eds.), JČMF, Praha, 55 – 75.

[3] Prchal L. (2004). *Neparametrické odhady pro analýzu funkcionálních dat.* Diplomová práce, MFF UK, Praha.

[4] Worsley K.J. (1986). *Confidence regions and test for a change point in a sequence of exponential family random variables.* Biometrika **73**, 91 – 104.

Poděkování: Autor děkuje prof. RNDr. Jaromíru Antochovi, CSc., za jeho nezištnou pomoc a neocenitelné rady v průběhu vzniku tohoto článku. Práce vznikla s podporou grantů GAČR 201/03/0945 a MSM 113200008.

Adresa: L. Prchal, KPMS MFF UK, Sokolovská 83, 186 75 Praha 8

E-mail: prchal@karlin.mff.cuni.cz

USPOŘÁDÁNÍ VÝSLEDKŮ ŠETŘENÍ REPREZENTOVANÝCH FUZZY ČÍSLY

Zdeněk Půlpán

Klíčová slova: Fuzzy relace, fuzzy čísla a jejich uspořádání.

Abstrakt: V příspěvku je diskutována otázka uspořádání fuzzy čísel; je navržena jedna z možností jak problém částečně řešit.

1 Úvod

Nemáme dosud univerzální definici uspořádání fuzzy čísel, která by uspokojivě řešila řadu různorodých aplikačních úloh (např. porovnávání variant zobrazených fuzzy čísly, varianty mohou být výsledkem expertního řízení). To vyplývá ze složitosti struktury fuzzy čísel, která je výsledkem naší snahy vložit do fuzzy čísel co nejvíce např. empirických informací. Definované uspořádání fuzzy čísel je pak většinou antisymetrické, ostře tranzitivní (tj. z použité definice vyplývá pro fuzzy čísla, zobrazující ostrá reálná čísla, klasická tranzitivita jejich uspořádání), které ale není úplné (nelze porovnat každá dvě fuzzy čísla) [1], [2], [3], [4].

Uvedeme prostřednictvím určité fuzzy relace porovnání dvou fuzzy čísel založené na průběhu „levé" a „pravé" části jejich funkce věrohodnosti. Protože však i fuzzy čísla je možné zavést v různé šíři obecnosti, musíme se nejdříve domluvit na tom, s jakou třídou fuzzy množin budeme pracovat. To vysvětlíme v následující definici.

Definice 1. Fuzzy množinu $\underset{\sim}{A}$ na R, kde R je množina reálných čísel, nazveme *fuzzy číslem*, když její funkci věrohodnosti $\mu_A : R \to \langle 0;1\rangle$ můžeme zapsat ve tvaru

$$\mu_A(x) = \left\{ \begin{array}{lll} L_A(x) & \text{pro} & x \in (-\infty, x_1^A) \\ 1 & \text{pro} & x \in \langle x_1^A, x_2^A\rangle, x_1^A \le x_2^A \\ P_A(x) & \text{pro} & x \in (x_2^A, \infty), \end{array} \right\} \qquad (1)$$

kde funkce $L_A : (-\infty, x_1^A) \to \langle 0;1)$ je neklesající a spojitá, $\lim\limits_{x\to-\infty} L_A(x) = 0$, $\lim\limits_{x\to x_1^A-} L_A(x) = 1$, a funkce $P_A : (x_2^A, +\infty) \to \langle 0;1)$ je nerostoucí a spojitá, $\lim\limits_{x\to x_2^A+} P_A(x) = 1$, $\lim\limits_{x\to\infty} P_A(x) = 0$. Budeme předpokládat, že $\int\limits_{-\infty}^{\infty} \mu_A(x)dx < +\infty$.

Poznámka 1. Funkci L_A (resp. P_A) z definice 1. nazýváme levou (resp. pravou) částí fuzzy čísla $\underset{\sim}{A}$.

Poznámka 2. Definiční obor funkce L_A (resp. P_A) je možné rozšířit na celé R dodefinováním $L_A(x) \equiv 0$ pro $x \notin (-\infty, x_1^A)$ (resp. $P_A(x) \equiv 0$ pro $x \notin (x_2^A, +\infty)$). S takto rozšířenou funkcí L_A (resp. P_A) budeme v další části pracovat.

Příklad 1. Fuzzy množina $\underset{\sim}{\mathrm{A}}$ na R, definovaná následujícím předpisem

$$\mu_A(x) = \begin{cases} e^{x+\varepsilon} & \text{pro} \quad x < -\varepsilon \\ 1 & \text{pro} \quad x \in \langle -\varepsilon, \varepsilon \rangle \\ e^{-x+\varepsilon} & \text{pro} \quad x > \varepsilon, \varepsilon > 0 \end{cases}$$

je fuzzy číslem.

Příklad 2. Fuzzy množina $\underset{\sim}{\mathrm{B}}$ na R, jejíž funkce věrohodnosti je tvaru

$$\mu_B(x) = \begin{cases} L_B(x) & \text{pro} \quad x \in (-\infty, b) \\ 1 & \text{pro} \quad x = b \\ P_B(x) & \text{pro} \quad x \in (b, +\infty), b \in R, \end{cases}$$

kde

$$L_B(x) = \begin{cases} 0 & \text{pro} \quad x \in (-\infty, a\rangle \\ \frac{x-b}{b-a} + 1 & \text{pro} \quad x \in (a, b) \\ 0 & \text{pro} \quad x \in \langle b, \infty), a \in R; \end{cases}$$

$$P_B(x) = \begin{cases} 0 & \text{pro} \quad x \in (-\infty, b) \\ \frac{x-b}{b-c} + 1 & \text{pro} \quad x \in \langle b, c\rangle \\ 0 & \text{pro} \quad x \in (c, +\infty); a < b < c; c \in R, \end{cases}$$

se nazývá trojúhelníkové fuzzy číslo. Je jednoznačně určeno reálnými čísly $a, b, c, -\infty < a < b < c < +\infty$.

Poznámka 3. Fuzzy množinu $\underset{\sim}{\mathrm{C}}$ na R, jejíž funkce věrohodnosti μ_c je tvaru

$$\mu_c(x) = \begin{cases} L_c(x) & \text{pro} \quad x \in (-\infty, b_1) \\ 1 & \text{pro} \quad x \in \langle b_1, b_2\rangle \\ P_c(x) & \text{pro} \quad x \in (b_2, +\infty), b_1 < b_2, \end{cases}$$

kde

$$L_c(x) = \begin{cases} 0 & \text{pro} \quad x \in (-\infty, a\rangle, a < b_1 \\ \frac{x-b_1}{b_1-a} + 1 & \text{pro} \quad x \in (a, b_1\rangle \\ 0 & \text{pro} \quad x \in (b_1, +\infty) \end{cases}$$

$$P_c(x) = \begin{cases} 0 & \text{pro} \quad x \in (-\infty, b_2) \\ \frac{x-b_2}{b_2-c} + 1 & \text{pro} \quad x \in \langle b_2, c) \\ 0 & \text{pro} \quad x \in \langle c, +\infty) \end{cases}$$

nazýváme lichoběžníkové fuzzy číslo. Je jednoznačně určeno reálnými čísly $a, b_1, b_2, c, -\infty < a < b_1 < b_2 < c < +\infty$. (Vzhledem k (1) je $x_1^c = b_1, x_2^c = b_2$.)

Základem pro porovnávání dvou fuzzy čísel $\underset{\sim}{\mathrm{A}}$, $\underset{\sim}{\mathrm{B}}$ bývá jejich průsek $\underset{\sim}{\mathrm{A}} \wedge \underset{\sim}{\mathrm{B}}$, resp. spojení $\underset{\sim}{\mathrm{A}} \vee \underset{\sim}{\mathrm{B}}$. Funkce věrohodnosti průseku $\mu_{A\wedge B}$ je definována

$$\mu_{A\wedge B}(z) = \sup\{\min\{\mu_A(x), \mu_B(y)\}; z = \min\{x, y\}, x, y \in R\},$$

resp. funkce věrohodnosti pro spojení $\mu_{A\vee B}$ je definována

$$\mu_{A\vee B}(z) = \sup\{\min\{\mu_A(x), \mu_B(y)\}; z = \max\{x, y\}, x, y \in R\}.$$

Řekneme pak, že fuzzy číslo $\underset{\sim}{\mathrm{A}}$ je větší nebo rovno fuzzy číslu $\underset{\sim}{\mathrm{B}}$ (označujeme $\underset{\sim}{\mathrm{A}} \geq \underset{\sim}{\mathrm{B}}$), jestliže platí

$$\underset{\sim}{\mathrm{A}} = \underset{\sim}{\mathrm{A}} \vee \underset{\sim}{\mathrm{B}}, \quad \text{resp.} \quad \underset{\sim}{\mathrm{B}} = \underset{\sim}{\mathrm{A}} \wedge \underset{\sim}{\mathrm{B}}.$$

Takto zavedená relace uspořádání mezi fuzzy čísly je reflexivní, antisymetrická a tranzitivní:

- $\underset{\sim}{A} \geq \underset{\sim}{A}$ je vyjádřením pravdivého tvrzení $\underset{\sim}{A} = \underset{\sim}{A} \vee \underset{\sim}{A}$;
- $\underset{\sim}{A} \geq \underset{\sim}{B}$ a $\underset{\sim}{B} \geq \underset{\sim}{A}$ znamená $\underset{\sim}{A} = \underset{\sim}{A} \vee \underset{\sim}{B}$ a $\underset{\sim}{B} = \underset{\sim}{B} \vee \underset{\sim}{A} = \underset{\sim}{A} \vee \underset{\sim}{B}$, pak ale $\underset{\sim}{A} = \underset{\sim}{B}$;
- $\underset{\sim}{A} \geq \underset{\sim}{B}$ a $\underset{\sim}{B} \geq \underset{\sim}{C}$ znamená $\underset{\sim}{A} = \underset{\sim}{A} \vee \underset{\sim}{B}$ a $\underset{\sim}{B} = \underset{\sim}{B} \vee \underset{\sim}{C}$, pak ale $\underset{\sim}{A} = \underset{\sim}{A} \vee (\underset{\sim}{B} \vee \underset{\sim}{C}) = (\underset{\sim}{A} \vee \underset{\sim}{B}) \vee \underset{\sim}{C} = \underset{\sim}{A} \vee \underset{\sim}{C}$, což je totéž jako $\underset{\sim}{A} \geq \underset{\sim}{C}$.

Dokonce množina fuzzy čísel spolu s operacemi $\wedge$ a $\vee$ tvoří distributivní svaz, který ale není úplný ([2]).

Výraz $\underset{\sim}{A} \geq \underset{\sim}{B}$ označuje totéž jako $\underset{\sim}{B} \leq \underset{\sim}{A}$; $\underset{\sim}{A} > \underset{\sim}{B}$ označuje $\underset{\sim}{A} \geq \underset{\sim}{B}$ a současně $\underset{\sim}{A} \neq \underset{\sim}{B}$.

Porovnávání fuzzy čísel pomocí průseku (resp. ekvivalentně pomocí spojení) nedává možnost posoudit významnost vztahu $\geq$ (nejde o fuzzy relaci). (Nenabízí také uspokojivou interpretaci v případě tzv. fuzzy symetrických fuzzy čísel, jak uvidíme dále.) Třída neporovnatelných fuzzy čísel (uvedenou metodou) je rozsáhlá, a to omezuje praktickou použitelnost.

2 První model uspořádání

Pokusíme se zde zavést fuzzy relaci v R, která by byla také jistým rozšířením klasické relace uspořádání reálných čísel a umožnila porovnávat kterákoliv dvě fuzzy čísla.

Označme znakem F množinu všech fuzzy čísel na R s riemannovsky integrovatelnými funkcemi věrohodnosti. Na množině $F \times F$ zaveďme nejprve fuzzy relaci $\precsim_1$ funkcí věrohodnosti $\mu_{\prec_1}$ vztahem

$$\mu_{\prec_1}(\underset{\sim}{A}, \underset{\sim}{B}) = \frac{\int_{-\infty}^{x_1^B} \max(\mu_A(x) - L_B(x), 0)dx}{\int_{-\infty}^{\infty} \mu_A(x)dx} \tag{2}$$

a podobně fuzzy relaci $\precsim_2$ funkcí věrohodnosti $\mu_{\precsim_2}$ vztahem

$$\mu_{\prec_2}(\underset{\sim}{B}, \underset{\sim}{A}) = \frac{\int_{x_2^B}^{\infty} \max(\mu_A(x) - P_B(x), 0)dx}{\int_{-\infty}^{\infty} \mu_A(x)dx}, \tag{3}$$

kde $\underset{\sim}{A}, \underset{\sim}{B} \in F$ (pro fuzzy množinu $\underset{\sim}{B}$ užíváme značení z (1)).

Pomocí relací $\precsim_1, \precsim_2$ můžeme stanovit fuzzy relaci $\precsim$ mezi libovolnými dvěma fuzzy čísly z F. K tomu zformulujeme následující definici 2.

Definice 2. Fuzzy relaci $\precsim$ mezi libovolnými dvěma fuzzy čísly $\underset{\sim}{A}, \underset{\sim}{B} \in F$ definujeme funkcí věrohodnosti $\mu_{\prec}$ ve tvaru

$$\mu_{\prec}(\underset{\sim}{A}, \underset{\sim}{B}) = \max(\mu_{\prec_1}(\underset{\sim}{A}, \underset{\sim}{B}) - \mu_{\prec_2}(\underset{\sim}{B}, \underset{\sim}{A}), 0). \tag{4}$$

Poznámka 4. Podmínka integrovatelnosti funkcí věrohodnosti porovnávaných fuzzy čísel $\underset{\sim}{A}, \underset{\sim}{B}$ z předchozí definice 2. podstatně neomezuje třídu

uvažovaných fuzzy čísel. Většinou užíváme trojúhelníková nebo lichoběžníková fuzzy čísla, u nichž tato podmínka je vždy splněna.

Příklad 3. Pro dvojice trojúhelníkových fuzzy čísel $\underset{\sim}{\mathrm{A}}^i$, $\underset{\sim}{\mathrm{B}}^i$, $i = 1, 2, 3$, definovaných takto (viz příklad 2.):

$$\begin{array}{ll} \underset{\sim}{\mathrm{A}}^1: \ a=1, b=2, c=3; & \underset{\sim}{\mathrm{B}}^1: \ a=4, b=5, c=6; \\ \underset{\sim}{\mathrm{A}}^2: \ a=0, b=1,5, c=3; & \underset{\sim}{\mathrm{B}}^2: \ a=1, b=1,5, c=2; \\ \underset{\sim}{\mathrm{A}}^3: \ a=4, b=5, c=6; & \underset{\sim}{\mathrm{B}}^3: \ a=1, b=2, c=3 \end{array}$$

postupně pro příslušné hodnoty věrohodnosti $\mu_{\prec_1}$, $\mu_{\prec_2}$ a $\mu_{\prec}$ dostáváme z (2), (3) a (4) v jednotlivých případech (v souladu s naší intuicí)

$$i=1: \ \ 1;\ 0;\ 1 \quad i=2: \ \ \tfrac{1}{3};\ \tfrac{1}{3};\ 0 \quad i=3: \ \ 0;\ 1;\ 0.$$

Pro vztah $<$ uspořádání se obvykle požaduje, aby stejný význam měl vztah $a < b$ jako $b > a$. Definice 2. však tomuto požadavku nevyhovuje. Užijeme-li analogie s předchozím postupem pro definování fuzzy relace $\precsim$, dostaneme novou fuzzy relaci (odlišnou od fuzzy relace $\precsim$). K definování fuzzy relace $\lesssim$ s požadovanou vlastností budeme však analogicky zavedenou fuzzy relaci $\succsim$ potřebovat. Zaveďme proto nejprve relace $\succsim_1$ a $\succsim_2$ funkcemi věrohodnosti $\mu_{\succ_1}$ a $\mu_{\succ_2}$ vztahy (5) a (6):

$$\mu_{\succ_1}(\underset{\sim}{\mathrm{B}}, \underset{\sim}{\mathrm{A}}) = \frac{\int_{x_2^A}^{\infty} \max(\mu_B(x) - P_A(x), 0)dx}{\int_{-\infty}^{\infty} \mu_B(x)dx} \tag{5}$$

$$\mu_{\succ_2}(\underset{\sim}{\mathrm{A}}, \underset{\sim}{\mathrm{B}}) = \frac{\int_{-\infty}^{x_1^A} \max(\mu_B(x) - L_A(x), 0)dx}{\int_{-\infty}^{\infty} \mu_B(x)dx}. \tag{6}$$

Analogicky s definicí 2. zavedeme nyní fuzzy relaci $\succsim$ na množině $F \times F$ funkcí věrohodností $\mu_{\succ}$.

Definice 3. Fuzzy relaci $\succsim$ mezi libovolnými dvěma fuzzy čísla $\underset{\sim}{\mathrm{A}}, \underset{\sim}{\mathrm{B}} \in F$ definujeme funkcí věrohodnosti $\mu_{\succ}$ ve tvaru

$$\mu_{\succ}(\underset{\sim}{\mathrm{B}}, \underset{\sim}{\mathrm{A}}) = \max(\mu_{\succ_1}(\underset{\sim}{\mathrm{B}}, \underset{\sim}{\mathrm{A}}) - \mu_{\succ_2}(\underset{\sim}{\mathrm{A}}, \underset{\sim}{\mathrm{B}}), 0). \tag{7}$$

Fuzzy relaci, kterou budeme považovat za jistý typ zobecnění klasické relace $<$ (a relace $>$) mezi reálnými čísly na fuzzy čísla, pak zavedeme následující definicí 4.

Definice 4. Na množině $F \times F$ dvojic fuzzy čísel definujeme fuzzy relace $\lesssim$ a $\gtrsim$ funkcemi věrohodnosti tvaru

$$\mu_{<}(\underset{\sim}{\mathrm{A}}, \underset{\sim}{\mathrm{B}}) = 0,5 \cdot (\mu_{\succ}(\underset{\sim}{\mathrm{B}}, \underset{\sim}{\mathrm{A}}) + \mu_{\prec}(\underset{\sim}{\mathrm{A}}, \underset{\sim}{\mathrm{B}})) \tag{8}$$

$$\mu_{>}(\underset{\sim}{\mathrm{A}}, \underset{\sim}{\mathrm{B}}) = 0,5 \cdot (\mu_{\prec}(\underset{\sim}{\mathrm{B}}, \underset{\sim}{\mathrm{A}}) + \mu_{\succ}(\underset{\sim}{\mathrm{A}}, \underset{\sim}{\mathrm{B}})). \tag{9}$$

Poznámka 5. Z definice 4. vyplývá, že pro každé $\underset{\sim}{\mathrm{A}}, \underset{\sim}{\mathrm{B}} \in F$ platí

$$\mu_{<}(\underset{\sim}{\mathrm{A}}, \underset{\sim}{\mathrm{B}}) \in \langle 0;1\rangle; \quad \mu_{>}(\underset{\sim}{\mathrm{A}}, \underset{\sim}{\mathrm{B}}) \in \langle 0;1\rangle \tag{10}$$

$$\mu_{<}(\underset{\sim}{\mathrm{A}}, \underset{\sim}{\mathrm{B}}) = \mu_{>}(\underset{\sim}{\mathrm{B}}, \underset{\sim}{\mathrm{A}}). \tag{11}$$

Příklad 4. Z dat příkladu 3. dostaneme pomocí vztahů (5), (6), (7) postupně pro jednotlivé případy $i = 1, 2, 3$ následující hodnoty

$$i = 1: \;\; 1,\; 0,\; 1; \quad i = 2: \;\; 0,\; 0,\; 0; \quad i = 3: \;\; 0,\; 1,\; 0.$$

Užitím definice 4. pak získáme z výsledku příkladu 3. a tohoto následující hodnoty pro $\mu_<(A^i, B^i)$, $i = 1, 2, 3$:

$$i = 1 : \mu_<(\underset{\sim}{\mathrm{A}}^1,\; \underset{\sim}{\mathrm{B}}^1) = 0,5 \cdot (1 + 1) = 1$$
$$i = 2 : \mu_<(\underset{\sim}{\mathrm{A}}^2,\; \underset{\sim}{\mathrm{B}}^2) = 0,5 \cdot (0 + 0) = 0$$
$$i = 3 : \mu_<(\underset{\sim}{\mathrm{A}}^3,\; \underset{\sim}{\mathrm{B}}^3) = 0,5 \cdot (0 + 0) = 0.$$

S ohledem na poznámku 5. máme také

$$\mu_>(B^1, A^1) = 1, \mu_>(B^2, A^2) = 0, \mu_>(B^3, A^3) = 0.$$

Vše opět v souhlase s naší intuicí.
Následující příklad 5. ozřejmuje interpretaci hodnot příslušnosti relací $\lesssim$ a $\gtrsim$.
Příklad 5. Uvažujme dvě trojúhelníková fuzzy čísla $\underset{\sim}{\mathrm{A}},\; \underset{\sim}{\mathrm{B}} \in F$, definovaná funkcemi věrohodnosti s následujícími parametry

$$\underset{\sim}{\mathrm{A}}: \;\; a = 1,\; b = 2,\; c = 4 \qquad \underset{\sim}{\mathrm{B}}: \;\; a = 0,\; b = 3,\; c = 3,5.$$

Pak pro jednotlivé funkce věrohodnosti máme

$$\mu_{\prec_1}(\underset{\sim}{\mathrm{A}},\; \underset{\sim}{\mathrm{B}}) = 0,10 \qquad \mu_{\succ_1}(\underset{\sim}{\mathrm{B}},\; \underset{\sim}{\mathrm{A}}) \doteq 0,13$$
$$\mu_{\prec_2}(\underset{\sim}{\mathrm{B}},\; \underset{\sim}{\mathrm{A}}) \doteq 0,06 \qquad \mu_{\succ_2}(\underset{\sim}{\mathrm{A}},\; \underset{\sim}{\mathrm{B}}) \doteq 0,14$$
$$\mu_{\prec}(\underset{\sim}{\mathrm{A}},\; \underset{\sim}{\mathrm{B}}) \doteq 0,04 \qquad \mu_{\succ}(\underset{\sim}{\mathrm{B}},\; \underset{\sim}{\mathrm{A}}) = 0$$
$$\mu_<(\underset{\sim}{\mathrm{A}},\; \underset{\sim}{\mathrm{B}}) \doteq 0,5 \cdot (0,04 + 0) = 0,02 \doteq \mu_>(\underset{\sim}{\mathrm{B}},\; \underset{\sim}{\mathrm{A}}).$$

Podobně ale také dostáváme

$$\mu_{\prec_1}(\underset{\sim}{\mathrm{B}},\; \underset{\sim}{\mathrm{A}}) \doteq 0,14 \qquad \mu_{\succ_1}(\underset{\sim}{\mathrm{A}},\; \underset{\sim}{\mathrm{B}}) = 0,05$$
$$\mu_{\prec_2}(\underset{\sim}{\mathrm{A}},\; \underset{\sim}{\mathrm{B}}) = 0,13 \qquad \mu_{\succ_2}(\underset{\sim}{\mathrm{B}},\; \underset{\sim}{\mathrm{A}}) = 0,10$$
$$\mu_{\prec}(\underset{\sim}{\mathrm{B}},\; \underset{\sim}{\mathrm{A}}) = 0,01 \qquad \mu_{\succ}(\underset{\sim}{\mathrm{A}},\; \underset{\sim}{\mathrm{B}}) = 0$$
$$\mu_<(\underset{\sim}{\mathrm{B}},\; \underset{\sim}{\mathrm{A}}) \doteq 0,5 \cdot (0,01 + 0) \doteq 0,005 \doteq \mu_>(\underset{\sim}{\mathrm{A}},\; \underset{\sim}{\mathrm{B}}).$$

Uvedená procedura odhadu vztahu fuzzy čísel $\underset{\sim}{\mathrm{A}}$ a $\underset{\sim}{\mathrm{B}}$ navrhuje chápat fuzzy číslo $\underset{\sim}{\mathrm{A}}$ jako menší než $\underset{\sim}{\mathrm{B}}$ s věrohodností 0,02 a větší s věrohodností jen 0,005.

Připomeňme si, že pro fuzzy čísla $\underset{\sim}{\mathrm{A}}^2,\; \underset{\sim}{\mathrm{B}}^2$ z příkladu 3. platí
$\mu_<(\underset{\sim}{\mathrm{A}}^2,\; \underset{\sim}{\mathrm{B}}^2) = 0 = \mu_>(\underset{\sim}{\mathrm{A}}^2,\; \underset{\sim}{\mathrm{B}}^2)$.

Ukazuje se užitečným pojmenovat třídy všech fuzzy čísel, jejichž míry věrohodnosti fuzzy relací $\lesssim$ a $\gtrsim$ jsou stejné, a rovny 0 (tyto třídy jsou zřejmě třídami ekvivalence). To uskutečníme v definici 5.
Definice 5. Každá dvě fuzzy čísla $\underset{\sim}{\mathrm{A}},\; \underset{\sim}{\mathrm{B}} \in F$ budeme nazývat fuzzy symetrická, když pro ně bude platit:

$$\mu_<(\underset{\sim}{\mathrm{A}},\; \underset{\sim}{\mathrm{B}}) = \mu_>(\underset{\sim}{\mathrm{A}},\; \underset{\sim}{\mathrm{B}}) = 0, \quad \underset{\sim}{\mathrm{A}},\; \underset{\sim}{\mathrm{B}} \in F \tag{12}$$

a fuzzy nesymetrická, když pro ně rovnost (12) nebude platit.

Následující věty popisují vztah dříve zavedených fuzzy relací. Některé dokážeme jen pro trojúhelníková fuzzy čísla.

Věta 1. Nechť $\underset{\sim}{A}, \underset{\sim}{B} \in F$ jsou libovolná fuzzy čísla. Pak je-li $\mu_{\succ_1}(\underset{\sim}{B}, \underset{\sim}{A})=1$, je $\mu_{\succ_2}(\underset{\sim}{A}, \underset{\sim}{B}) = 0$. Podobně, je-li $\mu_{\prec_1}(\underset{\sim}{A}, \underset{\sim}{B}) = 1$, je $\mu_{\prec_2}(\underset{\sim}{B}, \underset{\sim}{A}) = 0$. Obrácená tvrzení neplatí.

Důkaz: Přímo z definic (2) a (3) plyne platnost následujících nerovností pro libovolná fuzzy čísla $\underset{\sim}{A}, \underset{\sim}{B} \in F$

$$\begin{aligned} 0 &\leq \mu_{\prec_1}(\underset{\sim}{A}, \underset{\sim}{B}) + \mu_{\prec_2}(\underset{\sim}{B}, \underset{\sim}{A}) \leq 1 \\ 0 &\leq \mu_{\succ_1}(\underset{\sim}{A}, \underset{\sim}{B}) + \mu_{\succ_2}(\underset{\sim}{B}, \underset{\sim}{A}) \leq 1. \end{aligned}$$

Je-li tedy $\mu_{\prec_1}(\underset{\sim}{A}, \underset{\sim}{B}) = 1$, je $\mu_{\prec_2}(\underset{\sim}{B}, \underset{\sim}{A}) = 0$ a podle (4) také $\mu_{\prec}(\underset{\sim}{A}, \underset{\sim}{B})=1$. Podobně můžeme uvažovat užitím druhé nerovnosti i o fuzzy relaci $\succsim_1$ a $\succsim_2$. Neplatnost obráceného tvrzení vyplývá z předchozích nerovností nebo z příkladu 4 pro $i = 2$.

Věta 2. Nechť $\underset{\sim}{A}, \underset{\sim}{B} \in F$ jsou libovolná fuzzy čísla. Pak je-li $\mu_{\succ_1}(\underset{\sim}{B}, \underset{\sim}{A})=1$, je i $\mu_{\prec_1}(\underset{\sim}{A}, \underset{\sim}{B}) = 1$.

Důkaz: Nechť c je krajní bod intervalu, pro který platí

$$\begin{aligned} 0 < P_A(x) < 1 \qquad & \text{v } (x_2^A, c), \quad A \in F \\ P_A(x) = 0 \qquad & \text{pro } x \notin (x_2^A, c). \end{aligned}$$

Pro každou fuzzy množinu $\underset{\sim}{A}$ označujeme znakem $\mathrm{Supp}(\underset{\sim}{A})$ množinu $\mathrm{Supp}(\underset{\sim}{A}) = \{x \in R; \mu_A(x) > 0\}$. Z podmínky věty plyne, že $\mathrm{Supp}(\underset{\sim}{B})$ je interval, a proto musí být $c < +\infty$ a platí $\mathrm{Supp}(\underset{\sim}{B}) \subset (c, +\infty)$. Proto $\mathrm{Supp}(B) \cap \mathrm{Supp}(\underset{\sim}{A}) = \emptyset$ (nebo nejvýše jednobodová) a platí

$$\begin{aligned} \int_{-\infty}^{x_1^B} \max(\mu_A(x) - L_B(x), 0)dx &= \int_{-\infty}^{c} \max(\mu_A(x) - L_B(x), 0)dx = \\ &= \int_{-\infty}^{c} \mu_A(x)dx = \int_{-\infty}^{\infty} \mu_A(x)dx. \end{aligned}$$

Takže opravdu $\mu_{\prec_1}(\underset{\sim}{A}, \underset{\sim}{B}) = 1$.

Je jasné, že pak i podobně z $\mu_{\prec_1}(\underset{\sim}{B}, \underset{\sim}{A}) = 1$ plyne i $\mu_{\succ_1}(\underset{\sim}{B}, \underset{\sim}{A}) = 1$.

Platnost ostré tranzitivity mezi fuzzy čísly vzhledem k fuzzy relaci $\precsim_1$ vysvětluje následující věta.

Věta 3. Nechť $\underset{\sim}{A}, \underset{\sim}{B}, \underset{\sim}{C}$ jsou fuzzy čísla z množiny F. Je-li $\mu_{\prec_1}(\underset{\sim}{A}, \underset{\sim}{B}) = 1$ a $\mu_{\prec_1}(\underset{\sim}{B}, \underset{\sim}{C}) = 1$, pak i $\mu_{\prec_1}(\underset{\sim}{A}, \underset{\sim}{C}) = 1$.

$$\begin{aligned} &\mathrm{Supp}(\underset{\sim}{A}) \cap \mathrm{Supp}(\underset{\sim}{B}) = \emptyset \text{ (nebo nejvýše jednobodová)} \\ &\forall x \in \mathrm{Supp}(\underset{\sim}{A}) \forall y \in \mathrm{Supp}(\underset{\sim}{B}); x \leq y \\ &\mathrm{Supp}(\underset{\sim}{B}) \cap \mathrm{Supp}(\underset{\sim}{C}) = \emptyset \text{ (nebo nejvýše jednobodová)} \\ &\forall y \in \mathrm{Supp}(\underset{\sim}{B}) \forall z \in \mathrm{Supp}(\underset{\sim}{C}); y \leq z. \\ &\forall x \in \mathrm{Supp}(\underset{\sim}{A}) \forall z \in \mathrm{Supp}(\underset{\sim}{C}); x \leq z \\ &\mathrm{Supp}(\underset{\sim}{A}) \cap \mathrm{Supp}(\underset{\sim}{C}) = \emptyset \text{ (nebo nejvýše jednobodová)} \end{aligned}$$

a tedy $\mu_{\prec_1}(\underset{\sim}{A}, \underset{\sim}{C}) = 1$.

Poznámka 7. Zřejmě také platí podobné tvrzení i pro fuzzy relaci $\prec_2$:
Je-li $\mu_{\prec_2}(\underset{\sim}{A}, \underset{\sim}{B}) = 1$ a $\mu_{\prec_2}(\underset{\sim}{B}, \underset{\sim}{C}) = 1$, je i $\mu_{\prec_2}(\underset{\sim}{A}, \underset{\sim}{C}) = 1$.
Následující věta konstatuje ostrou tranzitivitu relace $\precsim$ na F.
Věta 4. Nechť $\underset{\sim}{A}, \underset{\sim}{B}, \underset{\sim}{C}$ jsou fuzzy čísla z množiny F.
Pak z $\mu_{\prec}(\underset{\sim}{A}, \underset{\sim}{B}) = 1$ a $\mu_{\prec}(\underset{\sim}{B}, \underset{\sim}{C}) = 1$ plyne $\mu_{\prec}(\underset{\sim}{A}, \underset{\sim}{C}) = 1$.
Důkaz: Nechť $\underset{\sim}{A}, \underset{\sim}{B}, \underset{\sim}{C} \in F$, pak platí přímo z definice (4)

$$\begin{aligned} \mu_{\prec}(\underset{\sim}{A}, \underset{\sim}{B}) &= 1 \Rightarrow \mu_{\prec_1}(\underset{\sim}{A}, \underset{\sim}{B}) = 1 \\ \mu_{\prec}(\underset{\sim}{B}, \underset{\sim}{C}) &= 1 \Rightarrow \mu_{\prec_1}(\underset{\sim}{B}, \underset{\sim}{C}) = 1. \end{aligned}$$

Z věty 3. pak také plyne $\mu_{\prec_1}(\underset{\sim}{A}, \underset{\sim}{C}) = 1$. Přímo z definice fuzzy relace $\precsim$ a věty 1 pak z poslední rovnosti pro relaci $\precsim_1$ dostáváme $\mu_{\prec}(\underset{\sim}{A}, \underset{\sim}{C}) = 1$.
Poznámka 8. Podobné tvrzení jako ve větě 4. platí i pro fuzzy relaci $\succsim$:
Jsou-li $\underset{\sim}{A}, \underset{\sim}{B}, \underset{\sim}{C} \in F$, pak z $\mu_{\succ}(\underset{\sim}{A}, \underset{\sim}{B}) = 1$ a $\mu_{\succ}(\underset{\sim}{B}, \underset{\sim}{C}) = 1$ platí, že i $\mu_{\succ}(\underset{\sim}{A}, \underset{\sim}{C}) = 1$.

Nyní ukážeme, že i pro fuzzy relaci $\lesssim$ mezi fuzzy čísly z F platí klasická tranzitivita.
Věta 5. Nechť $\underset{\sim}{A}, \underset{\sim}{B}, \underset{\sim}{C}$ jsou libovolná fuzzy čísla z množiny F. Je-li $\mu_{<}(\underset{\sim}{A}, \underset{\sim}{B}) = 1$, $\mu_{<}(\underset{\sim}{B}, \underset{\sim}{C}) = 1$, je i $\mu_{<}(\underset{\sim}{A}, \underset{\sim}{C}) = 1$.
Důkaz: Nechť $\underset{\sim}{A}, \underset{\sim}{B}, \underset{\sim}{C} \in F$. Pak

$$\begin{aligned} \mu_{<}(\underset{\sim}{A}, \underset{\sim}{B}) = 1 &\Rightarrow \mu_{\succ}(\underset{\sim}{B}, \underset{\sim}{A}) = 1 \wedge \mu_{\prec}(\underset{\sim}{A}, \underset{\sim}{B}) = 1 \\ \mu_{<}(\underset{\sim}{B}, \underset{\sim}{C}) = 1 &\Rightarrow \mu_{\succ}(\underset{\sim}{C}, \underset{\sim}{B}) = 1 \wedge \mu_{\prec}(\underset{\sim}{B}, \underset{\sim}{C}) = 1 \end{aligned}$$

a užitím věty 4. máme $\mu_{\succ}(\underset{\sim}{C}, \underset{\sim}{A}) = 1$ a $\mu_{\prec}(\underset{\sim}{A}, \underset{\sim}{C}) = 1$.
Z toho ihned podle definice fuzzy relace $\lesssim$ máme $\mu_{<}(\underset{\sim}{A}, \underset{\sim}{C}) = 1$.

V příkladu 5. jsme viděli, že míry věrohodnosti $\mu_{<}(\underset{\sim}{A}, \underset{\sim}{B})$ a $\mu_{>}(\underset{\sim}{A}, \underset{\sim}{B})$ dávají rozdílné výsledky. To v některých aplikacích způsobuje problém s interpretací (kdy se například očekává jednoznačné doporučení). Interpretaci si pak usnadníme zavedením totálního fuzzy uspořádání (jak je ukázáno v následující definici).
Definice 5. Nechť $\underset{\sim}{A}, \underset{\sim}{B}$ jsou libovolná dvě fuzzy čísla z množiny F. Totální fuzzy uspořádání dané funkcí věrohodnosti $\mu^T_{<}$ určíme ze vztahu

$$\mu^T_{<}(\underset{\sim}{A}, \underset{\sim}{B}) = \max(\mu_{<}(\underset{\sim}{A}, \underset{\sim}{B}) - \mu_{>}(\underset{\sim}{A}, \underset{\sim}{B}), 0). \tag{13}$$

Poznámka 9. Je-li $\mu^T_{<}(\underset{\sim}{A}, \underset{\sim}{B}) > 0$, je $\mu^T_{<}(\underset{\sim}{B}, \underset{\sim}{A}) = 0$. Pro fuzzy symetrická fuzzy čísla $\underset{\sim}{A}, \underset{\sim}{B} \in F$ je $\mu^T_{<}(\underset{\sim}{A}, \underset{\sim}{B}) = 0 = \mu^T_{>}(\underset{\sim}{A}, \underset{\sim}{B})$.
Poznámka 10. Podobně jako v definici 5. jsme určili totální fuzzy relaci uspořádání $\lesssim^T$, můžeme určit i totální fuzzy relaci uspořádání $\gtrsim^T$:
Pro libovolná fuzzy čísla $\underset{\sim}{A}, \underset{\sim}{B} \in F$ volíme

$$\mu^T_{>}(\underset{\sim}{A}, \underset{\sim}{B}) = \max(\mu_{>}(\underset{\sim}{A}, \underset{\sim}{B}) - \mu_{<}(\underset{\sim}{A}, \underset{\sim}{B}), 0). \tag{14}$$

Věta 6. Nechť $\underset{\sim}{A}, \underset{\sim}{B} \in F$ jsou libovolná fuzzy čísla. Pak

$$\mu^T_<(\underset{\sim}{B}, \underset{\sim}{A}) = \mu^T_>(\underset{\sim}{A}, \underset{\sim}{B}). \tag{15}$$

Důkaz: Postupně platí

$$\begin{aligned}\mu^T_<(\underset{\sim}{B}, \underset{\sim}{A}) &= \max(\mu_<(\underset{\sim}{B}, \underset{\sim}{A}) - \mu_>(\underset{\sim}{B}, \underset{\sim}{A}), 0) = \\ &= \max(\mu_>(\underset{\sim}{A}, \underset{\sim}{B}) - \mu_<(\underset{\sim}{A}, \underset{\sim}{B}), 0) = \mu^T_>(\underset{\sim}{A}, \underset{\sim}{B}).\end{aligned}$$

Věta 7. Fuzzy relace $\lesssim$ (resp. $\gtrsim$) na F je antireflexivní, tj.

$$\mu_<(\underset{\sim}{A}, \underset{\sim}{A}) = 0 \quad (\text{resp. } \mu_>(\underset{\sim}{A}, \underset{\sim}{A}) = 0) \tag{16}$$

pro každé $\underset{\sim}{A} \in F$

Důkaz: Pro každé $\underset{\sim}{A} \in F$ podle (8), (4) a (7) platí

$$\begin{aligned}\mu_<(\underset{\sim}{A}, \underset{\sim}{A}) &= 0,5 \cdot (\mu_\succ(\underset{\sim}{A}, \underset{\sim}{A}) + \mu_\prec(\underset{\sim}{A}, \underset{\sim}{A})) = \\ &= 0,5 \cdot (\max(\mu_{\succ_1}(\underset{\sim}{A}, \underset{\sim}{A}) - \mu_{\succ_2}(\underset{\sim}{A}, \underset{\sim}{A}), 0) + \\ &\quad \max(\mu_{\prec_1}(\underset{\sim}{A}, \underset{\sim}{A}) - \mu_{\prec_2}(\underset{\sim}{A}, \underset{\sim}{A}), 0) = \\ &= 0,5 \cdot (0 + 0) = 0.\end{aligned}$$

Z věty 7. přímo také vyplývá antireflexivita fuzzy relací $\gtrsim^T$, $\lesssim^T$:

$$\begin{aligned}\mu^T_>(\underset{\sim}{A}, \underset{\sim}{A}) &= \max(\mu_>(\underset{\sim}{A}, \underset{\sim}{A}) - \mu_<(\underset{\sim}{A}, \underset{\sim}{A}), 0) = \\ &= \max(0 - 0, 0) = 0 = \mu^T_<(\underset{\sim}{A}, \underset{\sim}{A}).\end{aligned}$$

Poznámka 11. Fuzzy relaci rovnosti $\underset{\sim}{=}$ mezi fuzzy čísly na F je možné definovat pomocí její funkce věrohodnosti ve tvaru

$$\mu_=(\underset{\sim}{A}, \underset{\sim}{B}) = \max(1 - \mu_<(\underset{\sim}{A}, \underset{\sim}{B}) - \mu_>(\underset{\sim}{A}, \underset{\sim}{B}), 0) \tag{17}$$

pro libovolná $\underset{\sim}{A}, \underset{\sim}{B} \in F$.

Následující příklad 6. ukazuje, že tranzitivita fuzzy relace $\prec_1$ mezi fuzzy čísly $\underset{\sim}{A}, \underset{\sim}{B}, \underset{\sim}{C} \in F$, definovaná podmínkou

$$\min(\mu_{\prec_1}(\underset{\sim}{A}, \underset{\sim}{B}), \mu_{\prec_1}(\underset{\sim}{B}, \underset{\sim}{C})) \le \mu_{\prec_1}(\underset{\sim}{A}, \underset{\sim}{C}) \tag{18}$$

obecně neplatí.

Příklad 6. Mějme tři trojúhelníková fuzzy čísla $\underset{\sim}{A}, \underset{\sim}{B}, \underset{\sim}{C}$, definovaná svými parametry a, b, c takto:

$\underset{\sim}{A}: \ a_1 = 1, b_1 = 2, c_1 = 3 \qquad \underset{\sim}{B}: \ a_2 = z, b_2 = 4, c_5 = 5, z \ \in R$

$\underset{\sim}{C}: \ a_3 = 0, b_3 = 3, c_3 = 4.$

Pak pro $z \in (-1,63; -1,33)$ vztah (15) neplatí. Například pro $z = -1,5$ dostáváme výpočtem

$$\mu_{\prec_1}(\underset{\sim}{\mathrm{A}}, \underset{\sim}{\mathrm{B}}) \doteq 0,137 \quad \mu_{\prec_1}(\underset{\sim}{\mathrm{B}}, \underset{\sim}{\mathrm{C}}) \doteq 0,128 \quad \mu_{\prec_1}(\underset{\sim}{\mathrm{A}}, \underset{\sim}{\mathrm{C}}) \doteq 0,125$$

a je tedy $\min(\mu_{\prec_1}(\underset{\sim}{\mathrm{A}}, \underset{\sim}{\mathrm{B}}), \mu_{\prec_1}(\underset{\sim}{\mathrm{B}}, \underset{\sim}{\mathrm{C}})) = \min(0,137; 0,128) > 0,125 \doteq \mu_{\prec_1}(\underset{\sim}{\mathrm{A}}, \underset{\sim}{\mathrm{C}})$.

Poznámka 12. Obdoba vztahu (16) pro fuzzy relace $\prec_2$, $\succ_1$ a $\succ_2$ pak také obecně neplatí.

Míry věrohodnosti pro všechny zde uvažované fuzzy relace jsou založeny na jistých integrálech věrohodností fuzzy čísel. S ohledem na co nejlepší souhlas naší intuice a zkušenosti s formulovanými relacemi je třeba dříve zavedené definice vztahů někdy ještě poněkud opravit (nebo alespoň vyslovit možnost takové opravy). Jaký vliv pak na míru věrohodnosti fuzzy relací může taková oprava mít, ukážeme na příkladě.

Definice 6. Řekneme, že fuzzy číslo $\underset{\sim}{\mathrm{A}} \in F$ normujeme na fuzzy množinu $\underset{\sim}{\mathrm{A}}^n$ s funkcí věrohodnosti μ_{A^n}, když $\int_{-\infty}^{\infty} \mu_A(x)dx \neq 0$ a volíme

$$\mu_{A^n}(x) = \frac{\mu_A(x)}{\int_{-\infty}^{\infty} \mu_A(x)dx}, \quad x \in R. \tag{19}$$

Poznámka 13. Normováním trojúhelníkového fuzzy čísla $\underset{\sim}{\mathrm{A}}$ normujeme jeho pravou i levou stranu; množina $\mathrm{Supp}(\underset{\sim}{\mathrm{A}})$ se normováním nezmění pro jakékoliv fuzzy číslo $\underset{\sim}{\mathrm{A}} \in F$

$$\mathrm{Supp}(A^n) = \mathrm{Supp}(\underset{\sim}{\mathrm{A}}). \tag{20}$$

Převedeme-li výrazy (2), (3), (5), (6) v našich definicích tak, aby všude, kde se vyskytuje míra věrohodnosti některého trojúhelníkového fuzzy čísla (nebo její restrikce) byla nahrazena podílem, ve kterém v čitateli zůstává původní míra a ve jmenovateli je hodnota integrálu z této míry přes R, dostaneme normalizované míry ve tvaru, např.

$$\mu^n_{\prec_1}(\underset{\sim}{\mathrm{A}}, \underset{\sim}{\mathrm{B}}) = \int_{-\infty}^{b_2} \max\left(\frac{\mu_A(x)}{A} - \frac{L_B(x)}{B}, 0\right) dx \tag{21}$$

$$A = \frac{c_1 - a_1}{2}; \quad B = \frac{c_2 - a_2}{2}. \tag{22}$$

Poznámka 14. V množině F jsou pro trojúhelníková fuzzy čísla ekvivalentní následující dvojice tvrzení:

$$\begin{aligned}
\underset{\sim}{\mathrm{B}} \geq \underset{\sim}{\mathrm{A}}: \quad & \mu_{\prec_1}(\underset{\sim}{\mathrm{A}}, \underset{\sim}{\mathrm{B}} \geq 0 \ \wedge \mu_{\prec_2}(\underset{\sim}{\mathrm{B}}, \underset{\sim}{\mathrm{A}}) = 0 \wedge \\
& \wedge \mu_{\succ_1}(\underset{\sim}{\mathrm{B}}, \underset{\sim}{\mathrm{A}}) \geq 0 \wedge \mu_{\succ_2}(\underset{\sim}{\mathrm{A}}, \underset{\sim}{\mathrm{B}}) = 0; \\
\underset{\sim}{\mathrm{B}} \leq \underset{\sim}{\mathrm{A}}: \quad & \mu_{\prec_1}(\underset{\sim}{\mathrm{A}}, \underset{\sim}{\mathrm{B}} = 0 \wedge \mu_{\prec_2}(\underset{\sim}{\mathrm{B}}, \underset{\sim}{\mathrm{A}}) \geq 0 \wedge \\
& \wedge \mu_{\succ_1}(\underset{\sim}{\mathrm{B}}, \underset{\sim}{\mathrm{A}}) = 0 \wedge \mu_{\succ_2}(\underset{\sim}{\mathrm{A}}, \underset{\sim}{\mathrm{B}}) \geq 0.
\end{aligned}$$

Nesplnění podmínky tranzitivity (18) pro fuzzy relaci $\precsim_1$ (resp. $\succsim_1$, $\precsim_2$, $\succsim_2$) na třídě všech fuzzy čísel z F omezuje silně možnosti aplikací. Protože $\precsim_1$ není fuzzy relací uspořádání splňujícího (18), zůstává mnohdy nerozřešen i problém určení optimální varianty ze tří možných. Otázkou je, jak asi často se to může stát v případě, že budeme používat trojúhelníková fuzzy čísla.

Příklad 7. Normujeme-li trojúhelníková fuzzy čísla $\underset{\sim}{\mathrm{A}}$, $\underset{\sim}{\mathrm{B}}$, $\underset{\sim}{\mathrm{C}}$ z příkladu 6., dostaneme tak normalizované míry

$$\mu^n_{\prec_1}(\underset{\sim}{\mathrm{A}}, \underset{\sim}{\mathrm{B}}) = 0,649 \quad \mu^n_{\prec}(\underset{\sim}{\mathrm{A}}, \underset{\sim}{\mathrm{C}}) = 0,457 \quad \mu^n_{\prec_1}(\underset{\sim}{\mathrm{B}}, \underset{\sim}{\mathrm{C}}) = 0,095.$$

Uvedená změna hodnot funkcí věrohodnosti způsobila, že pro uvažovaná fuzzy čísla a normalizovanou normu platí (16).

Nechť libovolná tři trojúhelníková fuzzy čísla $\underset{\sim}{\mathrm{A}}_1$, $\underset{\sim}{\mathrm{A}}_2$, $\underset{\sim}{\mathrm{A}}_3$ jsou určena postupně trojicemi čísel a_i, b_i, c_i, $a_i \le b_i \le c_i$, $i = 1, 2, 3$. Pak když pro ně platí (vzhledem k jejich průsekům resp. spojením)

$$\underset{\sim}{\mathrm{A}}_1 \le \underset{\sim}{\mathrm{A}}_2 \le \underset{\sim}{\mathrm{A}}_3, \tag{23}$$

je

$$a_1 \le a_2 \le a_3; \quad b_1 \le b_2 \le b_3; \quad c_1 \le c_2 \le c_3. \tag{24}$$

Platí ale také obráceně, jsou-li tři trojúhelníková fuzzy čísla $\underset{\sim}{\mathrm{A}}_1$, $\underset{\sim}{\mathrm{A}}_2$, $\underset{\sim}{\mathrm{A}}_3$ vázána podmínkou (24), platí pro ně pak (23). (Je to důsledek věty 1.10 v [2].)

Platí-li však pro libovolná tři trojúhelníková fuzzy čísla $\underset{\sim}{\mathrm{A}}_1$, $\underset{\sim}{\mathrm{A}}_2$, $\underset{\sim}{\mathrm{A}}_3$ podmínka (23), resp. (24), je splněna také podmínka fuzzy tranzitivity pro $\prec_1$ ($\prec_2$ i $\prec$). To vyplývá z toho, že za uvedených podmínek platí vždy

$$\int_{-\infty}^{b_2} \max(\mu_{A_1}(x) - L_{A_2}(x), 0)dx \le \int_{-\infty}^{b_3} \max(\mu_{A_1}(x) - L_{A_3}(x), 0)dx.$$

S ohledem na příklad 6. vidíme, že pojem fuzzy tranzitivity uvedených fuzzy relací je širší (splňuje ji rozsáhlejší třída fuzzy čísel). To je příznivá informace pro případné aplikace této metody porovnávání fuzzy čísel.

Poznámka 15. Jsou-li dvě libovolná trojúhelníková fuzzy čísla $\underset{\sim}{\mathrm{A}}_1$, $\underset{\sim}{\mathrm{A}}_2$ určena trojicemi čísel a_i, b_i, c_i, $a_i \le b_i \le c_i$, kde navíc je $b_i - a_i = c_i - b_i$, $i = 1, 2$, platí vždy buď $\underset{\sim}{\mathrm{A}}_1 \le \underset{\sim}{\mathrm{A}}_2$ nebo $\underset{\sim}{\mathrm{A}}_2 \le \underset{\sim}{\mathrm{A}}_1$.

Příklad 8. Byl sledován vliv pěti různých preparátů na léčbu určitého stadia jisté choroby. Úspěšnost léčby daným preparátem byla hodnocena na 11-bodové stupnici $0, 1, 2, \ldots, 10$. Přitom hodnota 10 měla označovat optimální úspěšnost, hodnota 0 nejnižší úroveň úspěšnosti. Pro každý preparát se výsledky šetření vyhodnocovaly tak, že se ze všech expertních odhadů, tvořících množinu S, vyhledala jak minimální, tak i maximální bodová hodnota a medián (případně modus). Každému preparátu tak byla přiřazena úspěšnost léčby jako trojúhelníkové fuzzy číslo, určené trojicí $a = \min\{S\}$,

$b = \mathit{medián}\{S\}$, $c = \max\{S\}$. Bylo dohodnuto, že ze dvou možných preparátů P_i, P_j, $i \neq j$, s úspěšnostmi $\underset{\sim}{\mathrm{P}}_i$, $\underset{\sim}{\mathrm{P}}_j$ bude považován za lepší preparát P_j když $\mu^T_<(\underset{\sim}{\mathrm{P}}_i, \underset{\sim}{\mathrm{P}}_j) > 0$.

Výsledky šetření pro preparáty $P_1, P_2, \cdots, P_5$ byly zaznamenány do tabulky 1. Vypočtené míry věrohodnosti fuzzy relace $\lesssim^T$ jsou obsahem tabulky 2.

preparát	a $\min\{S\}$	b $\mathit{medián}\{S\}$	c $\max\{S\}$
P_1	3	5	8
P_2	4	6	7
P_3	7	8	10
P_4	1	3	6
P_5	5	8	9

Tabulka 1:

$\lesssim^T$	$\underset{\sim}{\mathrm{P}}_1$	$\underset{\sim}{\mathrm{P}}_2$	$\underset{\sim}{\mathrm{P}}_3$	$\underset{\sim}{\mathrm{P}}_4$	$\underset{\sim}{\mathrm{P}}_5$
$\underset{\sim}{\mathrm{P}}_1$	0	0,14	0,94	0	0,66
$\underset{\sim}{\mathrm{P}}_2$	0	0	1,00	0	0,71
$\underset{\sim}{\mathrm{P}}_3$	0	0	0	0	0
$\underset{\sim}{\mathrm{P}}_4$	0,64	0,80	1,00	0	0,96
$\underset{\sim}{\mathrm{P}}_5$	0	0	0,40	0	0

Tabulka 2:

Z tabulky Tab. 2. byly sestaveny následující řetězce úrovně úspěšnosti zkoumaných preparátů: P_4, P_1, P_2, P_3 a dále P_4, P_1, P_2, P_5 a P_5, P_3. Nejúspěšnější se jeví pak preparát P_3.

Průsek reprezentujících fuzzy množin nám nabízí informaci o vztahu $\leq$ úspěšností léčby, uvedenou v tabulce Tab. 3. Zde znakem „?“ registrujeme neporovnatelnost úrovní úspěšnosti pro preparáty P_1 a P_2.

$\leq$	$\underset{\sim}{\mathrm{P}}_1$	$\underset{\sim}{\mathrm{P}}_2$	$\underset{\sim}{\mathrm{P}}_3$	$\underset{\sim}{\mathrm{P}}_4$	$\underset{\sim}{\mathrm{P}}_5$
$\underset{\sim}{\mathrm{P}}_1$	+	?	+	-	+
$\underset{\sim}{\mathrm{P}}_2$	?	+	+	-	+
$\underset{\sim}{\mathrm{P}}_3$	-	-	+	-	-
$\underset{\sim}{\mathrm{P}}_4$	+	+	+	+	+
$\underset{\sim}{\mathrm{P}}_5$	-	-	+	-	+

Tabulka 3:

Vynecháním výsledku pro preparát P_2 (resp. P_1) dostaneme již úplné uspořádání ve zbývající množině fuzzy množin úspěšností: $\underset{\sim}{\mathrm{P}}_4$, $\underset{\sim}{\mathrm{P}}_1$, $(\underset{\sim}{\mathrm{P}}_2)$, $\underset{\sim}{\mathrm{P}}_5$, $\underset{\sim}{\mathrm{P}}_3$.

Někdy je vhodné uspořádat trojúhelníková fuzzy čísla pomocí x-ových souřadnic těžišť trojúhelníků, tvořených grafem nenulových hodnot funkce věrohodnosti příslušného fuzzy čísla $\underset{\sim}{\mathrm{A}}$ a jeho Supp($\underset{\sim}{\mathrm{A}}$). V našem případě bychom tak získali následující úplné uspořádání fuzzy množin pěti výsledných úspěšností preparátů (v závorkách je uvedena x-ová souřadnice těžiště, zaokrouhlená na jedno desetinné místo):

$$\underset{\sim}{\mathrm{P}}_4(3,3),\ \underset{\sim}{\mathrm{P}}_1(5,3),\ \underset{\sim}{\mathrm{P}}_2(5,7),\ \underset{\sim}{\mathrm{P}}_5(7,3),\ \underset{\sim}{\mathrm{P}}_3(8,3).$$

Tento typ uspořádání vždy splňuje podmínku tranzitivnosti (a je úplné).

3 Druhý model uspořádání

Pro porovnání dvou fuzzy čísel se nabízí ještě jedna cesta, analogická k definování uspořádání reálných čísel. Fuzzy číslo $\underset{\sim}{A} \in F$ nazveme *kladné* (resp. *záporné*), když $\mathrm{Sup}(\underset{\sim}{A}) \subset (0, +\infty)$ (resp. $\mathrm{Supp}(\underset{\sim}{A}) \subset (-\infty, 0)$). Není-li fuzzy číslo $\underset{\sim}{A}$ kladné ani záporné, nazveme ho *nulové* ([4]). Mějme nyní dvě libovolná fuzzy čísla $\underset{\sim}{A}, \underset{\sim}{B} \in F$ a určeme fuzzy číslo $\underset{\sim}{C} = \underset{\sim}{B} + (-\underset{\sim}{A})$. Pro míru věrohodnosti tohoto fuzzy čísla $\underset{\sim}{C}$ pak platí podle principu rozšíření

$$\mu_C(z) = \sup_x \min(\mu_A(x), \mu_B(x+z)) \tag{25}$$

a je pak i $\underset{\sim}{C} \in F$. Indikátor vztahu $\underset{\sim}{B} > \underset{\sim}{A}$ (označme ho $\mathrm{Pref}(\underset{\sim}{B}, \underset{\sim}{A})$) pak definujeme pro všechna $\underset{\sim}{A}, \underset{\sim}{B} \in F$ pomocí míry věrohodnosti μ_C takto

$$\mathrm{Pref}(\underset{\sim}{B}, \underset{\sim}{A}) = \frac{\int_0^\infty \mu_C(z)dz}{\int_{-\infty}^\infty \mu_C(z)dz}, \quad \text{když} \quad \int_{-\infty}^{\infty} \mu_c(z)dz \neq 0. \tag{26}$$

Zřejmě je $\mathrm{Pref} \in \langle 0; 1\rangle$. Dále platí $\mathrm{Pref}(\underset{\sim}{B}, \underset{\sim}{A}) = 1 - \mathrm{Pref}(\underset{\sim}{A}, \underset{\sim}{B})$ a z toho $\mathrm{Pref}(\underset{\sim}{A}, \underset{\sim}{A}) = 0,5$. Je-li např. $\mathrm{Pref}(\underset{\sim}{B}, \underset{\sim}{A}) > 0,5$, můžeme to označit $\underset{\sim}{B} >^P \underset{\sim}{A}$ (obrácené tvrzení nemusí platit).

Znakem $\underset{\sim}{B} >^K \underset{\sim}{A}$ pro libovolná dvě fuzzy čísla $\underset{\sim}{A}, \underset{\sim}{B}$ můžeme označit situaci, kdy fuzzy číslo $\underset{\sim}{B} - \underset{\sim}{A}$ bude kladné. (Zřejmě tedy platí $\underset{\sim}{B} >^K \underset{\sim}{A} \rightarrow \underset{\sim}{B} >^P \underset{\sim}{A}$ pro libovolná dvě fuzzy čísla $\underset{\sim}{A}, \underset{\sim}{B} \in F$.)

Pro praxi je důležitá otázka, zda vztahy $>^P, >^K$ splňují podmínku tranzitivity:

$$((\underset{\sim}{A} > \underset{\sim}{B}) \wedge (\underset{\sim}{B} > \underset{\sim}{C})) \rightarrow (\underset{\sim}{A} > \underset{\sim}{C}).$$

Uvažujme nyní jen gaussovská fuzzy čísla; míra věrohodnosti gaussovského fuzzy čísla $\underset{\sim}{A}$ je určena dvojicí parametrů a, σ_a ve tvaru (27)

$$\mu_A(x) = \exp\left(-\frac{(x-a)^2}{2\sigma_a^2}\right), \tag{27}$$

kde a je libovolné reálné číslo, $\sigma_a > 0$. Vzhledem k uvedené definici platí pro každé gaussovské fuzzy číslo $\underset{\sim}{A} \in F$.

Máme-li dvě gaussovská fuzzy čísla $\underset{\sim}{A}, \underset{\sim}{B}$, určená postupně parametry $a, \sigma_a > 0, b, \sigma_b > 0$, pak fuzzy číslo $\underset{\sim}{C} = \underset{\sim}{B} + (-\underset{\sim}{A})$ je podle principu rozšíření opět gaussovské, určené parametry $b - a$, $\sqrt{\sigma_a^2 + \sigma_b^2}$. [1]

[1] Uvedené tvrzení vyplývá snadno z následujících vztahů a jejich úpravy:

$$\begin{aligned} \exp\left(-\frac{(x-a)^2}{2\sigma_a^2}\right) &= \exp\left(-\frac{(y-b)^2}{2\sigma_b^2}\right) = \alpha; \quad 0 < \alpha \leq 1; \\ |x-a| &= \sigma_a \cdot \sqrt{-2\ln\alpha}, |y-b| = \sigma_b \cdot \sqrt{-2\ln\alpha}; \end{aligned}$$

Pro hodnotu výrazu Pref($\underset{\sim}{\mathrm{B}}$, $\underset{\sim}{\mathrm{A}}$) pak podle (26) dostaneme (28)

$$\mathrm{Pref}(\underset{\sim}{\mathrm{B}}, \underset{\sim}{\mathrm{A}}) = \Phi\left(\frac{b-a}{\sqrt{\sigma_a^2+\sigma_b^2}}\right). \tag{28}$$

Předpokládejme pro tři gaussovská fuzzy čísla platnost vztahů $\underset{\sim}{\mathrm{B}} >^P \underset{\sim}{\mathrm{A}}$, $\underset{\sim}{\mathrm{C}} >^P \underset{\sim}{\mathrm{B}}$ a zjišťujme, zda pak také platí $\underset{\sim}{\mathrm{C}} >^P \underset{\sim}{\mathrm{A}}$. Z definice vztahu $>^P$ máme následující vzájemně si odpovídající podmínky:

$$\underset{\sim}{\mathrm{B}} >^P \underset{\sim}{\mathrm{A}} \equiv \Phi\left(\frac{b-a}{\sqrt{\sigma_a^2+\sigma_b^2}}\right) > 0,5 \; a \; \underset{\sim}{\mathrm{C}} >^P \underset{\sim}{\mathrm{B}} \equiv \Phi\left(\frac{c-b}{\sqrt{\sigma_c^2+\sigma_b^2}}\right) > 0,5.$$

a tedy $b-a>0$, $c-b>0$ a $c-a>0$. To je však ekvivalentní tvrzení $\underset{\sim}{\mathrm{C}} >^P \underset{\sim}{\mathrm{A}}$. Relace $>^P$ je tedy na množině všech gaussovských fuzzy čísel tranzitivní.

Příklad 9. Mějme tři gaussovská fuzzy čísla $\underset{\sim}{\mathrm{A}}$, $\underset{\sim}{\mathrm{B}}$, $\underset{\sim}{\mathrm{C}}$, určená následujícími hodnotami svých parametrů

$$\underset{\sim}{\mathrm{A}} : a=0; \sigma_a=1 \quad \underset{\sim}{\mathrm{B}} : b=1; \sigma_b=2 \quad \underset{\sim}{\mathrm{C}} : c=0; \sigma_c=3$$

Výpočtem a pomocí tabulek hodnot funkce ϕ dostaneme

$$\begin{aligned}
\mathrm{Pref}(\underset{\sim}{\mathrm{B}}, \underset{\sim}{\mathrm{A}}) &= \phi\left(\frac{1}{\sqrt{1+2^2}}\right) \doteq 0,67 \\
\mathrm{Pref}(\underset{\sim}{\mathrm{C}}, \underset{\sim}{\mathrm{B}}) &= \phi\left(\frac{0,5}{\sqrt{2^2+3^2}}\right) \doteq 0,55 \\
\mathrm{Pref}(\underset{\sim}{\mathrm{C}}, \underset{\sim}{\mathrm{A}}) &= \phi\left(\frac{1,5}{\sqrt{3^2+1^2}}\right) \doteq 0,68.
\end{aligned}$$

Tedy $\underset{\sim}{\mathrm{B}} >^P \underset{\sim}{\mathrm{A}}$, $\underset{\sim}{\mathrm{C}} >^P \underset{\sim}{\mathrm{B}}$ i $\underset{\sim}{\mathrm{C}} >^P \underset{\sim}{\mathrm{A}}$.

Uvažujme dále nesymetrická gaussovská fuzzy čísla $\underset{\sim}{\mathrm{A}}$ s funkcí věrohodnosti (29) s parametry $a, \sigma_{a1}, \sigma_{a2}$ (kde a je libovolné reálné, σ_{a1}, σ_{a2} jsou kladná reálná čísla)

$$\mu_A(x) = \begin{cases} \exp\left(-\frac{(x-a)^2}{2\sigma_{a1}^2}\right) & \text{pro} \quad x < a \\ \exp\left(-\frac{(x-a)^2}{2\sigma_{a2}^2}\right) & \text{pro} \quad x > a. \end{cases} \tag{29}$$

Je tedy také $\underset{\sim}{\mathrm{A}} \in F$. Pro $\sigma_{a1} = \sigma_{a2}$ je příslušné nesymetrické gaussovské fuzzy číslo symetrickým gaussovským fuzzy číslem. K fuzzy číslu $\underset{\sim}{\mathrm{A}}$ z (29) přísluší fuzzy číslo $-\underset{\sim}{\mathrm{A}}$ s parametry $-a, \sigma_{a2}, \sigma_{a1}$.

---

pro fuzzy číslo $\underset{\sim}{\mathrm{B}} + \underset{\sim}{\mathrm{A}}$ tak dostáváme podmínku

$$-\frac{[x+y-(a+b)]^2}{2\cdot(\sigma_a^2+\sigma_b^2)} = \ln\alpha,$$

tedy je gaussovské s parametry $(a+b), \sqrt{\sigma_a^2+\sigma_b^2}$. Fuzzy číslo $-\underset{\sim}{\mathrm{A}}$ je také gaussovské (s parametry $-a, \sigma_a$), když $\underset{\sim}{\mathrm{A}}$ je gaussovské.

Pro fuzzy číslo $\underset{\sim}{C} = \underset{\sim}{B} + (-\underset{\sim}{A})$ pak dostaneme funkci věrohodnosti (30) ve tvaru

$$\mu_c(x) = \begin{cases} \exp\left(-\frac{(x-(b-a))^2}{2(\sigma_{b1}^2+\sigma_{a2}^2)}\right) & \text{pro} \quad x < b-a \\ \exp\left(-\frac{(x-(b-a))^2}{2(\sigma_{b2}^2+\sigma_{a1}^2)}\right) & \text{pro} \quad x > b-a. \end{cases} \tag{30}$$

Výsledné fuzzy číslo $\underset{\sim}{C}$ je opět obecně nesymetrické gaussovské. Pro hodnotu výrazu $\mathrm{Pref}(\underset{\sim}{B}, \underset{\sim}{A})$ pak máme

$$\begin{aligned} \mathrm{Pref}(\underset{\sim}{B}, \underset{\sim}{A}) &= 2 \cdot \frac{\sqrt{\sigma_{b2}^2+\sigma_{a1}^2}}{\sqrt{\sigma_{b1}^2+\sigma_{a2}^2}+\sqrt{\sigma_{b2}^2+\sigma_{a1}^2}} \cdot \phi\left(\frac{b-a}{\sqrt{\sigma_{b2}^2+\sigma_{a1}^2}}\right), \text{ když } b < a, \\ \mathrm{Pref}(\underset{\sim}{B}, \underset{\sim}{A}) &= 2 \cdot \frac{\sqrt{\sigma_{b1}^2+\sigma_{a2}^2} \cdot \left(\phi\left(\frac{b-a}{\sqrt{\sigma_{b1}^2+\sigma_{a2}^2}}\right) - 0,5\right) + 0,5 \cdot \sqrt{\sigma_{b2}^2+\sigma_{a1}^2}}{\sqrt{\sigma_{b1}^2+\sigma_{a2}^2}+\sqrt{\sigma_{b2}^2+\sigma_{a1}^2}}, \\ &\quad \text{když } b > a. \end{aligned} \tag{31}$$

Příklad 10. Mějme tři nesymetrická gaussovská fuzzy čísla $\underset{\sim}{A}$, $\underset{\sim}{B}$, $\underset{\sim}{C}$, určená následujícími trojicemi svých parametrů

$$\begin{aligned} \underset{\sim}{A}: &\quad a = 0; \sigma_{a1}^2 = 1; \sigma_{a2}^2 = 1,5 \\ \underset{\sim}{B}: &\quad b = 1; \sigma_{b1}^2 = 1; \sigma_{b2}^2 = 2,5 \\ \underset{\sim}{C}: &\quad C = 1,5; \sigma_{c1}^2 = 1; \sigma_{c2}^2 = 2,5. \end{aligned}$$

Pak máme

$$\mathrm{Pref}(\underset{\sim}{B}, \underset{\sim}{A}) \doteq 0,75; \ \mathrm{Pref}(\underset{\sim}{C}, \underset{\sim}{B}) \doteq 0,61; \ \mathrm{Pref}(\underset{\sim}{C}, \underset{\sim}{A}) \doteq 0,84.$$

Tedy $\underset{\sim}{B} >^P \underset{\sim}{A}$, $\underset{\sim}{C} >^P \underset{\sim}{B}$ i $\underset{\sim}{C} >^P \underset{\sim}{A}$.

Předchozí příklad 10. navozuje otázku o tranzitivnosti relace $>^P$. Náhodně proto byla generována pro každé gaussovské fuzzy číslo trojice určujících reálných čísel z intervalu (0; 100). Z asi 892 882 trojic zkoumaných gaussovských fuzzy čísel nesplňovalo podmínku tranzitivnosti jen 927 trojic, např.

$$\begin{aligned} \underset{\sim}{A}: &\quad 0,50; 0,90; 0,70 \quad \mathrm{Pref}(\underset{\sim}{A}, \underset{\sim}{B}) = 0,504008 \\ \underset{\sim}{B}: &\quad 0,25; 0,01; 0,65 \quad \mathrm{Pref}(\underset{\sim}{B}, \underset{\sim}{C}) = 0,509616 \\ \underset{\sim}{C}: &\quad 0,40; 0,97; 0,97 \quad \mathrm{Pref}(\underset{\sim}{A}, \underset{\sim}{C}) = 0,493568. \end{aligned}$$

Relace $>^P$ mezi gaussovskými fuzzy čísly tedy není sice tranzitivní, ale pravděpodobnost této netranzitivity je velmi malá, asi 0,00104.

Vraťme se zpět k trojúhelníkovým fuzzy číslům. Snadno nahlédneme, že relace $>^P$ je také tranzitivní i na množině všech symetrických trojúhelníkových fuzzy čísel (která nedegenerují na singletony). (Relace $>^P$ musí být dokonce tranzitivní pro jakákoliv symetrická fuzzy čísla.) Protože součet (resp. rozdíl) dvou trojúhelníkových fuzzy čísel

$$\begin{aligned} \underset{\sim}{A}: &\quad a - s_{a1}, a, a + s_{a2}; s_{a1}, s_{a2} > 0 \\ \underset{\sim}{B}: &\quad b - s_{b1}, b, b + s_{b2}; s_{b1}, s_{b2} > 0 \end{aligned}$$

je opět trojúhelníkové fuzzy číslo, dostáváme pro výpočet Pref($\underset{\sim}{\mathrm{B}}$, $\underset{\sim}{\mathrm{A}}$) následující vyjádření (32)

$$\begin{aligned}
\mathrm{Pref}(\underset{\sim}{\mathrm{B}},\underset{\sim}{\mathrm{A}}) &= 0, \text{ když } (b-a)+(s_{a1}+s_{b2})<0,\\
\mathrm{Pref}(\underset{\sim}{\mathrm{B}},\underset{\sim}{\mathrm{A}}) &= \frac{[(b-a)+(s_{a1}+s_{b2})]^2}{(s_{a1}+s_{a2}+s_{b1}+s_{b2})\cdot(s_{a1}+s_{b2})}, \qquad (32)\\
&\quad \text{když } b-a<0<(b-a)+(s_{a1}+s_{b2}),\\
\mathrm{Pref}(\underset{\sim}{\mathrm{B}},\underset{\sim}{\mathrm{A}}) &= 1-\frac{[(a-b)+(s_{a2}+s_{b1})]^2}{(s_{a1}+s_{a2}+s_{b1}+s_{b2})\cdot(s_{a2}+s_{b1})},\\
&\quad \text{když } (b-a)-(s_{a2}+s_{b1})<0<b-a,\\
\mathrm{Pref}(\underset{\sim}{\mathrm{B}},\underset{\sim}{\mathrm{A}}) &= 1 \text{ když } 0<(b-a)+(s_{a2}+s_{b1}). \qquad (33)
\end{aligned}$$

(Pro symetrická trojúhelníková fuzzy čísla $\underset{\sim}{\mathrm{A}}$, $\underset{\sim}{\mathrm{B}}$ je $s_{a1}=s_{a2}=s_a$, $s_{b1}=s_{b2}=s_b$.)

Ve třídě obecně nesymetrických trojúhelníkových fuzzy čísel neplatí také pro relaci $>^P$ podmínka tranzitivity. Náhodným generováním uvedených typů fuzzy čísel jsme však zjistili, že podmínku tranzitivity nesplňovalo jen $1,067\cdot 10^{-3}\%$ z generovaných (přesněji: generováno bylo 86 483 810 trojic trojúhelníkových fuzzy čísel s parametry z intervalu $\langle 0;100\rangle$ a podmínku tranzitivity nesplnilo 923 trojic trojúhelníkových fuzzy čísel).
Podmínku tranzitivity například nesplňuje trojice trojúhelníkových fuzzy čísel určených pro fuzzy číslo $\underset{\sim}{\mathrm{X}}$ trojicí $x, s_{x1}, s_{x2}, s_{x1}>0,\ s_{x2}>0$:

$$\begin{aligned}
\underset{\sim}{\mathrm{A}}: &\quad 39,4707;\ 26,5723;\ 1,82885,\\
\underset{\sim}{\mathrm{B}}: &\quad 26,5561;\ 19,1287;\ 41,4242,\\
\underset{\sim}{\mathrm{C}}: &\quad 735,7175;\ 88,5758;\ 76,7576,
\end{aligned}$$

$\mathrm{Pref}(\underset{\sim}{\mathrm{B}},\underset{\sim}{\mathrm{A}})=0,501609;\ \ \mathrm{Pref}(\underset{\sim}{\mathrm{C}},\underset{\sim}{\mathrm{B}})=0,502746;\ \ \mathrm{Pref}(\underset{\sim}{\mathrm{C}},\underset{\sim}{\mathrm{A}})=0,495316.)$

Vraťme se opět k příkladu 8. Do tabulky 4 zaneseme hodnoty Pref($\underset{\sim_i}{\mathrm{P}}$, $\underset{\sim_j}{\mathrm{P}}$) mezi úspěšnostmi každých dvou preparátů P_i, P_j, $i,j=1,2,\cdots,5$. Do tabulky 5 pak zaneseme výsledky posuzování vztahu $\underset{\sim_i}{\mathrm{P}} >^K \underset{\sim_j}{\mathrm{P}}$ mezi dvěma úspěšnostmi preparátů P_i, P_j tak, že znakem 0 v tabulce označíme situaci, kdy fuzzy číslo $\underset{\sim_i}{\mathrm{P}} - \underset{\sim_j}{\mathrm{P}}$ bude nulové a znakem + (resp. −) situaci, kdy bude kladné (resp. záporné). Z tabulky 4 dostaneme následující řetězce vztahů mezi úspěšnostmi preparátů:

$$\underset{\sim_3}{\mathrm{P}} >^P \underset{\sim_5}{\mathrm{P}} >^P \underset{\sim_2}{\mathrm{P}} >^P \underset{\sim_1}{\mathrm{P}} >^P \underset{\sim_4}{\mathrm{P}}.$$

Z tabulky 5. ale získáme méně informace, pouze, že $\underset{\sim_3}{\mathrm{P}} >^K \underset{\sim_2}{\mathrm{P}}$ a $\underset{\sim_3}{\mathrm{P}} >^K \underset{\sim_4}{\mathrm{P}}$. Relace $>^P$, generovaná pomocí (26) v množině (trojúhelníkových) fuzzy čísel z F je ostrá, antisymetrická i ostře tranzitivní. V množině fuzzy čísel z F, která nejsou singletony, je relace $>^P$ i úplná.

Pref	$P_{\sim_1}$	$P_{\sim_2}$	$P_{\sim_3}$	$P_{\sim_4}$	$P_{\sim_5}$
$P_{\sim_1}$	0,50	0,40	0,03	0,76	0,17
$P_{\sim_2}$	0,60	0,50	0,00	0,90	0,14
$P_{\sim_3}$	0,97	1,00	0,50	1,00	0,71
$P_{\sim_4}$	0,24	0,10	0,00	0,50	0,02
$P_{\sim_5}$	0,83	0,86	0,29	0,98	0,50

Tabulka 4:

$>^K$	$P_{\sim_1}$	$P_{\sim_2}$	$P_{\sim_3}$	$P_{\sim_4}$	$P_{\sim_5}$
$P_{\sim_1}$	0	0	0	0	0
r $P_{\sim_2}$	0	0	-	0	0
$P_{\sim_3}$	0	+	0	+	0
$P_{\sim_4}$	0	0	-	0	0
$P_{\sim_5}$	0	0	0	0	0

Tabulka 5:

4 Závěr

Ukázali jsme, jak lze zavést antisymetrickou, ostře tranzitivní a úplně uspořádanou fuzzy relaci na (trojúhelníkových) fuzzy číslech s riemannovsky integrovatelnou mírou věrohodnosti. Přitom pod úplností fuzzy relace jsme rozuměli jen to, že definiční obor její funkce věrohodnosti je $F \times F$. Tato (tyto) fuzzy relace $\lesssim^T$ ($\precsim_1, \precsim_2, \precsim, \lesssim, \gtrsim_1, \gtrsim_2, \gtrsim$) umožňuje (umožňují) porovnat kterákoliv dvě fuzzy čísla z množiny F na základě věrohodnostní funkce a na rozdíl od velmi často užívaného kategorického porovnávání fuzzy čísel metodou průseku a spojení je (jsou) úplná (úplné). Podmínku fuzzy tranzitivity ale obecně nesplňuje (nesplňují ji však ani jiné fuzzy relace na jistých množinách fuzzy čísel, např. fuzzy relace preference ([1], [2])).

Relace $>^P$, zavedená prostřednictvím (26), se ukazuje jako velmi výhodná, řešící dobře problémy, reprezentované příkladem 8.

Reference

[1] Mareš M. (1994). *Computation over fuzzy quantities.* CRC Press, Boca Raton.

[2] Talašová J. (2003). *Fuzzy metody vícekriteriálního hodnocení a rozhodování.* Univerzita Palackého v Olomouci, Olomouc.

[3] Jang J.,-S.R, Sun C.- T, Mizutani E. (1997). *Neuro – fuzzy and soft computing.* Prentice Hall, NJ 07458, USA.

[4] Půlpán Z. (2000). *K problematice měření v humanitních vědách.* Academia, Praha.

[5] Mareš M. (2001). *Počítání s vágností II.* Automatizace **2**, 96 – 99.

[6] Xuzhu Wang, Etienne E. Kerre (2001). *Reasonable properties for the ordering of fuzzy quantities.* (I) Fuzzy Sets and Systems **118**, 375 – 385; (II) Fuzzy Sets and Systems **118**, 387 – 405.

Adresa: Z. Půlpán, Katedra matematiky, Pedagogická fakulta Univerzity Hradec Králové, Rokitanského 62, 500 03 Hradec Králové 3

E-mail: `zdenek.pulpan@uhk.cz`

ANALÝZA PŘEŽITÍ A COXŮV MODEL PRO DISKRÉTNÍ ČAS

Soňa Reisnerová

Klíčová slova: Coxův regresní model, proporcionální rizika, analýza nezaměstnanosti.

Abstrakt: Tento článek se zabývá Coxovým regresním modelem a jeho aplikací pro diskrétní povahu dat. Ukazuje, jakými metodami se získávají odhady parametrů modelu a jaké jsou jejich asymptotické vlastnosti. Zmiňuje testy významnosti parametrů modelu, test dobré shody a test proporcionality rizik. Vše je v závěru článku demonstrováno na datech týkajících se vývoje nezaměstnanosti v České republice.

1 Úvod

V tomto článku budeme pomocí stochastických procesů modelovat data, která měří čas do nějaké události. Jsou to data související s analýzou přežití, u nichž nás bude zajímat zejména intenzita, s jakou události nastávají. V regresním modelu proporcionálních rizik tato data závisejí na vysvětlujících proměnných (kovariátách). Jelikož je základní riziková funkce, která představuje obecnou na čase závislou intenzitu výskytu událostí, libovolná, dostáváme semiparametrický model. Nejdříve uděláme stručný úvod do analýzy přežití, pak přejdeme k mnohorozměrnému čítacímu procesu a od něj se již dostaneme ke Coxově regresnímu modelu proporcionálních rizik.

2 Analýza přežití

Nejznámější je použití analýzy přežití v lékařských studiích zkoumajících časy úmrtí pro skupinu pacientů trpících stejnou chorobou. Zabývejme se nyní odhadem a testováním hypotéz týkajících se rizikové funkce λ, v tomto případě intenzitou smrti, která je obecně definována takto: Předpokládejme, že máme data týkající se homogenní populace. Nechť X je absolutně spojitá nezáporná náhodná veličina s distribuční funkcí $F(t) = P(X \leq t)$ a hustotou $f(t) = F'(t) = -S'(t)$, kde S je funkce přežití definovaná vztahem $S(t) = P(X > t) = 1 - F(t)$. X představuje čas úmrtí (čas výskytu události) pro každého jedince, pak

$$\lambda(t) = \lim_{\Delta t \to 0+} \frac{P(X \leq t + \Delta t \mid X \geq t)}{\Delta t} = \frac{f(t)}{1 - F(t)} = -\frac{\mathrm{d}}{\mathrm{d}t} \log S(t). \qquad (1)$$

Kumulativní riziková funkce je definovaná

$$\Lambda(t) = \int_0^t \lambda(s)\mathrm{d}s, \qquad \text{potom} \qquad S(t) = 1 - F(t) = \exp\{-\Lambda(t)\}. \qquad (2)$$

3 Mnohorozměrný čítací proces

Pozorujme časy událostí skupiny n jedinců, kteří pocházejí z homogenní populace. Uvažme náhodné cenzorování zprava a konečný čas studie. $X_i \overset{iid}{\sim} X$ čas události, $C_i \overset{iid}{\sim} C$ čas cenzorování pro jedince $i = 1, \ldots, n$ a jsou pro něj nezávislé. Ve skutečnosti pozorujeme pouze (T_i^*, δ_i), kde $T_i^* = \min(X_i, C_i)$ je cenzorovaný čas události a δ_i je 1 pro necenzorované, 0 pro cenzorované pozorování. $N(t) = \sum_{i=1}^n N_i(t)$ je počet pozorovaných událostí do času t v celé populaci, kde $N_i(t) = I(T_i^* \leq s, \delta_i = 1)$. N je mnohorozměrný čítací proces. $Y(t) = \sum_{i=1}^n Y_i(t)$ je počet pozorovaných jedinců, u kterých může nastat událost v čase t, kde $Y_i(t) = I(T_i^* \geq t)$. Y je indikátorový stochastický proces.

Pro odhad parametrů budeme potřebovat věrohodnostní funkci, část obsahující charakteristiky náhodné veličiny X_i (f, F, λ) má tvar:

$$L = \prod_{i=1}^n f(T_i^*)^{\delta_i} P(X_i > T_i^*)^{1-\delta_i} = \prod_{i=1}^n f(T_i^*)^{\delta_i} (1 - F(T_i^*))^{1-\delta_i}.$$

S využitím vztahů (1) a (2) dostáváme

$$L = \prod_{i=1}^n \lambda(T_i^*)^{\delta_i} (1 - F(T_i^*)) = \prod_{i=1}^n \lambda(T_i^*)^{\delta_i} \exp\left\{ - \int_0^{T_i^*} \lambda(s) \mathrm{d}s \right\}.$$

Tato situace je popsána v rozsáhlé literatuře, uveďme např. [1], [5], [6]. My budeme nadále předpokládat, že máme **časově diskrétní pozorování** v periodách jednotkové délky, délka celé studie je T period a $T_i^* \in \{1, \ldots, T\}$. Riziková funkce je skokovitá, ale konstantní pro každou periodu. Za těchto předpokladů a pokud označíme $\mathrm{d}N_i(t)$ přírůstky N_i v čase $(t-1, t\rangle$ můžeme věrohodnost přepsat

$$L = \prod_{i=1}^n \lambda(T_i^*)^{\delta_i} \exp\left\{ - \sum_{s=1}^{T_i^*} \lambda(s) \right\} = \prod_{i=1}^n \prod_{t=1}^T \lambda(t)^{\mathrm{d}N_i(t)} \exp\left\{ - \sum_{s=1}^T \lambda(s) Y_i(s) \right\}. \tag{3}$$

4 Coxův regresní model proporcionálních rizik

V jistých případech potřebujeme do modelu zahrnout vliv vysvětlujících proměnných, čehož lze dosáhnout přidáním individuálních kovariát Z_i (v našem případě nezávislých na čase) do rizikové funkce. Mějme populaci n jedinců a u každého pozorujeme (T_i^*, δ_i, Z_i). Předpokládejme, že jsou X_i a C_i nezávislé při daném Z_i. Vztah kovariát a rizikové funkce vysvětlíme multiplikativním regresním modelem

$$\lambda_i(t) = \lambda_0(t) \exp\{\beta^T Z_i\}, \qquad t \in \{1, \ldots, T\}, \qquad i = 1, \ldots, n, \tag{4}$$

nazývaným Coxův regresní model proporcionálních rizik, kde λ_0 je neznámá nezáporná základní riziková funkce a β je p-rozměrný vektor regresních parametrů. Nechť je základní riziková funkce po částech konstantní, pak věrohodnostní funkce pro Coxův regresní model odvozená z (3) má tvar:

$$L = \prod_{i=1}^{n} \prod_{t=1}^{T} (\lambda_0(t) \exp\left\{\beta^T Z_i\right\})^{\mathrm{d}N_i(t)} \exp\left\{ - \sum_{s=1}^{T} \lambda_0(s) \exp\{\beta^T Z_i\} Y_i(s) \right\}. \tag{5}$$

V našem případě je tento model plně parametrizovaný, a proto mohou být jeho parametry odhadnuté maximalizací logaritmu věrohodnostní funkce. Pokud parciální derivaci log-věrohodnostní funkce vzhledem k $\lambda_0(t)$ položíme rovnou nule a vyřešíme, dostaneme Breslow - Crowleyho odhad základní rizikové funkce:

$$\widehat{\lambda_0}(t) = \frac{\sum_{j=1}^{n} \mathrm{d}N_j(t)}{\sum_{i=1}^{n} Y_i(t) \exp\{\beta^T Z_i\}}, \tag{6}$$

kde parameter β může být nahrazen odhadem. Ve [4] bylo navrženo odvození β z parciální věrohodnostní funkce $L_p(\beta)$, která je věrohodnostní profilovou funkcí ve smyslu $L_p(\beta) = \max_\lambda L$. V našem případě má tvar:

$$L_p(\beta) = \prod_{i=1}^{n} \prod_{t=1}^{T} \left(\frac{\exp\{\beta^T Z_i\}}{\sum_{j=1}^{n} Y_j(t) \exp\{\beta^T Z_j\}} \right)^{\mathrm{d}N_i(t)}. \tag{7}$$

Její odvození lze najít např. v [5]. Označme $C(\beta, T) = \log L_p(\beta)$. Odhad $\hat{\beta}$ parametru β pro obecný případ (spojitý čas) je definován ve [2] řešením rovnice, která má v našem případě tvar

$$0 = \frac{\partial C(\beta, T)}{\partial \beta} = \sum_{i=1}^{n} \sum_{t=1}^{T} \mathrm{d}N_i(t) \left[Z_i - \frac{\sum_{j=1}^{n} Y_j(t) Z_j \exp\{\beta^T Z_j\}}{\sum_{j=1}^{n} Y_j(t) \exp\{\beta^T Z_j\}} \right]. \tag{8}$$

Tato rovnice se řeší iterativním algoritmem. Stejnou rovnici bychom dostali, kdybychom dosadili $\widehat{\lambda_0}(t)$ z (6) do $\lambda_0(t)$ v $\partial \log L / \partial \beta$.

Pokud nás zajímají statistické vlastnosti $\hat{\beta}$ a $\widehat{\lambda_0}(t)$ můžeme vyjít z plně parametrizovaného modelu a přímo použít vlastností maximálně věrohodných odhadů. Také lze využít obecných výsledků ve [2] a modifikovat je pro náš případ: Abychom mohli vyjádřit matici druhých parciálních derivací z log-parciální věrohodnostní funkce, potřebujeme zavést následující pomocné funkce:

$$S_n^{(0)}(\beta, t) = \sum_{j=1}^{n} Y_j(t) \exp\{\beta^T Z_j\} \qquad S_n^{(1)}(\beta, t) = \sum_{j=1}^{n} Y_j(t) Z_j \exp\{\beta^T Z_j\}$$

$$S_n^{(2)}(\beta, t) = \sum_{j=1}^{n} Y_j(t) Z_j Z_j^T \exp\{\beta^T Z_j\}. \tag{9}$$

Matice druhých derivací $\Sigma_n(\beta, T) = \partial^2 C(\beta, T)/\partial\beta\beta^T$ má tvar

$$-\sum_{i=1}^{n}\sum_{t=1}^{T} \mathrm{d}N_i(t) \left[\frac{S_n^{(2)}(\beta,t)S_n^{(0)}(\beta,t) - (S_n^{(1)}(\beta,t))(S_n^{(1)}(\beta,t))^T}{(S_n^{(0)}(\beta,t))^2} \right]. \qquad (10)$$

Asymptotická stabilita, předpoklad na pomocné sumy v (9), Lindebergova podmínka a předpoklad regularity (podrobně pro obecný případ viz. [2], 1982, str. 1105) jsou postačujícími předpoklady pro konzistenci β, $\hat{\beta} \xrightarrow{\wp} \beta$ a asymptotickou normalitu, $\sqrt{n}(\hat{\beta} - \beta)$ má asymptoticky mnohorozměrné rozdělení $N_p(0, \Sigma^{-1})$, kde Σ je pozitivně definitní a lze stejnoměrně konzistentně odhadnout $-\frac{1}{n}\Sigma_n(\hat{\beta}, T)$.

Nyní když dosadíme odhad β do Breslow - Crowleyho odhadu (6), dostaneme odhad základní rizikové funkce. Kumulativní funkce, která je definovaná v (2), se odhadne jako

$$\widehat{\Lambda_0}(t) = \sum_{s=1}^{t} \frac{\sum_{j=1}^{n} \mathrm{d}N_j(s)}{\sum_{i=1}^{n} Y_i(s) \exp\{\hat{\beta}^T Z_i\}}. \qquad (11)$$

Dle [2] lze za zmíněných podmínek odvodit slabou konvergenci $\sqrt{n}(\widehat{\Lambda_0} - \Lambda_0)$ ke Gaussovskému procesu s nulovou střední hodnotou a kovarianční funkcí, která může být stejnoměrně konzistentně odhadnuta

$$\sum_{s=1}^{t}\sum_{i=1}^{n} \frac{\mathrm{d}N_i(s)}{(S_n^{(0)}(\hat{\beta}, s))^2} + \qquad (12)$$

$$+ \quad \left(\sum_{s=1}^{t}\sum_{i=1}^{n} \frac{S_n^{(1)}(\hat{\beta}, s)\mathrm{d}N_i(s)}{(S_n^{(0)}(\hat{\beta}, s))^2} \right) \Sigma_n(\hat{\beta}, T)^{-1} \left(\sum_{s=1}^{t}\sum_{i=1}^{n} \frac{S_n^{(1)}(\hat{\beta}, s)\mathrm{d}N_i(s)}{(S_n^{(0)}(\hat{\beta}, s))^2} \right)^T.$$

Na základě předchozích tvrzení je již snadné zkonstruovat konfidenční interval pro regresní parametry β a základní rizikovou funkci $\lambda_0(t),\ t = 1, \ldots, T$.

Statistické testy

Vlastnosti klasických maximálně věrohodných odhadů nám umožní snadno odvodit statistické testy pro parametr β. Waldova statistika pro H: $\beta = \beta_0$ je rovna $(\hat{\beta} - \beta_0)^T \Sigma_n(\hat{\beta}, T)(\hat{\beta} - \beta_0)$, a má přibližně χ^2 rozdělení s p stupni volnosti za platnosti hypotézy H. Statistika založená na věrohodnostním poměru pro H: $\beta = \beta_0$ má tvar $2(C(\hat{\beta}, T) - C(\beta_0, T))$. Přestože je asymptoticky ekvivalentní k Waldově statistice, mohou se jejich vlastnosti na konečném souboru mírně lišit.

Grafický test dobré shody navržený v [3] je založený na součtu reziduí ve stratu $I \subset \{1, \ldots, n\}$

$$R_I(\hat{\beta}, t) = \sum_{i \subset I} N_i(t) - \sum_{s=1}^{t} \frac{\sum_{i=1}^{n} \mathrm{d}N_i(s) \sum_{j \subset I} Y_j(s) \exp\{\hat{\beta}^T Z_j\}}{\sum_{i=1}^{n} Y_i(s) \exp\{\hat{\beta}^T Z_i\}}, \qquad (13)$$

pro $t \in \{1, \ldots, T\}$. Znázorňuje rozdíl mezi pozorovanými a předpokládanými počty událostí ve stratu I.

Jedním z klíčových předpokladů Coxova modelu je proporcionalita rizik. Poměr rizikových funkcí dvou objektů k a l by měl splňovat vztah

$$\frac{\lambda_0(t)\exp\{\beta^T Z_k\}}{\lambda_0(t)\exp\{\beta^T Z_l\}} = \frac{\exp\{\beta^T Z_k\}}{\exp\{\beta^T Z_l\}}, \tag{14}$$

který je nezávislý na čase. Jednoduchý grafický test tohoto předpokladu je navržen v [6]. Nejprve odhadneme kumulativní rizikovou funkci bez znalosti speciálního tvaru základní rizikové funkce. Toto umožňuje Nelson-Aalenův odhad (viz. [1])

$$\widehat{\Lambda}(t) = \sum_{s=1}^{t} \frac{\sum_{i=1}^{n} \mathrm{d}N_i(s)}{\sum_{i=1}^{n} Y_i(s)}. \tag{15}$$

Nechť $I, J \subset \{1, \ldots, n\}$ jsou dvě disjunktní strata, pak součtem přes i v (15) procházející množinami I resp. J dostaneme $\widehat{\Lambda}_I(t)$ resp. $\widehat{\Lambda}_J(t)$. Vhodným grafickým testem je zobrazení $-\log(\widehat{\Lambda}_m(t))$, kde $m \in \{I, J\}$. Pokud je předpoklad proporcionality rizik splněn, křivky jsou přibližně paralelní.

5 Analýza nezaměstnanosti v České republice

Představme si ideální data týkající se nezaměstnanosti. Pro každou skupinu a periodu by obsahovala počet lidí ohrožených nezaměstnaností, počet nově nezaměstnaných a počet lidí, kteří si znovu našli práci. Pro taková data by logickými modely byly Poissonovy, jako speciální případ standardní analýzy přežití. Bohužel taková data nebyla k dispozici, proto jsme použili Poissonův model s regresí na souhrnná data o nezaměstnanosti v České republice z Českého statistického úřadu, která nemají tak bohatou strukturu.

Data obsahují pozorování z 11 kvartálů (1/2001 - 3/2003) rozdělených do 14 krajů, podle pohlaví a do 6 věkových kategorií (15-24, 25-29, 30-34, 35-44, 45-54, 55 a více). Dohromady máme $14 \times 2 \times 6 = 168$ tříd. Data jsou seřazena do matice $\{N_{ti}\}_{11\times 168}$, kde N_{ti} je počet nezaměstnaných v třídě i v čase t. Řádky matice jsou seřazené podle kraje, pohlaví a nakonec dle věku, sloupce ukazují vývoj v čase. Máme $14 + 2 + 6 = 22$ kovariát, které jsou ale jen indikátory příslušnosti do určité skupiny. Postačující velikost parametru β je 19 (prvních 13 odpovídá krajům, β_{14} ženám, β_{15} až β_{19} věkovým kategoriím), protože parametr odpovídající prvnímu kraji, pohlaví a věkové kategorii je nulový. Data byla analyzována v softwaru Matlab verze 6.0.

Předpokládejme, že pro $t = 1, \ldots, 11, i = 1, \ldots, 168$ se data řídí modelem (podobně viz. [7])

$$N_{ti} \sim Poiss\{N_{0,i}\lambda(t,i)\} \quad \text{a} \quad \lambda(t,i) = \lambda_0(t)\exp\{\beta^T Z_i\}, \tag{16}$$

kde $N_{0,i}$ je počet obyvatel v jednotlivých krajích ČR (ke 3. kvartálu 2003), $\lambda_0(t)$ je po částech konstantní funkce. Věrohodnostní funkce má pro náš model tvar

$$L = \prod_{i=1}^{168}\prod_{t=1}^{11}(N_{0,i}\lambda_0(t)\exp\{\beta^T Z_i\})^{N_{ti}}\left(\frac{\exp\{-N_{0,i}\lambda_0(t)\exp\{\beta^T Z_i\}\}}{N_{ti}!}\right), \quad (17)$$

odpovídající parciální věrohodnostní funkce je

$$L_p(\beta) = \prod_{i=1}^{168}\prod_{t=1}^{11}\left(\frac{\exp\{\beta^T Z_i\}}{\sum_{j=1}^{168} N_{0,i}\exp\{\beta^T Z_j\}}\right)^{N_{ti}}. \quad (18)$$

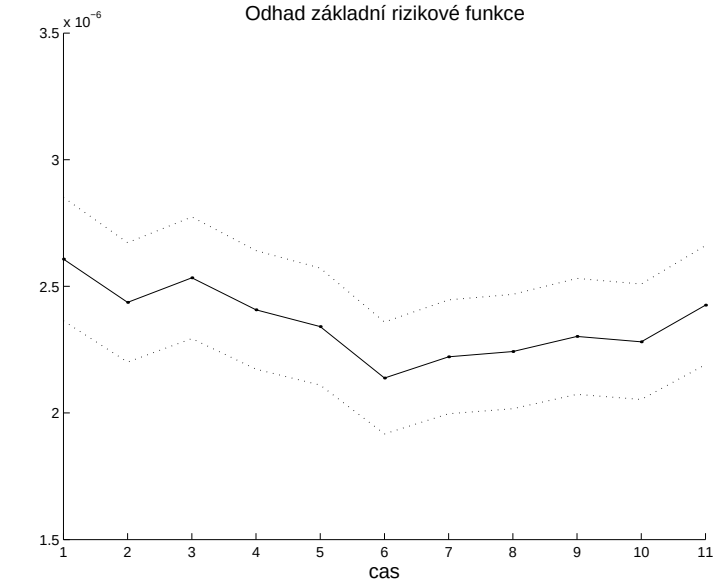
Použití tohoto modelu má jisté problémy, neboť předpokládá Poissonovo rozdělení, aniž bychom měli nezávislá pozorování. Na druhou stranu výsledky testů jsou uspokojivé a ukazuje se, že model poměrně dobře popisuje realitu.

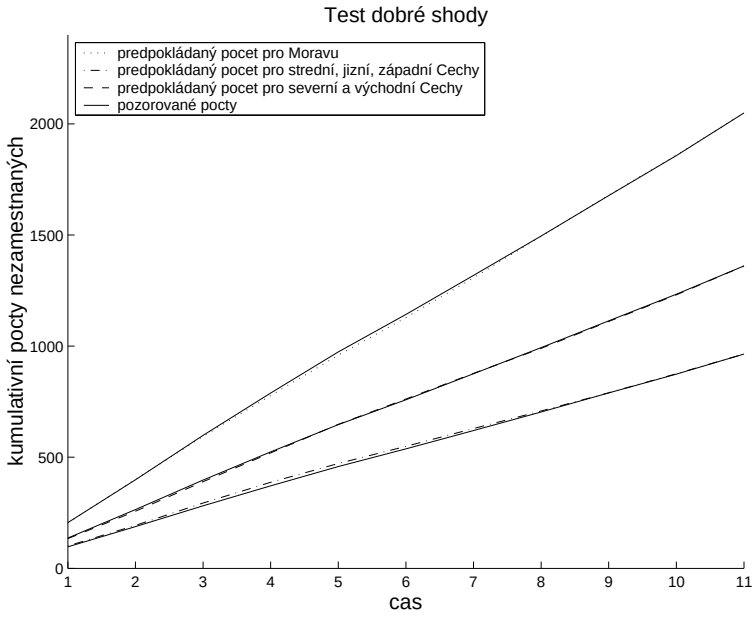
Numerické výsledky

Odhad parametru β jsme získali z parciální věrohodnostní funkce pomocí Newton - Raphsonova iteračního algoritmu, $\widehat{\lambda_0}(t)$ jsme dostali z (6) a dále jsme využili poznatky o Coxově regresním modelu. Všechny složky parametru β máme statisticky významné. Waldova statistika týkající se významnosti parametru β je 1 471 a statistika věrohodnostního poměru je 1 477. Odpovídající hodnota χ^2 statistiky o 19 stupních volnosti je 30.14, takže můžeme zamítnout hypotézu o nulovosti parametru β. Následující tabulka obsahuje odhady parametrů β s 95% konfidenčním intervalem.

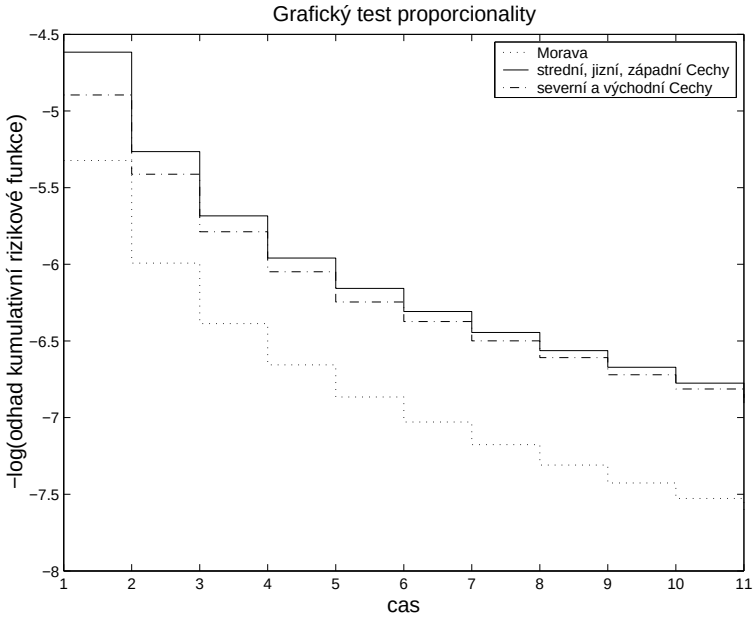
β	odhad	konfidenční interval	
Středočeský kraj	0.2890	0.1487	0.4292
Budějovický kraj	0.2210	0.1802	0.2617
Plzeňský kraj	0.2458	0.0619	0.4297
Karlovarský kraj	0.5600	0.3576	0.7625
Ústecký kraj	1.1486	1.0257	1.2714
Liberecký kraj	0.2887	0.0928	0.4847
Královéhradecký kraj	0.2387	0.0673	0.4100
Pardubický kraj	0.4975	0.3222	0.6728
Jihlavský kraj	0.2359	0.0479	0.4240
Brněnský kraj	0.6418	0.5143	0.7694
Olomoucký kraj	0.8389	0.6991	0.9787
Zlínský kraj	0.6222	0.4695	0.7749
Ostravský kraj	1.1880	1.0747	1.3014
ženy	0.1940	0.1345	0.2535
věková kategorie 25 - 29	-0.4919	-0.5879	-0.3958
věková kategorie 30 - 34	-0.7587	-0.8638	-0.6536
věková kategorie 35 - 44	-0.1840	-0.2727	-0.0954
věková kategorie 45 - 54	-0.1361	-0.2232	-0.0490
věková kategorie 55 a více	-1.4020	-1.5354	-1.2686

Nejhorší situace je v Ústeckém a Ostravském kraji. Lépe než ostatní kraje si vede Praha hl. m. Naopak horší situace je pro ženy a věkovou skupinu 15-24 let. Na následujících obrázcích je znázorněn odhad základní rizikové funkce s 95% intervalovými odhady a grafický test dobré shody pro tři strata.





Následuje grafický test proporcionality rizik.



Křivky $-\log(\widehat{\Lambda}(t))$ jsou pro daná tři strata přibližně paralelní.

Model nabízí několik modifikací. Mohli bychom do něj zahrnout vliv interakcí kovariát, či zohlednit vliv sousedních krajů.

Reference

[1] Andersen P.K., Borgan O. (1985). *Counting process models for life history data: a review*. Scand.J.Statist. **12**, 97 – 158.

[2] Andersen P.K., Gill R.D. (1982). *Cox's regression model for counting processes: a large sample study*. Ann. Statist. **10**, 1100 – 1120.

[3] Arjas E. (1988). *A graphical method for assessing goodness of fit in Cox's proportional hazards model*. J. Amer. Statist. Assoc. **83**, 204 – 212.

[4] Cox D.R. (1972). *Regression models and life tables (with discussion)*. J. Roy. Statist. Soc. B **34**, 187 – 220.

[5] Fleming T.R., Harrington D.P. (1991). *Counting processes and survival analysis*. Wiley, New York.

[6] Therneau T.M., Grambsch P.M. (2000). *Modeling survival data: Extending the Cox model*. Springer, New York.

[7] Volf P. (2003). *Cox's regression models for dynamics of grouped unemployment data*. Bulletin of the Czech Econometric Society **10**, 19, 151 – 162.

Poděkování: Tato práce vznikla s podporou grantu GAČR č. 402/04/1294.

Adresa: S. Reisnerová, KPMS MFF UK, Sokolovská 83, 186 75 Praha 8

E-mail: `sona.reisnerova@centrum.cz`

EXTRÉMY V TEPLOTNÍCH ŘADÁCH

Monika Rencová

Klíčová slova: Teorie extrémů, teplotní řady, tříparametrické Weibullovo rozdělení.

Abstrakt: Ze statistického hlediska je užitečné studovat chování maximálních nebo minimálních ročních teplot. Z teorie extrémů vyplývá, že rozdělení ročních minimálních/maximálních teplot by mělo odpovídat jednomu z extremálních rozdělení. Existuje však několik důvodů, proč dostupná data neodpovídají Gumbelovu rozdělení. Pro modelování těchto dat může být vhodnější tříparametrické Weibullovo rozdělení.

1 Úvod

Široce rozšířená hypotéza o globálním oteplování vyvolala zájem o studium teplotních řad. Ke změně však nedochází pouze u průměrných teplot, ale mnoho klimatologů tvrdí, že změna klimatu se může projevovat především vznikem extremálních událostí.

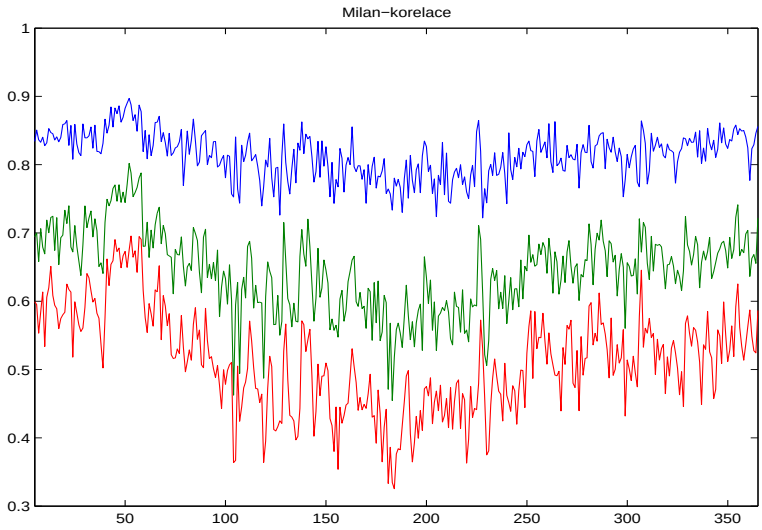
V příspěvku jsou sledovány dlouhé řady průměrných denních teplot naměřených v sedmi meteorologických stanicích - v Padově (1766 - 1997), Milánu (1763 - 1998), Uppsale (1722 - 2000), Stockholmu (1756 - 2000), Cadizu (1786 - 2000), St. Petersburgu (1743 - 1997) a Belgii (1767 - 1998). Tato data pocházejí z knihy D. Camuffa a P. Jonese "Improved understanding of past climatic variability from early daily European instrumental sources".

Podle známé extremální teorie bychom mohli usuzovat, že maximální/ minimální roční teploty budou rozděleny podle jednoho z extremálních rozdělení. Avšak naše data tento závěr nepotvrzují.

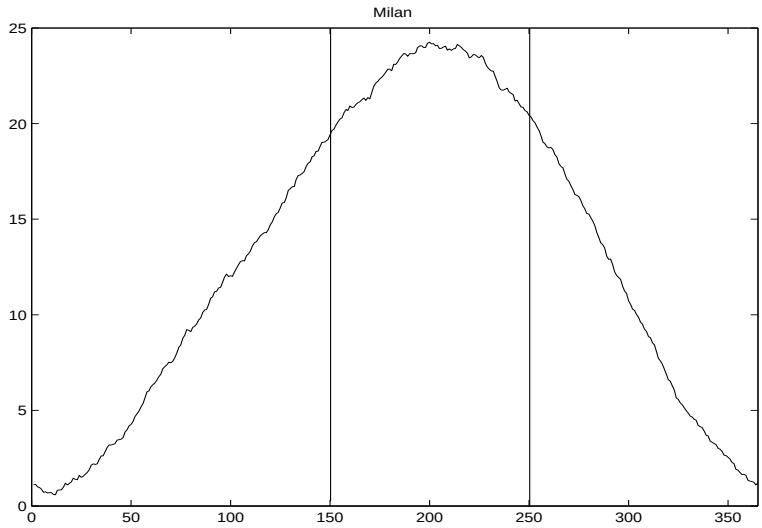
2 Základní vlastnosti dat

Ukážeme si několik důvodů, proč Gumbelovo rozdělení v našem případě není vhodné pro modelování těchto dat.

1. Gumbelovo rozdělení vzniká jako asymptotické rozdělení maxima/ minima posloupnosti nezávislých veličin, případně maxima/minima stacionárních posloupností s krátkou pamětí (např. ARMA posloupností), viz článek Daniely Jaruškové. Konvergence je však velmi pomalá, pokud mezi po sobě jdoucími členy posloupnosti je silná závislost. Tak tomu je např. u našich dat. Korelační koeficient průměrných denních teplot mezi dvěma po sobě jdoucími dny se pohybuje u všech stanic kolem hodnoty 0,8. Na obrázku 1 jsou znázorněny korelační koeficienty průměrných denních teplot mezi jednotlivými po sobě jdoucími dny v roce. Horní křivka odpovídá dvěma po sobě jdoucím dnům, další křivky směrem shora dolů odpovídají korelačním koeficientům průměrných denních teplot mezi dny, jejichž rozdíl jsou 2 dny



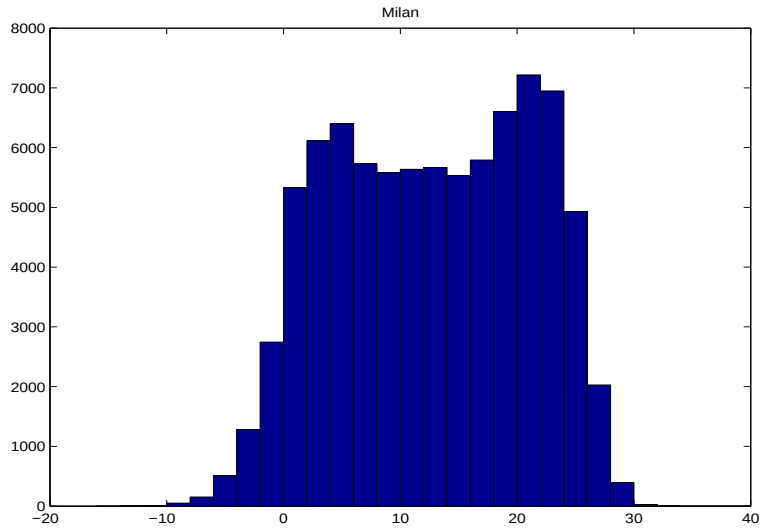
Obrázek 1: Korelace mezi jednotlivými dny.



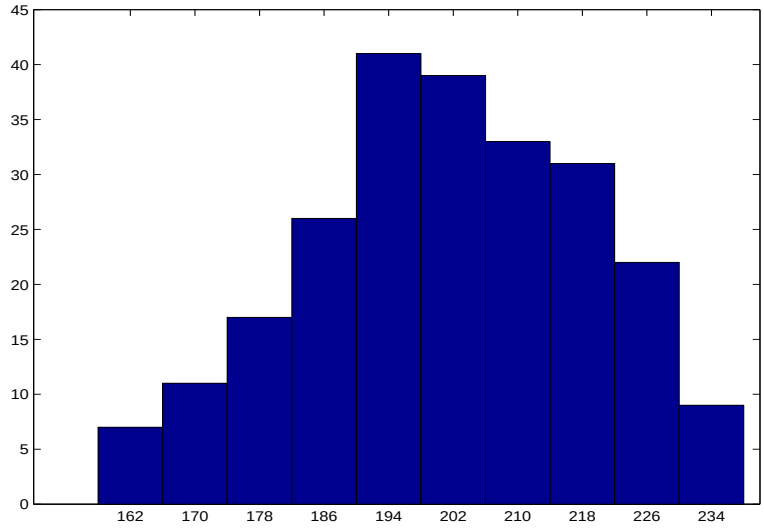
Obrázek 2: Závislost průměrné denní teploty na dni v roce.

a 3 dny, t.j. u druhé křivky byly spočteny korelační koeficienty mezi 1. lednem a 3. lednem, 2. lednem a 4. lednem atd.

2. Dalším důvodem proč asymptotická teorie selhává, je nestejné rozdělení veličin, z nichž počítáme maximum/minimum. Nestejné rozdělení je způso-



Obrázek 3: Histogram všech dostupných dat.

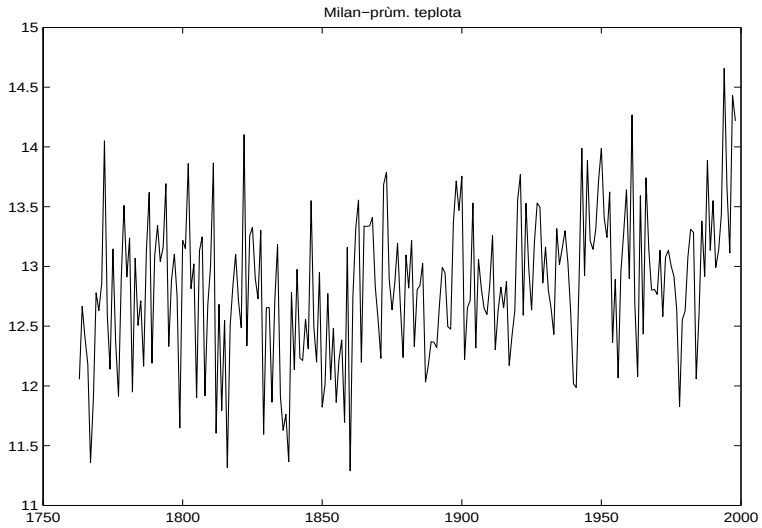


Obrázek 4: Histogram z dat, kdy bylo dosaženo maximum.

beno ročním chodem. Zjistili jsme např., že v našich datech nastal nejteplejší den v rozmezí přibližně 100 dnů. To znamená, že pro nalezení maxima není nutno sledovat období mimo tento časový úsek (přibližně 150. - 250. den v roce), viz obrázek 2.

3. Vzhledem k tomu, že uvažujeme maxima z relativně malého počtu dat, hraje výraznou roli původní rozdělení veličin. Na obrázku 3 je histogram všech denních teplot, na obrázku 4 je příklad histogramu z dat, ve kterých alespoň v některém roce byla dosažena maximální teplota.

4. Rozdělení maxim/minim mohou být také ovlivněna celkovým růstem teploty, což ilustruje graf vývoje průměrné teploty, viz obrázek 5.



Obrázek 5: Řada průměrných ročních teplot.

Základní informace o průměrných, minimálních a maximálních hodnotách jsou uvedeny v tabulkách 1 až 3.

	Průměr		
	$\bar{x}$	σ_{n-1}	šikmost
Miláno	12.8297	0.6123	-0.0043
Padova	13.0023	0.5850	-0.1799
Belgie	9.5808	0.7513	-0.2845
Cadiz	17.520	0.5603	0.2616
Uppsala	5.1719	0.9909	0.0301
Stockholm	5.7945	0.9478	0.0128
St.Petersburg	4.2306	1.1771	-0.0297

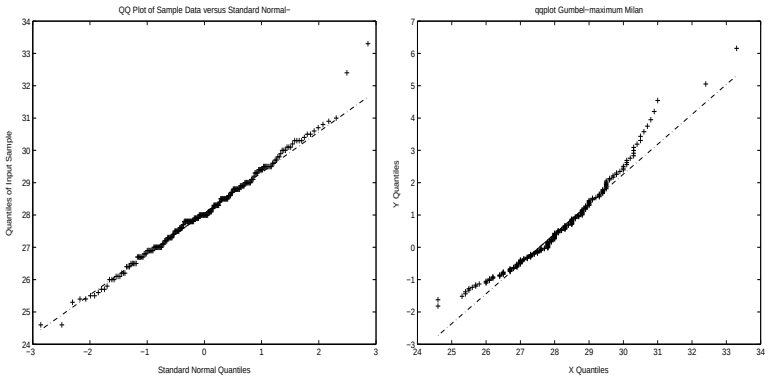
Tabulka 1: Průměr – vlastnosti.

	Maximum		
	$\bar{x}$	σ_{n-1}	šikmost
Miláno	28.0936	1.3309	0.2635
Padova	27.7079	1.2843	0.4224
Belgie	23.6008	1.9263	0.1700
Cadiz	29.3642	1.4383	0.1011
Uppsala	22.1272	1.9472	0.1020
Stockholm	22.3939	1.9406	0.1060
St.Petersburg	23.8034	1.8190	-0.0098

Tabulka 2: Maximum – vlastnosti.

	Minimum		
	$\bar{x}$	σ_{n-1}	šikmost
Miláno	-4.4263	2.7081	-0.6107
Padova	-3.9374	2.6492	-0.7707
Belgium	-6.7985	3.3769	-0.2411
Cadiz	5.6300	2.2947	-0.8921
Uppsala	-17.8075	4.9149	-0.0585
Stockholm	-15.0208	4.2407	-0.4257
St.Petersburg	-23.0882	5.0987	-0.1833

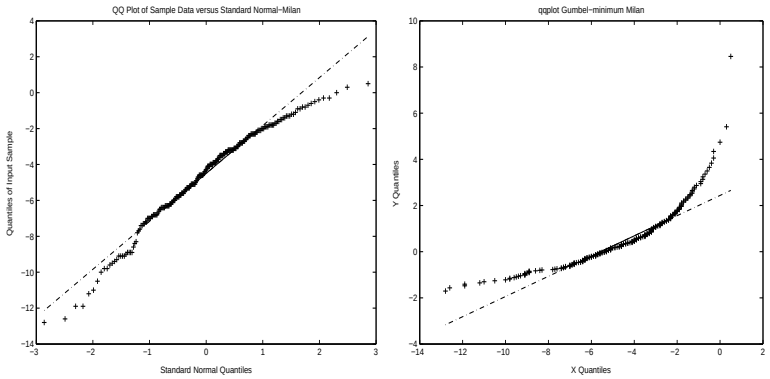
Tabulka 3: Minimum – vlastnosti.



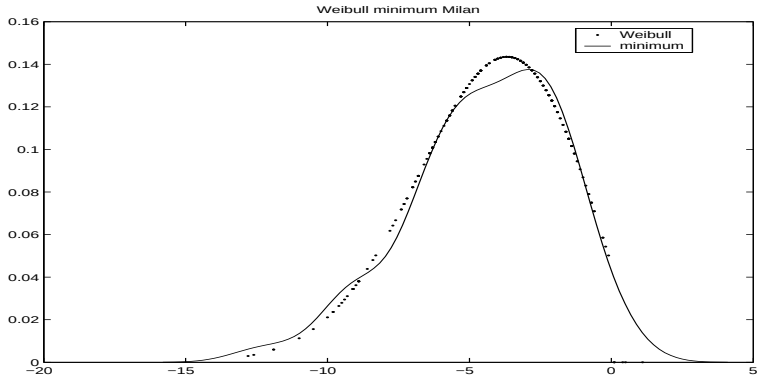
Obrázek 6: QQ plot - maximum Miláno, vlevo-normální rozdělení, vpravo-Gumbelovo rozdělení.

3 Maxima and minima

Když porovnáme tabulky 2 a 3, zjistíme, že pro minimální teploty vychází šikmost záporná a nabývá vyšších hodnot, zatímco pro maximální hodnoty vychází šikmost kladná s hodnotami nižšími, blízkými nule. Mohli bychom tedy usuzovat, že rozdělení maximální teploty je spíše normální než Gumbelovo. Zatímco minimální teploty se neřídí ani normálním ani Gumbelovým rozdělením. O tom se můžeme přesvědčit také srovnáním příslušných grafů na obrázcích 6 a 7.



Obrázek 7: QQ plot - minimum Miláno, vlevo-normální rozdělení, vpravo-Gumbelovo rozdělení.



Obrázek 8: Srovnání odhadnutého tříparametrického Weibullova rozdělení s původními daty.

4 Weibullovo rozdělení

Pro modelování nesymetrických dat, k nimž patří také maximální/minimální teploty, se často používá tříparametrické Weibullovo rozdělení, pro jehož hustotu platí

$$f(x;\theta,\alpha,\beta) = \alpha\beta(x-\theta)^{\alpha-1}\mathrm{e}^{-\beta(x-\theta)^{\alpha}} \qquad (\theta < x < \infty, \alpha > 0, \beta > 0).$$

Pro každou teplotní řadu byly nalezeny odhady tříparametrického Weibullova rozdělení. Např. pro Miláno jsme dostali odhady

$$\hat{\alpha} = 2.24521, \hat{\beta} = 0.01506, \hat{\theta} = -1.31.$$

Z obrázku 8 vyplývá, že tento model popisuje naše data poměrně dobře.

Reference

[1] Camuffo D., Jones P. (2002). *Improved understanding of past climatic variability from early daily European instrumental sources*. Climatic Change **53**, 1 – 3.

Adresa: M. Rencová, Katedra matematiky, FSV ČVUT, Thákurova 7, Praha 6

E-mail: rencova@mat.fsv.cvut.cz

TESTY A KONFIDENČNÉ INTERVALY PRE STREDNÚ HODNOTU V MODELOCH JEDNODUCHÉHO TRIEDENIA

Alexander Savin

Kľúčové slová: Medzilaboratórne štúdie, modely jednoduchého triedenia, heteroskedasticita.

Abstrakt: Kenward a Roger [8] navrhli metódu, pre testovanie hypotéz o strednej hodnote pomocou úpravy odhadu variancie neznámych parametrov strednej hodnoty a následnou aproximáciou Waldovej štatistiky v zovšeobecnenom lineárnom modeli. V texte je pojednávaná táto metóda pre model jednoduchého triedenia s pevnými efektmi v prípade heteroskedasticity, a porovnaná už so známymi metódami pre inferenciu o strednej hodnote spomenutých v [4].

1 Úvod

V medzilaboratórnych štúdiách uvažujeme, že merania na tom istom objekte záujmu sú vykonávané v k laboratóriach. V i-tom laboratóriu sa meranie opakuje n_i krát, $n_i \geq 2$. Laboratóriá sa môžu medzi sebou líšiť variabilitou, ako aj rozličnými medzi-laboratórnymi varianciami (heteroskedasticita). Dané laboratóriá môžeme mať buď pevne špecifikované (pevné efekty) alebo môžu byť vybrané náhodne z nejakeho počtu možných laboratórií(náhodné efekty, pozri [14]). V modeloch budeme uvažovať normálnu distribúciu meraní.

Ak máme pevne zadané laboratóriá, teda v modeli uvažujeme pevné efekty, predpokladáme $E(Y_{ij}) = \mu + \alpha_i = \mu_i$ pre $\forall i, j$.

Uvažujeme nasledujúci model s jedným faktorom:

$$Y_{ij} = \mu_i + \varepsilon_{ij} \tag{1}$$

$$= \mu + \alpha_i + \varepsilon_{ij}, \tag{2}$$

$i = 1, \ldots, k$ a $j = 1, \ldots, n_i$, pre $n_i \geq 1$, a kde predpokladáme: $E(e_{ij}) = 0$, $Var(\varepsilon_{ij}) = \sigma_i^2 \ \forall \ j,$ pre $i = 1, \ldots, k$ a $Cov(\varepsilon_{ij}, \varepsilon_{uv}) = 0, \ i \neq u, j \neq v$.

Pre túto situáciu by sme potrebovali zistiť, či objekt záujmu je meraný rovnako v jednotlivých laboratóriách, teda testujeme hypotézu o homogenite strednej hodnoty:

$$H_0 : \ \mu_1 = \mu_2 = \cdots = \mu_k = \mu \text{ oproti } H_1 : \ \exists \text{ také } i, j \text{ že } \mu_i \neq \mu_j. \tag{3}$$

Tejto historicky známej problematike sa niekedy hovoríaj zovšeobecnený Behrens–Fisherov problém: testovanie homogenity strednej hodnoty pri rozličných varianciách pre jednotlivé triedy (laboratóriá). V článku v [4] sú spomenuté niektoré testy. Náš príspevok do tejto problematiky je využitie meódy

Kenwarda a Rogera [8], navrhnutej pre zovšeobecnený lineárny model, pre tento model. Testovacia štatistika je:

$$F_{KR} = \frac{1}{k-1} \sum_{i=1}^{k} \frac{n_i}{S_i^2} \left[\bar{Y}_{i.} - \left(\sum_{j=1}^{k} \frac{n_j \bar{Y}_{j.}}{S_j^2} \right) \Big/ \left(\sum_{j=1}^{k} \frac{n_j}{S_j^2} \right) \right]^2, \qquad (4)$$

kde

$$\bar{Y}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}, \qquad S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}_{i.} \right)^2, \qquad \text{pre } i = 1, \ldots, k.$$

λF_{KR} má približné $\mathcal{F}$ rozdelenie s $k-1$ a m stupňami voľnosti, kde λ a m sa odhadujú nasledovne:

$$\lambda = \frac{m}{E[F]\,(m-2)}, \qquad (5)$$

$$m = 4 + \frac{k+1}{\left[Var[F]\,(k-1) \right] / \left(2E[F]^2 \right) - 1}, \qquad (6)$$

kde

$$E[F] = \left(1 - \frac{A}{k-1} \right)^{-1}, \qquad Var[F] = \frac{2\,(1 + c_1 B)}{(1 - c_2 B)^2\,(1 - c_3 B)},$$

$$A = \sum_{i=1}^{k} \frac{2}{n_i - 1} \left[1 - \left(\sum_{i=1}^{k} \frac{n_i}{\hat{\sigma}_i^2} \right)^{-1} \frac{n_i}{\hat{\sigma}_i^2} \right]^2, \qquad B = \frac{7A}{2\,(k-1)},$$

$$c_1 = \frac{-3}{3k^2 + 2k + 5}, \qquad c_2 = \frac{k^2 + 2}{3k^2 + 2k + 5}, \qquad c_3 = \frac{k^2 + 2k + 4}{3k^2 + 2k + 5}.$$

Na porovnanie uvedieme už známe používané metódy:

ANOVA *F*-test: Test je daný

$$F = \frac{N-k}{k-1} \frac{\sum_{i=1}^{k} n_i \left(\bar{Y}_{i.} - \bar{Y}_{..} \right)^2}{\sum_{i=1}^{k} (n_i - 1)\, S_i^2}, \qquad (7)$$

kde

$$\bar{Y}_{..} = \frac{\sum_{i=1}^{k} n_i Y_{i.}}{\sum_{i=1}^{k} n_i}.$$

Tento test testuje rovnosť stredov z jednotlivých tried má za platnosti hypotézy (3) H_0 $\mathcal{F}$ rozdelenie s $k-1$ a $N-k$ stupňami voľnosti. Tento test je pre danú problematiku nevhodný, aj keď sa v praxi môže vyskytnúť, nakoľko test predpokladá variančnú homogenitu, ktorá v heteroskedastickom modeli nie je splnená.

Brownov–Forsytheov test: Tento test je tiež známy ako modifikovaný F-test, je daný

$$B = \frac{\sum_{i=1}^{k} n_i \left(\bar{Y}_{i.} - \bar{Y}_{..}\right)^2}{\sum_{i=1}^{k} \left(1 - n_i/N\right) S_i^2}. \tag{8}$$

Brown a Forsythe [2] použili pre odvodenie svojho testu maximalizáciu cez kontrasty podobne, ako je odvodený ANOVA F-test. Pre odvodenie distribúcie štatistiky B, použili Satterthwaiteovu aproximáciu stupňov voľnosti. Ak je hypotéza (3) H_0 pravdivá B je distribuovaná približne ako $\mathcal{F}$ náhodná premenná s $k-1$ a ν stupňami voľnosti, kde

$$\nu = \frac{\left[\sum_{i=1}^{k} \left(1 - n_i/N\right) S_i^2\right]^2}{\sum_{i=1}^{k} \left(1 - n_i/N\right)^2 S_i^4 / \left(n_i - 1\right)}.$$

Modifikovaný Brownov–Forsytheov test: Mehrotra [9] v pokuse opraviť "chybu" v pôvodnom Brownovom–Forsytheovom teste vypracoval nasledujúci test

$$B^* = \frac{\sum_{i=1}^{k} n_i \left(\bar{Y}_{i.} - \bar{Y}_{..}\right)^2}{\sum_{i=1}^{k} \left(1 - n_i/N\right) S_i^2}, \tag{9}$$

Chyba v Brownovej–Forsytheovej testovacej procedúre, identifikovaná Mehrotrom [9], je v špecifikácii čitateľa stupňov voľnosti. Brown a Forsythe používajú v čitateli $k-1$ stupňov voľnosti, naproti tomu Mehrotra [9] používa Boxovu (1954) aproximáciu. Teda B^* má $\mathcal{F}$-distribúciu s ν_1 a ν stupňami voľnosti, kde

$$\nu_1 = \frac{\left[\sum_{i=1}^{k} \left(1 - n_i/N\right) S_i^2\right]}{\sum_{i=1}^{k} S_i^4 + \left[\sum_{i=1}^{k} n_i S_i^2/N\right]^2 - 2\sum_{i=1}^{k} n_i S_i^4/N}$$

a ν je dané vyššie.

Cochranov test: K odvodeniu testovacej štatistiky môžeme pristupovať aj iným spôsobom ako pomocou kontrastov v predchádzajúcich prípadoch. Ak uvažujeme neznáme parametre $\mu_1, \ldots, \mu_k$, ich odhady sú $\bar{Y}_{1.}, \ldots, \bar{Y}_{k.}$ s variancami $\sigma_1^2/n_1, \ldots, \sigma_k^2/n_k$. Ak sú variancie $\sigma_1^2/n_1, \ldots, \sigma_k^2/n_k$ známe, potom štatistika

$$Q = \sum_{i=1}^{k} \omega_i \left(\bar{Y}_{i.} - \sum_{i=1}^{k} h_i \bar{Y}_{i.}\right)^2, \tag{10}$$

kde $\omega_i = n_i/\sigma_i^2$ a $h_i = \omega_i / \sum_{j=1}^{k} \omega_j$, má za platnosti hypotézy (3) H_0 χ^2 distribúciu s $k-1$ stupňami voľnosti.

Po nahradenínezn´amych varianciíσ$_1^2/n_1, \ldots, \sigma_k^2/n_k$ ich odhadmi $S_1^2/n_1, \ldots, S_k^2/n_k$, dostávame štatistiku

$$C = \hat{Q} = \sum_{i=1}^{k} \hat{\omega}_i \left(\bar{Y}_{i.} - \sum_{i=1}^{k} \hat{h}_i \bar{Y}_{i.} \right)^2, \tag{11}$$

kde $\hat{\omega}_i = n_i/S_i^2$ a $\hat{h}_i = \hat{\omega}_i / \sum_{j=1}^{k} \hat{\omega}_j$, ktorá má za platnosti hypotézy (3) H_0 približné rozdelenie ako χ^2 náhodná premenná s $k-1$ stupňami voľnosti. Štatistika C bola uverejnená Cochranom [3] a neskôr modifikovaná Welchom [13].

Welchov test: Test je daný

$$W = \frac{\sum_{i=1}^{k} \hat{\omega}_i \left(\bar{Y}_{i.} - \sum_{i=1}^{k} \hat{h}_i \bar{Y}_{i.} \right)^2}{k - 1 + (k-2)(k+1)^{-1} \sum_{i=1}^{k} \left(1 - \hat{h}_i\right)^2 / (n_i - 1)}. \tag{12}$$

Welch [13] aproximoval rozdelenie W pomocou $\mathcal{F}$ náhodnej premennej. Za platnosti hypotézy (3) H_0, má štatistika W približne $\mathcal{F}$ distribúciu s $k-1$ a ν_W stupňami voľnosti, kde

$$\nu_W = \frac{k^2 - 1}{3 \sum_{i=1}^{k} \left(1 - \hat{h}_i\right)^2 / (n_i - 1)}.$$

2 Simulačné štúdie

V tejto simulačnej štúdii chceme porovnať sily už známych testov pre testovanie *hypotézy o homogenite strednej hodnoty* s testom odvodeným metódou Kenwarda a Rogera [8].

Realizovali sme $N = 10000$ krát postačujúce štatistiky Y_i a S_i^2 pre výpočet testu pre našu hypotézu o homogenite, kde parametre do simuláciísme vyberali týmito spôsobmi:
$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)'$ sme uvažovali nasledovne

- pre nasimulovanie hladiny významnosti za platnosti hypotézy (3) H_0

$$\mu_i = 0 \ \forall \ i = 1, \ldots, k;$$

- pre nasimulovanie sily jednotlivých testov za platnosti alternatívnej hypotézy (3) H_1

$$\mu_i = \begin{cases} 2, & \text{pre } i = 3j+1 \\ 0, & \text{pre } i \neq 3j+1 \end{cases} \quad \text{a} \quad \mu_i = \begin{cases} -1, & \text{pre } i = 3j+1 \\ 0, & \text{pre } i = 3j+2 \\ 1, & \text{pre} i = 3j+3, \end{cases}$$

kde $j = 0, \ldots, k/3 - 1$.

V ďalšom sme uvažovali $k \in \{3, 6, 9, 15, 21\}$. Vzory σ_i^2 a n_i boli brané podľa nasledujúcej tabuľky:

	j	1	2	3		j	1	2	3
A	n_j	5	5	5	C	n_j	10	20	30
	σ_j^2	4	4	4		σ_j^2	4	4	4
	n_j	5	5	5		n_j	10	20	30
	σ_j^2	2	6	10		σ_j^2	2	6	10
	n_j	5	5	5		n_j	10	20	30
	σ_j^2	10	6	2		σ_j^2	10	6	2
B	n_j	5	10	15	D	n_j	30	30	30
	σ_j^2	4	4	4		σ_j^2	4	4	4
	n_j	5	10	15		n_j	30	30	30
	σ_j^2	2	6	10		σ_j^2	2	6	10
	n_j	5	10	15		n_j	30	30	30
	σ_j^2	10	6	2		σ_j^2	10	6	2

Tabulka 1: Tabuľka použitých dizajnov.

Výsledky simuláciísú uvedené v nasledujúcich tabuľkách. Pre rozsiahlosť výsledkov a počet tabuliek, uvedieme výsledky simulaciípre $k \in \{3, 9\}$:

	$\hat{\alpha}\%$					
	F	B	B^*	C	W	F_{KR}
A	4.88	3.96	3.68	12.24	4.52	4.02
	6.42	5.17	4.67	13.16	5.11	4.59
	5.72	4.49	4.09	12.99	4.89	4.30
B	5.01	4.96	4.61	10.03	5.36	5.19
	2.36	5.47	4.58	8.82	5.13	5.02
	12.49	5.19	4.94	11.75	5.71	5.38
C	5.04	5.05	4.90	7.54	5.12	5.06
	2.27	5.57	4.62	6.52	4.76	4.73
	12.96	5.61	5.19	8.59	5.20	5.15
D	5.13	5.12	5.03	6.07	5.18	5.16
	6.23	6.08	5.33	6.40	5.20	5.19
	5.97	5.81	5.15	6.48	5.36	5.35

Tabulka 2: Aktuálne nasimulované hladiny významnosti (5%) pre $k = 3$.

3 Diskusia

Simulácie ukázali, že klasický ANOVA F-test a Cochranov test nám dávajú zlé vysledky. Ostatné testy nám dávajú dobré a porovnateľné výsledky. Welchov test sa pri malom počte opakovanív triedach a narastajúcom počte tried chová liberárne.

Test založený na metóde Kenwarda a Rogera [8] je porovnateľný s už známymi testmi, (Brownov–Forsytheov test, Welchov test). Oproti nim má síce nižšiu silofunkciu, pri malom počte opakovanív triedach a narastajúcom počte tried sa chová konzervatívne, ale vykazuje stabilitu pri odhadovaníchýb prvého druhu.

	$\hat{\alpha}\%$					
	F	B	B^*	C	W	F_{KR}
A	5.21	4.17	3.18	29.42	7.82	3.23
	6.72	5.00	3.34	29.58	7.62	3.51
	7.33	5.54	3.89	29.18	8.06	3.78
B	5.29	4.69	3.77	19.73	7.33	5.63
	2.03	6.15	4.08	17.88	6.39	4.96
	18.19	5.46	4.06	20.28	7.77	5.99
C	4.92	4.75	4.07	10.74	5.26	4.97
	2.45	7.05	4.96	10.53	5.51	5.22
	18.70	6.52	4.88	11.24	5.36	4.94
D	5.29	5.26	4.84	7.63	5.29	5.13
	6.88	6.72	4.97	8.07	5.40	5.24
	6.61	6.48	4.86	7.72	5.35	5.26

Tabulka 3: Aktuálne nasimulované hladiny významnosti (5%) pre $k = 9$.

	$\hat{\alpha}\%$					
	F	B	B^*	C	W	F_{KR}
A	28.21	25.06	24.02	24.02	44.41	22.50
	20.60	16.39	15.34	15.34	47.05	22.63
	20.73	17.43	16.39	16.39	30.07	13.56
B	39.59	37.01	35.85	38.91	49.38	35.30
	14.47	28.12	24.46	24.11	58.09	45.88
	39.36	21.44	20.43	24.39	30.86	17.84
C	70.53	69.13	68.35	69.89	73.04	66.86
	41.25	62.57	58.11	57.74	85.66	81.90
	60.91	42.77	41.01	43.50	43.74	34.91
D	98.03	98.02	97.94	97.94	98.29	97.87
	94.44	94.26	93.13	93.13	98.62	98.26
	86.48	86.15	84.74	84.74	82.59	80.16

Tabulka 4: Nasimulovaná sila testov pre $k = 3$, $\mu_j = 2$, $\mu_{j+1} = 0$, $\mu_{j+2} = 0$.

	$\hat{\alpha}\%$					
	F	B	B^*	C	W	F_{KR}
A	22.75	20.44	19.44	19.44	36.89	17.61
	17.74	14.61	13.56	13.56	32.56	14.64
	16.76	13.57	12.60	12.60	32.10	13.91
B	38.27	35.28	33.99	37.19	48.36	32.91
	16.41	28.29	25.47	25.08	43.82	32.13
	40.91	21.30	20.29	24.20	40.16	23.39
C	71.23	69.71	68.88	70.39	74.70	67.57
	39.86	57.54	53.67	53.37	70.97	65.67
	65.23	42.97	40.99	44.16	59.60	50.11
D	93.65	93.62	93.49	93.49	94.29	93.29
	80.15	79.66	77.55	77.55	87.20	85.34
	79.40	78.95	76.93	76.93	87.09	85.08

Tabulka 5: Nasimulovaná sila testov pre $k = 3$, $\mu_j = -1$, $\mu_{j+1} = 0$, $\mu_{j+2} = 1$.

	$\hat{\alpha}\%$					
	F	B	B^*	C	W	F_{KR}
A	48.65	44.28	39.00	39.00	78.86	25.80
	33.13	26.66	20.26	20.26	80.31	27.60
	32.69	27.08	21.62	21.62	59.34	13.77
B	65.52	62.55	58.31	61.14	81.62	54.73
	24.55	46.11	36.94	36.37	90.30	69.94
	64.46	36.11	30.28	34.34	56.06	28.14
C	96.09	95.70	95.29	95.45	97.38	93.54
	71.66	89.49	84.39	84.11	99.65	98.83
	89.33	73.92	68.55	70.54	73.58	59.58
D	100.0	100.0	100.0	100.0	100.0	100.0
	99.99	99.99	99.98	99.98	100.0	100.0
	99.47	99.47	99.21	99.21	99.23	98.57

Tabulka 6: Nasimulovaná sila testov pre $k = 9$, $\mu_j = 2$, $\mu_{j+1} = 0$, $\mu_{j+2} = 0$.

	$\hat{\alpha}\%$					
	F	B	B^*	C	W	F_{KR}
A	36.63	32.44	27.75	27.75	69.62	18.95
	25.52	20.71	15.59	15.59	62.34	15.17
	24.93	19.55	15.22	15.22	63.06	14.83
B	64.50	61.00	56.85	59.57	80.40	52.28
	26.82	46.22	38.94	38.57	77.28	50.05
	66.97	34.89	29.00	33.29	68.56	36.72
C	95.80	95.60	95.06	95.27	97.33	93.80
	66.47	83.81	78.66	78.44	95.96	91.68
	92.06	76.60	70.69	72.72	89.69	80.50
D	99.95	99.95	99.93	99.93	99.96	99.94
	98.77	98.71	97.98	97.98	99.72	99.49
	98.83	98.76	97.90	97.90	99.76	99.53

Tabulka 7: Nasimulovaná sila testov pre $k = 9$, $\mu_j = -1$, $\mu_{j+1} = 0$, $\mu_{j+2} = 1$.

Ďalšia možnosť pri riešenítohoto problému je rozvinúť už známe odvodené metódy ako Merhotra [9] Brownov–Forsytheov test. Alternatívne možnosti, ktoré sa otvárajú je použitie metód založených na zovšeobecnených p-hodnotách.

Reference

[1] Böckenhoff A., Hartung J. (1998). *Some corrections of the significance level in meta-analysis*. Biometrical Journal **40**, 937 – 947.

[2] Brown M.B., Forsythe A.B. (1974). *The small sample behavior of some statistics which test the equality of several means*. Technometrics **16**, 129 – 132.

[3] Cochran W.G. (1937). *Problems arising in the analysis of a series of similar experiments*. J. Roy. Stat. Soc. Supp. **4**, 102 – 118.

[4] Hartung J., Argaç D., Makambi K.H. (2002). *Small sample properites of test on homogenity in one-way ANOVA and meta-analysis*. Preprint. Department of Statistics, University of Dortmund.

[5] Hartung J., Makambi K.H. (2002). *Alternative test procedures and confidence intervals on the common mean in the fixed effects model for meta-analysis*. Preprint. Department of Statistics, University of Dortmund.

[6] Iyer H.K., Wang C.M., Mathew T. (2002). *Models and confidence intervals for true values in interlaboratory trials*. Submitted to the Journal of the American Statistical Association.

[7] Kackar A.N., Harville D.A. (1984). *Approximations for standard errors of estimators of fixed and random effects in mixed linear models*. Journal of the American Statistical Association **79**, 853 – 862.

[8] Kenward M.G., Roger J.H. (1997). *Small sample inference for fixed effects from restricted maximum likelihood*. Biometrics **53**, 983 – 997.

[9] Mehrotra D.V. (1997). *Improving the Brown–Forsythe solution to the generalized Behrens–Fisher problem*. Commun. Statist.–Simula. **26**, 1139 – 1145.

[10] Rukhin A.L., Biggerstaff B.J., Vangel M.G. (2000). *Restricted maximum likelihood estimation of a common mean and the Mandel–Paule algorithm*. Journal of Statistical Planning and Inference **83**, 319 – 330.

[11] Rukhin A.L., Vangel M.G. (1998). *Estimation of a common mean and weighted means statistics*. Journal of the American Statistical Association **93**, 303 – 308.

[12] Savin A., Wimmer G., Witkovský V. (2003). *On Kenward–Roger confidence intervals for common mean in interlaboratory trials*. Measurement Science Review, Theoretical Problems Of Measurement Vol. **3**, 53 – 56, `http://www.measurement.sk`.

[13] Welch B.L. (1951). *On the comparison of several mean values: An alternative approach*. Biometrika **38**, 330 – 336.

[14] Witkovský V., Savin A., Wimmer G. (2003). *On small sample inference for common mean in heteroscedastic one-way model*. Discussiones Mathematicae, Probability and Statistic **23**, 123 – 145.

Poďakovanie: Výskum bol podporený grantmi z Vedeckej grantovej agentúry Slovenskej Republiky 1/0264/03 a 2/4026/04.

Adresa: A. Savin, Ústav merania, Slovenská akadémia vied,
Dúbravská cesta 9, 841 04 Bratislava, Slovenská Republika

E-mail: `savin@savba.sk`

HISTORIE GRAFICKÉHO ZOBRAZOVÁNÍ STATISTICKÝCH DAT

Ivan Saxl, Lucia Ilucová

Klíčová slova: tématická kartografie, statistická grafika.

Abstrakt: *Tématická kartografie* (mapy doplněné daty) je stará pouze několik málo staletí. *Statistická grafika* začíná v XVII. století, k jejímu systematickému rozvoji dochází však až koncem století XVIII. Zpočátku mají grafy vesměs politicko-ekonomickou tématiku. Velký rozmach grafického zobrazování dat probíhá v XIX. století a je dílem francouzských stavebních inženýrů, většinou žáků Gasparda Monge. Souběžně se grafika uplatňuje i ve společenských studiích, v epidemiologii, v biologii a grafy se objevují již i ve školních učebnicích.

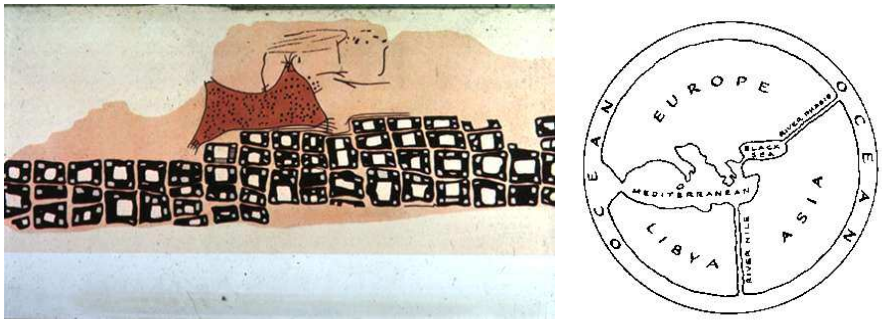
1 Úvod

Grafická reprezentace dat se v současné době těší mimořádné pozornosti. Vedle jejího praktického rozvíjení sofistikovanými počítačovými programy probíhá také podrobné studium její minulosti. Na Internetu lze nalézt skoro každý významnější graf z minulosti a existuje řada adres obsahujících detailní chronologické přehledy umožňující prohlédnutí a obvykle i stažení stovek komentovaných grafů včetně popisu okolností jejich vzniku a životopisných medailonů jejich autorů (viz [1]-[7]). Z nejvýznamějších časopisecky či knižně publikovaných prací lze uvést především [6], [8], [11]-[13]. Na požadavek „graphical statistics" poskytne vyhledávač Google 988 000 odkazů, další tisíce produkují hesla „statistical graphics", „statistical graphs" atd. Moderní přístupy jsou zachyceny např. v [5], [7], [9], [14], [15].

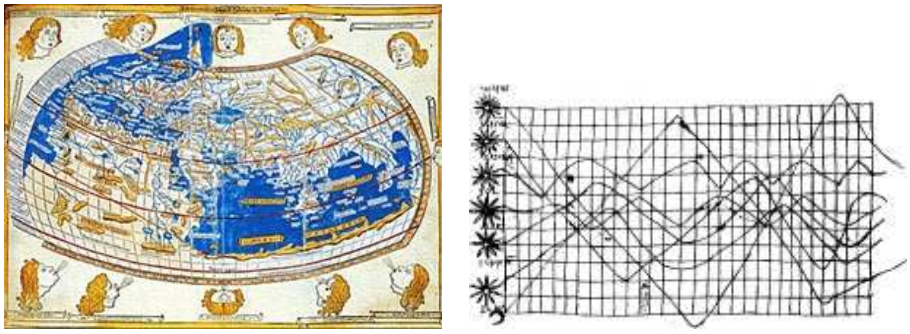
Proč graficky zobrazovat statistická data? W.S. Jevons komentuje své časové diagramy, v nichž sleduje změny cen základních i méně běžných produktů v závislosti na „komerčních bouřích" typu objevení australského zlata v roce 1849 takto: „Jejich smyslem není ani odkaz ke konkrétním číslům, která lze lépe zjistit z odpovídajících tabulek, jako předvést očím obecné výsledky vyplývající z velkého množství číslic, jež nemohou být zachyceny jinak než graficky. Mé diagramy ukazují i ty nejmenší detaily tabulek, ale předčí i výpočty středních hodnot, protože oko či mysl samy zaznamenají obecný trend číselných souborů. Pouze tato reprezentace může být základem politicko-ekonomických debat a přesto většina statistických zdůvodnění závisí na pár číslech více či méně náhodně vybraných."[1]

Samotné slovo *graf* je poměrně nové (objevilo se až koncem XIX. století), předtím se pro grafickou prezentaci dat převážně používalo slov *mapa* a *diagram*.

[1] R. D. Block (edit.) *Papers and correspondence of William Stanley Jevons*, vols. 1–7, Macmillan, London 1972–1981, vol. 2, 450. Dopis R. Huttonovi z 1. 9. 1862.



Obrázek 1: Rekonstrukce fresky nalezené v Catal Hüyüku a Anaximandrovy mapy světa.



Obrázek 2: Ptolemaiova mapa světa a diagram poloh planet (pořadí:Venuše, Merkur, Saturn, Slunce!, Mars, Jupiter, Měsíc).

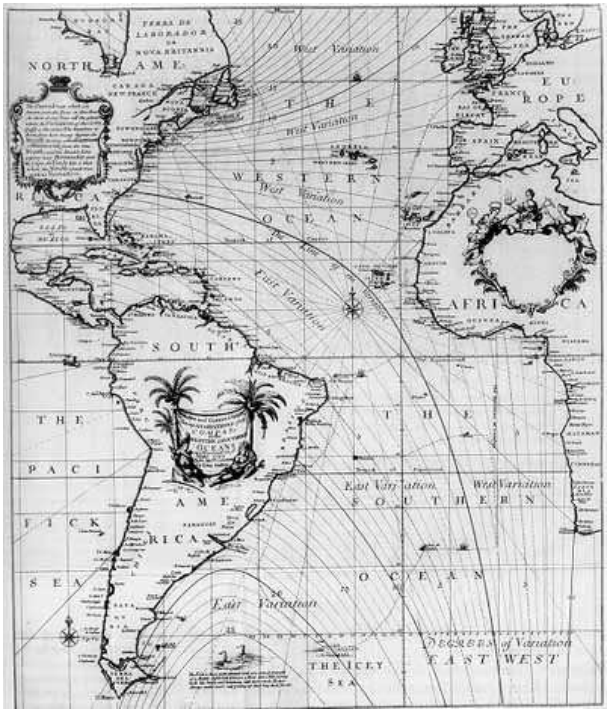
Práce jako celek je vychází především z citací [1]-[4], [6], [8], [10], [11], [13], které nejsou v textu explicitně uváděny. Citace úzce související s konkrétním textem jsou uváděny v poznámkách.

2 Kartografie a tématická kartografie

Nejstarší formou grafického znázornění dat jsou mapy po tisíciletí zobrazující jednak pozemské oblasti, jednak výřezy hvězdné oblohy. Nejstarší známou mapou je část plánu města (patrně Catal Hüyük) objeveného jako freska při vykopávkách v letech 1961 až 1965 na pláni Konya v Anatolii a datovaného uhlíkovou metodou do let 6250 až 6400 př. Kr., podle letokruhů dokonce 7100 až 7200 př. Kr. – Obr. 1. Autorem údajně první (podle Hérodotova popisu rekonstruované) mapy světa je Anaximandros z Milétu [2] – Obr.1.

Klaudios Ptolemaios je tvůrcem prvních map se zakreslenými poledníky i rovnoběžkami – Obr. 2. Nejstarší zobrazení planetárních pohybů pochází

[2]Stručné životopisné údaje o autorech jsou uvedeny v Dodatku.

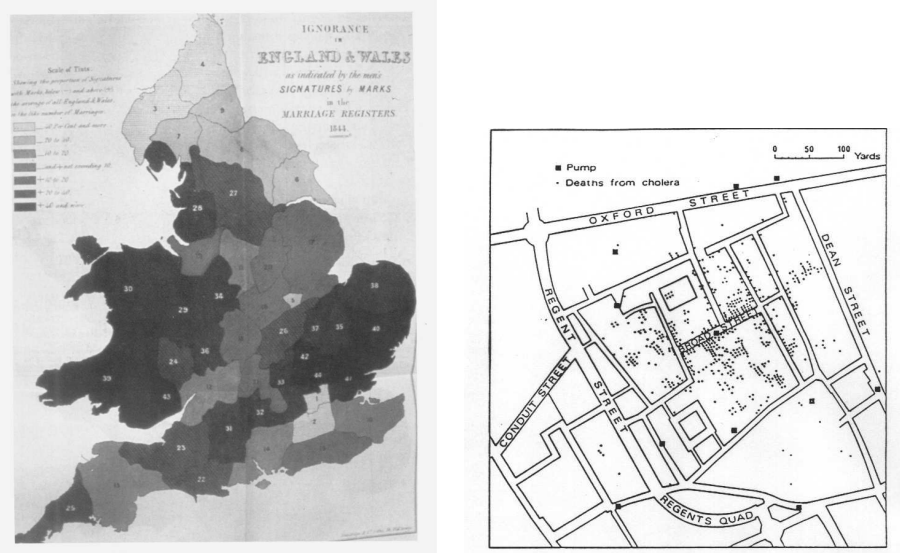


Obrázek 3: Halleyova mapa isogonál v Atlantickém oceánu (1701).

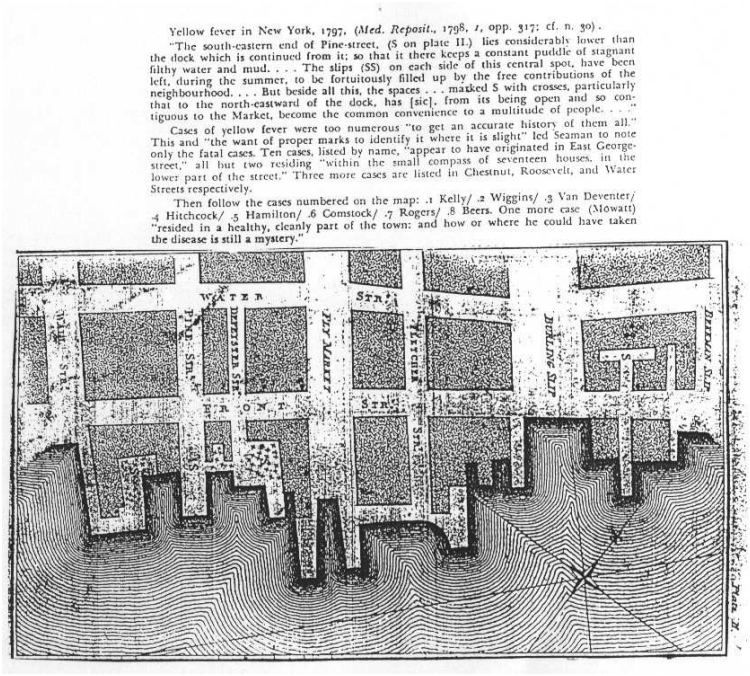
od neznámého autora z doby kolem roku 950 a zachycuje změny poloh Slunce a hvězd v průběhu roku – Obr. 2.

Dalším vývojovým momentem je zakreslování doplňujících charakteristik; E. Halley roku 1701 publikuje mapu se zakreslenými isogonálami spojujícími místa se stejnou magnetickou deklinací – Obr. 3. Tím začíná obor *tématické kartografie*, v níž jsou do map vedle územního členění zanášena data vztahující se k obyvatelstvu, obchodu, dopravě i k historickým událostem. Na jejím počátku jsou mapy analfabetismu ve Francii (P. Ch. F. Dupin: *Carte de la France eclairée et de la France obscure*, 1819) a v Anglii (J. Fletcher: *Distribution of ignorance in England*, 1834) založené na průzkumu matrik (záznamy sňatků analfabetů mají značky místo podpisů) – Obr. 4.

Zřejmě sem patří také slavný plán Londýna vytvořený Johnem Snowem v roce 1854 za účelem objasnění příčiny cholerové epidemie – Obr. 4. Zakreslením poloh studní a bydlišť nemocných se podařilo lokalizovat nakaženou studnu a zjistit způsob šíření nákazy, který do té doby nebyl bezpečně znám. Prvenství v mapování šíření epidemie však patří Valentinu Seamanovi, který publikoval podobnou mapu jako J. Snow v souvislosti s epidemií žluté horečky v New Yorku v roce 1795 (Obr. 5) a několik map výskytu cholery bylo publikováno v Anglii již v první polovině XIX. století.



Obrázek 4: Fletcherova mapa analfabetismu v Anglii (1834) a Snowův plán Londýna v době cholerové epidemie (1854).

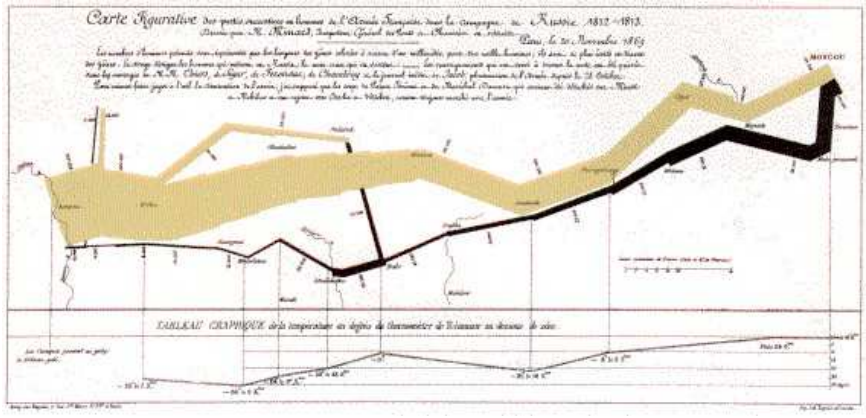

Yellow fever in New York, 1797, (*Med. Reposit.*, 1798, *1*, opp. 317; cf. n. 30).

"The south-eastern end of Pine-street, (S on plate II.) lies considerably lower than the dock which is continued from it; so that it there keeps a constant puddle of stagnant filthy water and mud. . . . The slips (SS) on each side of this central spot, have been left, during the summer, to be fortuitously filled up by the free contributions of the neighbourhood. . . . But beside all this, the spaces . . . marked S with crosses, particularly that to the north-eastward of the dock, has [sic], from its being open and so contiguous to the Market, become the common convenience to a multitude of people. . . ."

Cases of yellow fever were too numerous "to get an accurate history of them all." This and "the want of proper marks to identify it where it is slight" led Seaman to note only the fatal cases. Ten cases, listed by name, "appear to have originated in East George-street," all but two residing "within the small compass of seventeen houses, in the lower part of the street." Three more cases are listed in Chestnut, Roosevelt, and Water Streets respectively.

Then follow the cases numbered on the map: .1 Kelly/ .2 Wiggins/ .3 Van Deventer/ .4 Hitchcock/ .5 Hamilton/ .6 Comstock/ .7 Rogers/ .8 Beers. One more case (Mowatt) "resided in a healthy, cleanly part of the town: and how or where he could have taken the disease is still a mystery."

Obrázek 5: Seamanova mapa výskytu žluté horečky v New Yorku (1795).



Obrázek 6: Napoleonovo tažení na Moskvu od Ch. J. Minarda (1869). Šířka stopy znázorňuje početní sílu armády, graf ve spodní části mapy udává průběh teploty při jejím ústupu.

Nepřekonaným vrcholem co do emocionální působnosti je „nejslavnější mapa všech dob“, *Napoleonovo tažení na Moskvu* Charlese Josepha Minarda z roku 1869 – Obr. 6. V současné době jsou nejběžnějším produktem tématické kartografie mapy zachycující okamžitý stav počasí, publikované v denním tisku a v televizi.

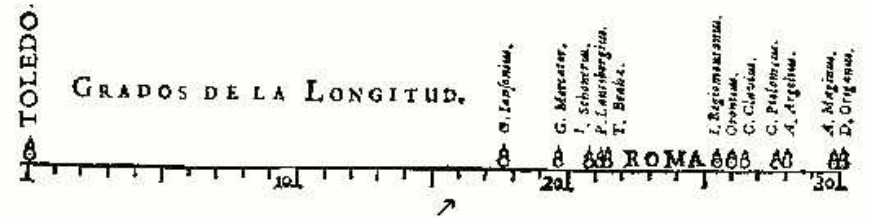
3 Statistická grafika

Statistická grafika prezentuje nejrůznější data v závislosti na zvoleném parametru, jímž bývá velmi často čas. Samotné slovo *graf* do angličtiny zavedl J. J. Sylvester v roce 1878 v souvislosti s konstatováním podobnosti mezi schématy molekulárních vazeb a grafickou reprezentací algebraických invariantů. Zhruba v téže době definuje graf Charles S. Peirce jako „plošný diagram sestávající z bodů či jejich ekvivalentů a jejich spojnice na omezené ploše“. Potřeba takové definice ukazuje, do jaké míry byly grafy ještě koncem XIX. století málo běžným informačním prostředkem.

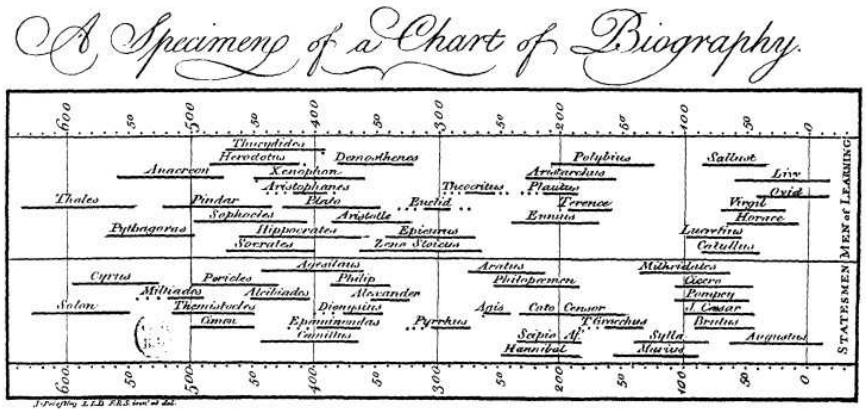
Jejich počáteční rozvoj byl do značné míry ovlivněn, ne-li podmíněn, několika vynálezy umožňujícími grafické zaznamenávání kontinuálně probíhajících fyzikálních procesů. Prvním z nich je Christopherem Wrenem vynalezený zapisovač počasí (*weather-clock*) zaznamenávající teplotu a směr větru v polárních souřadnicích, dalším Wattův indikátor tlaku v parním stroji.

3.1 Vývoj grafického zobrazování v XVIII. století

Za zakladatele statistické grafiky je obecně považován William Playfaire. Ve svých grafikách, které nazýval „čárovou aritmetikou“ (*lineal arithmetics*),



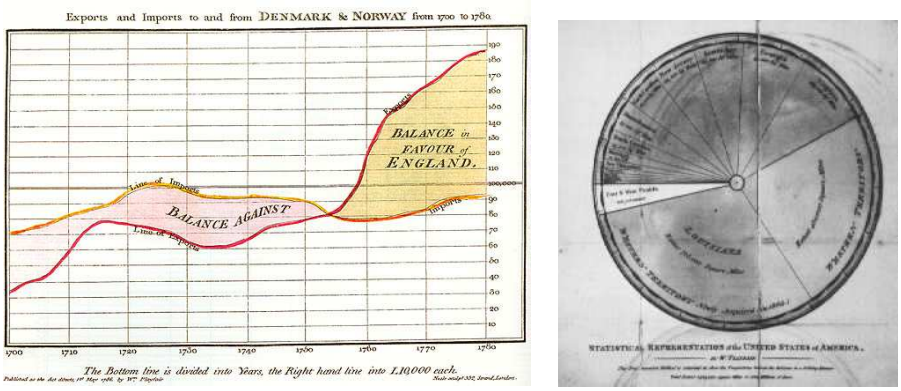
Obrázek 7: Van Langrenův diagram rozdílů zeměpisných délek Toleda a Říma podle různých autorů (1644); šipka ukazuje ke správné hodnotě 16°30', nápis ROMA je v místě vychýlené střední hodnoty historických kartografických údajů.



Obrázek 8: Druhá čtvrtina diagramu J. Priestleye (1765).

využíval převážně kartézské souřadné soustavy, v níž znázorňoval závislosti jedné i více zvolených veličin na vybraném parametru, jímž byl nezřídka čas. Měl však řadu předchůdců. Prvním z nich je Michael F. van Langren publikující v roce 1644 srovnání rozdílů zeměpisných délek Říma a Toleda – Obr. 7. Údaj známý v jeho současnosti je srovnán s vesměs pozitivně vychýlenými odhady získanými z různých historických map.

Druhým ranným grafem je *Životopisná mapa* (*Chart of Biography*, 1765) – Obr. 8, jejíž autorem je neobyčejně plodný vynálezce, vědec, teolog a politik Joseph Priestley. Opět se jedná v podstatě o jednorozměrný diagram prezentující životní rozpětí 2000 významných osobností žijících v letech 1200 př. Kr. až 1750. Priestley na čtyřech stránkách komentáře přesvědčuje čtenáře, že toto znázornění času je možné a účelné. Zatímco dnes je tento přístup považován za zcela přirozený, v polovině XVIII. století tomu bylo jinak. Kartézské souřadnice byly obecně přijaty jako systém vhodný pro znázornění prostoru, v němž existuje pozorovatelný materiální svět (nezavedl je však Descartes, ale

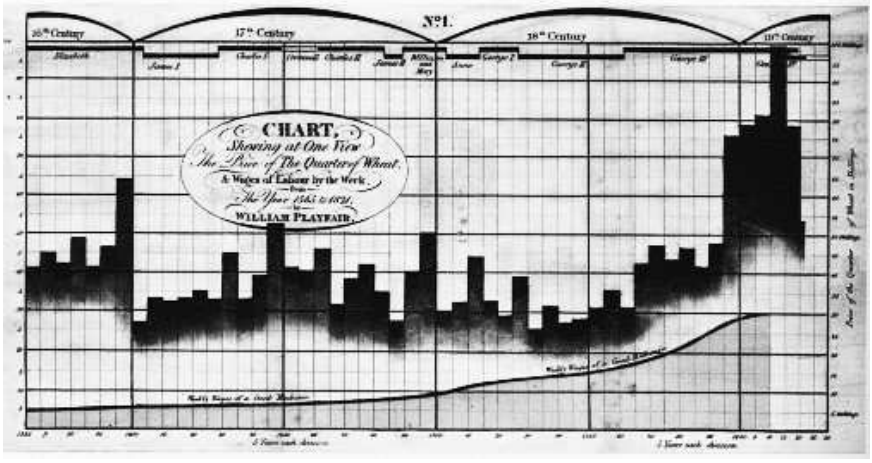


Obrázek 9: Obchodní bilance mezi Anglií a Norskem s Dánskem (1786) a kruhový diagram amerických států (1805) od W. Playfaira.

Leonardo da Vinci kolem roku 1500 pro analýzu rychlosti padání objektů). Historický čas však byl považován za jev subjektivní, vázaný na schopnost myšlení a sám Descartes zdůrazňoval „nezbytnost úplného abstrahování od analogií s hmotou při studiu zákonitostí Mysli".

Skotský filosof Dugald Stewart ve stati *A general View of the Progress of Metaphysical, Ethical, and Political Philosophy since the Revival of Letters* (1811) napsané pro Dodatky k Britské encyklopedii konstatuje, že historie, jako znalost určitých faktů a dějů, je především záležitostí naší paměti, která je subjektivní. Historické děje (a s nimi také ekonomické, populační aj.) sice mohou být a nejspíš jsou podřízeny nějakým zákonům, ty však nelze zjistit pozorováním jako zákony přírodní, ale pouze reflexí, uvažováním. Speciálně ekonomický stav státu je důsledkem subjektivního jednání lidí v jejich soukromých životech; to může probíhat např. na základě „zdravého rozumu". Protože první kroky grafické statistiky se odbývaly právě na půdě historie a politické ekonomie, byl pro ni význam chápaní historického času zcela podstatný a jeho subjektivní chápaní bylo velkou překážkou jejího obecného rozšíření.

William Playfair byl schopný vynálezce, ale jeho hlavní zájmy byly finance a obchod, v nichž však byl spíše neúspěšný, a dále publicistika, která jej dovedla ke statistické grafice, jíž se proslavil. V této oblasti mohla být jeho inspirací jednak spolupráce s J. Wattem, u nějž pracoval jako kreslič a návrhář, jednak rady jeho bratra, matematika a geologa. Od něj se podle vlastního sdělení naučil, že všechno, co lze vyjádřit čísly, může být vyjádřeno také rovnými čarami. Mezi jeho významné práce patří graf růstu britského národního dluhu v letech 1699 až 1800, grafy vzájemného obchodu mezi Anglií a různými státy (např. s Německem, s Dánskem a Norskem – Obr. 9), histogram zahraničního obchodu Skotska aj. Populární jsou také jeho grafy, v nichž upozorňoval na vysoké daňové zatížení Angličanů (Obr. 11) a první



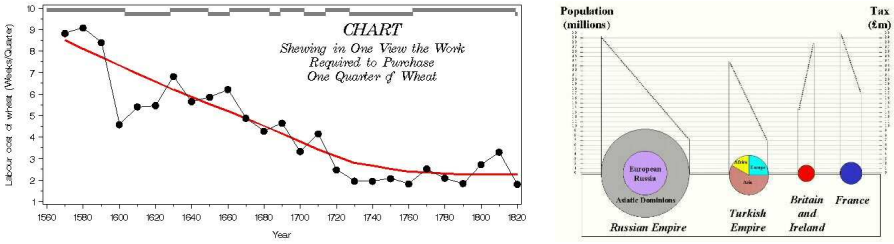
Obrázek 10: Srovnání cen pšenice (histogram uprostřed) a mzdy mechanika (graf v dolní části) za vlády různých panovníků (graf v horní části) od W. Playfaira (1821).

kruhový diagram rozlohy amerických států – Obr. 9. K nejznámějším patří mimořádně sugestivní graf porovnávající ceny pšenice a mzdy řemeslníků na pozadí vlád jednotlivých britských panovníků v letech 1665 až 1821 – Obr. 10. Pozoruhodné je, že právě tento graf na první pohled nesděluje autorův záměr a hrozivě rostoucí černý histogram (termín *histogram* však zavedl až K. Pearson[3]) mu spíše protiřečí – Obr. 11. Playfairovou snahou bylo totiž podle jeho vlastního vyjádření[4] ukázat, že nikdy nebyla pšenice tak levná jako na počátku XIX. století. To je však patrné teprve tehdy, když je vynesen graf poměru cen a mezd, který skutečně klesá od devíti ke dvěma. Playfairův graf tak ukazuje jednu z charakteristických vlastností grafického zobrazení, totiž na možnost vytvoření dojmu na první pohled opačného, než odpovídá skutečnému obsahu dat.

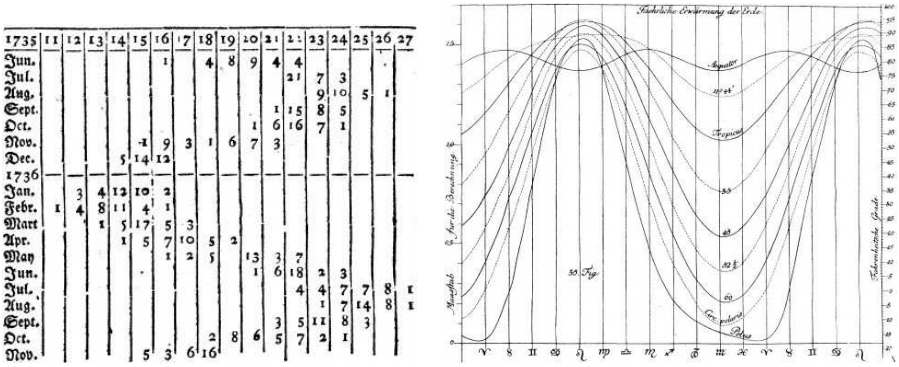
Playfairovy práce jsou shrnuty v knize *The Commerical and Political Atlas* vydané v Londýně roku 1786 a obsahující 44 diagramů; s výjimkou jediného se jedná o časové závislosti. Tím je sloupcový diagram zachycující obchod mezi Skotskem a 13 jinými státy a Playfairovi se podařilo získat data pouze pro jediný rok (1780), takže nemohl vynést časovou závislost. V úvodu to komentuje jako nedostatek („... it does not comprehend any portion of time, and is much inferior in utility to those that do."). Ve třetím vydání *Atlasu* v roce 1801 však sloupcový graf již vyzdvihuje jako typický produkt

[3] Viz poznámku na str. 399 v jeho článku ve Phil. Trans. Roy. Soc. A 186 (1895).

[4] Z Playfairova komentáře ke grafu (citováno podle H. Wainer: Visual revelations, Chance 17 (2004), 51–54, který je také autorem ukázaného poměrového grafu – Obr. 11): „...the main fact deserving consideration is, that never at any former period was wheat so cheap, in proportion to mechanical labour, as it is in the present time...".



Obrázek 11: Poměr ceny pšenice a mzdy řemeslníka podle předcházejícího Playfairova grafu (viz pozn.[4]) a jeho srovnání daňového zatížení občanů čtyřech evropských států (překreslená část původního grafu z roku 1801).

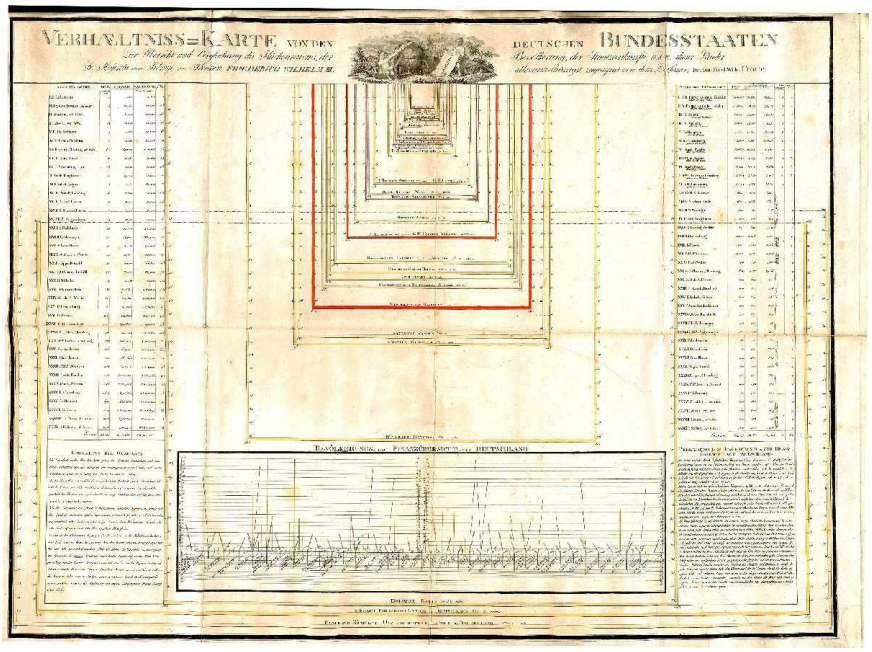


Obrázek 12: Lambertův „číselný graf“ (1779) udávající počet dnů v měsíci, kdy bylo dosaženo určité teploty (horizontální osa ve °C) a graf závislosti ohřevu půdy v průběhu roku na zeměpisné šířce (1779).

své „čárové aritmetiky“. Jako příklad uvádí muže, který denně vydělá jistý sloupec guinejí a jehož výsledná výška je potom rovna součtu výdělků za určitý čas, který je tak v zobrazení implicitně zahrnut.

Jedním z prvních uživatelů grafického zobrazení dat byl také alsaský přírodovědec Johann Heinrich Lambert, jehož hlavním zájmem byla fotometrie a fyzikální či astronomická měření. Byl patrně první, kdo vytvořil „číselný graf“ vhodným rozmístěním číselných hodnot v rovině – Obr. 12.

S dalším propagátorem grafických metod se vrací problematika sociálních a politických věd. August Friedrich Wilhelm Crome byl profesorem politických věd v Gießenu a je známý jednak svými knihami (např. *Über die Große und Bevölkerung der europäischen Staaten* z roku 1785), jednak řadou pamfletů, v nichž vedl vášnivé politické diskuse a své názory často dokazoval graficky zpracovanými statistickými údaji. Pomocí diagramů různých typů porovnával situaci v jednotlivých státech, např. velikost států znázorňuje pomocí pravidelných obrazců (čtverců, obdélníků či kruhů) o plochách



Obrázek 13: Srovnání společenských a hospodářských charakteristik německých států od A. F. W. Crome (1821); plochy obdélníků jsou úměrné jejich rozlohám.

úměrných rozlohám států, takže optický dojem není zkreslen komplikovaným průběhem hranic – Obr. 13 (autorem prvního takového grafu byl však Charles de Fourcroy; v práci *l'Essay d'une table poléographique* z roku 1782 srovnává rozlohy evropských měst čtvercovým diagramem podobným obr. 13 - viz [12]). Slavná je také jeho mapa *Produkten-Karte von Europa* z roku 1782, znázorňující vedle měst a přístavů také přírodní a průmyslovou produkci v jednotlivých zemích.

3.2 Nové směry statistické grafiky v XIX. století

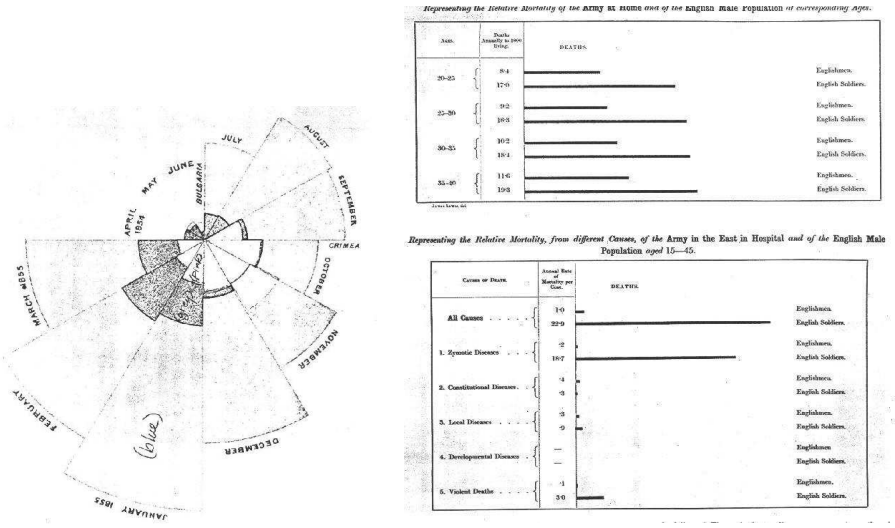
Poté, co se grafické zobrazování začalo v širší míře používat, vyskytla se potřeba technických prostředků, které by usnadňovaly jeho realizaci a šíření. V Anglii v roce 1794 začíná Dr. Buxton vyrábět rastrovaný papír, v Německu v roce 1798 pražský rodák Aloys Senefelder vynalézá litografickou techniku pro tisk map a diagramů (své výsledky shrnuje v knize *Vollstandiges Lehrbuch der Steindruckerei*, 1818). Ve Francii v roce 1843 Léon Lalanne (viz níže) začíná používat sférické souřadnice a v roce 1846 zavádí logaritmickou stupnici na obě pravoúhlé osy. Semilogaritmickou stupnici používá jako první pro své diagramy W. S. Jevons v roce 1863.

Playfairovy grafy byly patrně inspirací pro anglickou statističku Florence Nightingaleovou. Přihlásila se jako dobrovolná zdravotní sestra v době krymské války, sestavovala časové tabulky úmrtí pacientů podle příčin a jimi dokazovala nedostatečnost nemocniční hygieny v polních podmínkách. V prvním provedení byly počty úmrtí úměrné úsekům poloměrů výsečí a tedy zkreslené, poté si uvědomila svou chybu a jako první zavedla *radiální graf* – Obr. 14. Vedle podrobné zprávy pro vojenské kruhy vydala stručný souhrn svých výsledků také jako malou brožurku (*Mortality of the British Army*, 1858) s cílem ovlivnit veřejné mínění. Ať již její grafické zpracování přesvědčilo velení armády či veřejnost, která uplatnila svůj vliv, hygieně v nemocnicích začala být věnována podstatně větší pozornost, a to nejen v armádě. Po návratu do Anglie měla F. Nightingaleová značný (údajně dodnes přetrvávající) podíl na celkovém zlepšení nemocniční péče, jíž věnovala veškerou svou pozornost po zbytek života. Její radiální grafy bývají v literatuře nazývány *kohoutími hřebínky* (*coxcombs*), jedná se však o jeden z historických omylů; kohoutím hřebínkem nazvala F. Nightingalová v průvodním dopise z 25. 12. 1857 k výše zmíněné brožurce presidentovi Královské armádní komise Sidney Herbertovi[5] právě tuto brožurku, nikoliv svůj radiální graf.

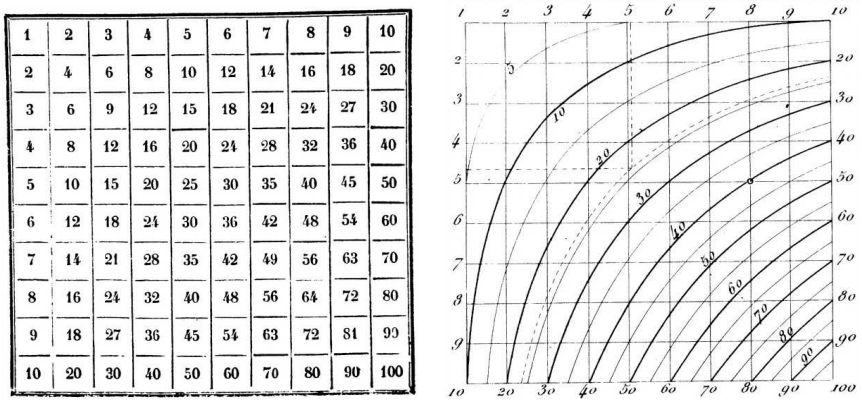
Od začátku XIX. století se střediskem vývoje grafického znázorňování dat stává Francie. Jejich technické využití a rozvoj jsou svázány s odvozem městských odpadků, který byl aktuální již v XVIII. století a vynucoval si stále rozsáhlejší stavbu silnic. Maximální efektivností této problematiky se zabývaly dva přední francouzské vzdělávací ústavy: vojenská École de Génie v Meziéres (s těžištěm v likvidaci pevnostního odpadu) a École des Ponts et Chausées v Paříži zaměřená civilně. Profesorem na první škole byl Gaspard Monge, zakladatel deskriptivní geometrie, a právě z jeho žáků a následovníků se rekrutovali významní propagátoři grafického zobrazování. Na druhé z uvedených škol zase vyučoval již zmíněný Ch. J. Minard. Záměr pokrýt celou Francii vyhovující sítí silnic hvězdicovitě vycházejících z Paříže se stává aktuální kolem roku 1842. Při jeho realizaci opět přichází ke slovu grafická kartografie, zvláště díky Ch. J. Minardovi, který se snažil prosadit decentralizovanější dopravní síť, jejíž výhodnost demonstroval čarami s tloušťkou úměrnou přepravním nárokům; tato forma grafického znázornění vyvrcholila posléze jeho Napoleonovým tažením. Výstavba dopravní sítě však byla svěřena centrální státní organizaci Corps des Ponts et Chausées řízené Victorem Legrandem; její charakter vyjadřoval hovorový název „Legrandova hvězda“ a byla spojena s obrovskými přesuny půdy díky přísným požadavkům na povolené maximální stoupání a minimální poloměry křivosti. Již v letech 1835 a 1837 byly vypracovány tabulky pro výpočet nezbytných přesunů zeminy, platily však pouze pro jeden pevný profil silničního uložení.

Grafické konverze výpočetních tabulek se ujal Léon Lalanne. Vyšel při-

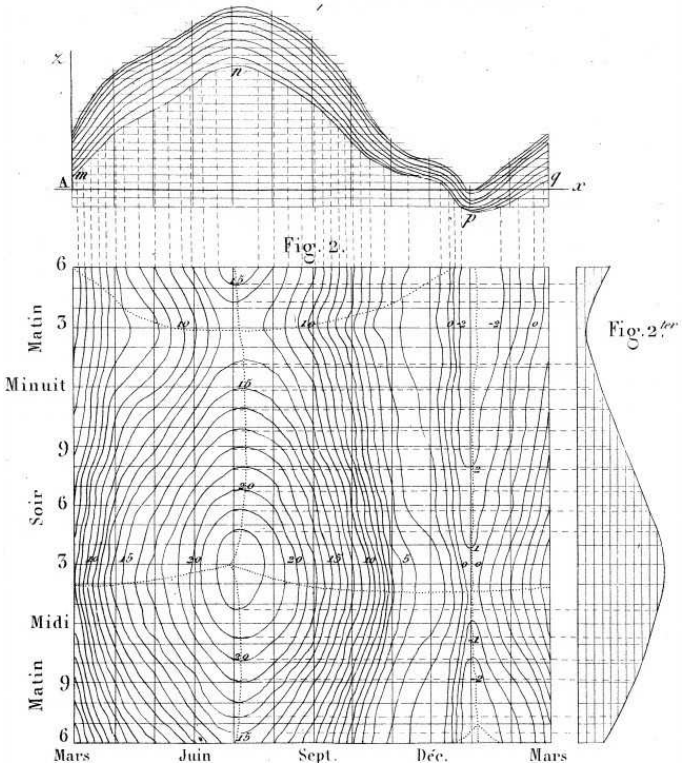
[5] „Dear Mr. Herbert, I send you one of the „coxcombs“ …“, citováno podle příspěvku H. Small: Florence Nightingale's statistical diagrams, předneseného 18. 3. 1998 na konferenci, kterou pořádalo Museum Florence Nightingaleové v Londýně.



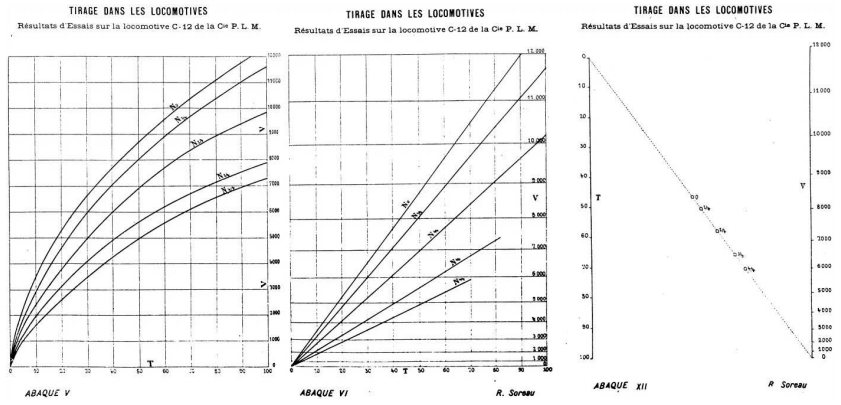
Obrázek 14: Radiální graf F. Nightingaleové (1858) znázorňuje příčiny úmrtí vojáků (počet úmrtí je úměrný ploše) v krymské válce (1854–55). Vnitřní malé světlé výseče zachycují po jednotlivých měsících úmrtí na zranění, velké světlé výseče úmrtí na nakažlivé choroby vyvolané nedostatečnou hygienou a vnitřní malé tmavé výseče libovolné jiné příčiny. Sloupcové diagramy porovnávají procentuální úmrtnost v různých věkových kategoriích (horní diagram) a podle příčin (spodní diagram) u běžných anglických mužů a u vojáků (vždy spodní sloupec v páru).



Obrázek 15: Pouchetova Pytagorejská tabulka a isočáry $xy = 5k, k = 0, 1, \ldots, 19$ (1795).



Obrázek 16: Lalanneův prostorový graf {měsíc × hodina × teplota} (1845).



Obrázek 17: Srovnání grafu v kartézských a logaritmicko-logaritmických souřadnicích s nomogramem od M. d'Ocagne (1891). T je tažná síla francouzské lokomotivy, V je váha páry spotřebované za hodinu a N je relativní doba, po niž je parní ventil otevřený.

tom z tzv. pytagorejské tabulky typu 10×10^6, kterou Louis-Ézechiel Pouchet (v souvislosti se snahami francouzské vlády přejít na decimální soustavu jednotek) v roce 1795 doplnil isočarami (hyperbolami) $xy = 5k, k = 1, 2, \ldots, 19$ – Obr. 15. Tabulka se sice obecně neprosadila, byla však používána pro inženýrské výpočty k převodu různých měr, např. při kalibraci děl. Lalanne nejdříve upozornil, že čáry $xy = konst.$ můžeme chápat jako ortogonální projekce čar konstantní výšky na 3D ploše $z = xy$ a pro demonstraci této myšlenky vytvořil projekci isoterm v 3D grafu typu {měsíc×hodina×teplota} s projekcí do roviny {měsíc × teplota} a řezem rovinou {hodina × teplota} (Mongeova škola se nedala nezapřít) – Obr. 16.

Druhou inovací bylo zavedení logaritmických souřadnic (Pouchetovy hyperboly se pak staly přímkami) a v roce 1846 již Lalanne publikuje grafickou tabulku s lineárními závislostmi půdních přenosů pro dvoukolejnou železnici. Vývoj dovršuje v roce 1884 Maurice d'Ocagne vytvořením *nomogramu*. Pravoúhlé osy nahrazuje osami rovnoběžnými a využívá principu duality z projektivní geometrie, podle nějž lze body zobrazit jako přímky a přímky jako body. Soubor přímek z Lalanneova grafu pak přechází v přímku jedinou – Obr. 17.

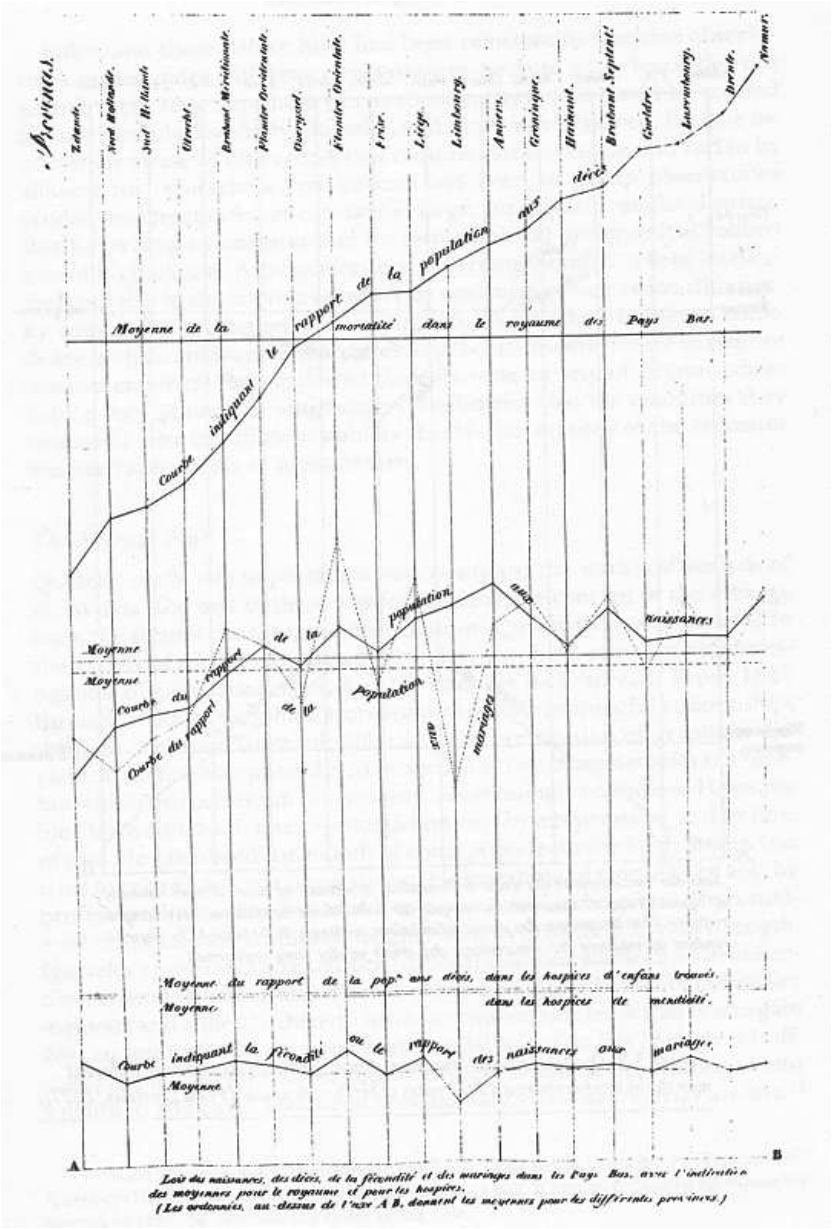
Zásluhy L. A. J. Queteleta o rozvoj statistiky v sociální oblasti jsou dostatečně známé: jeho názory jsou různě vykládány, interpretovány i kritizovány, jeho podíl na vzniku statistických společností v evropských státech i v Americe je však nesporný, stejně jako inspirativní vliv na celou řadu statistických aktivit. Z Queteletových grafických prací si všimneme aspoň jednoho okruhu studií včetně okolností, za nichž vznikly. Sčítání lidu je velmi nákladná akce, a když se v porevoluční Francii o ní začalo uvažovat, přišel P. S. Laplace s návrhem určité formy výběrového šetření. Doporučil využít přesně vedených matrik narozených dětí v celé zemi a celkový počet obyvatel N_O určit ze vztahu $N_O = r_D N_D$, kde N_D je počet všech narozených dětí za nějaké období a $r_D = n_O/n_D$ je pečlivě stanovený poměr počtu obyvatel a narozených dětí ve vybraných „reprezentativních" oblastech, rovnoměrně rozložených po celé ploše státu a s pozorností k jednotlivým skupinám obyvatel[7]. Quetelet byl nejprve (v roce 1824) nakloněn použití této metody i v Belgii a Nizozemí, avšak v roce 1829 podává návrh na kompletní sčítání. Byl totiž zřejmě ovlivněn pamětním spisem, který mu poslal baron de Keverberg v roce 1827[8] a v němž zpochybňuje možnost dostatečně vhodného výběru podoblastí pro odhad poměru r_D, protože relace mezi n_O a n_D závisí nesnadno definovatelným způsobem na množství lokálních proměnných.

Patrně inspirován de Keverbergovým spisem, provedl Quetelet v 19 oblastech Belgie, Holandska a Lucemburku vlastní výběrové odhady následují-

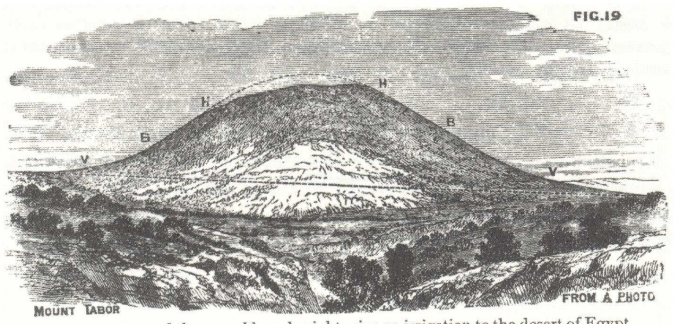
[6] Celočíselná tabulka s hodnotami $\{x_{ij} = ij\}, i, j = 1, 2, \ldots, 10$.

[7] Viz P. S. Laplace: *Essai philosophique sur les probabilités*, Paris, Courcie 1840 (6. vydání knihy z roku 1814), 100–101. Laplace svou metodu navrhl již v roce 1780, byla použita v r. 1802 a diskutuje ji K. Pearson v Biometrika 20A (l928), 165–174.

[8] Citace z de Keverbergova spisu jsou v S. M. Stiegler: *The History of Statistics*, Harvard University Press, Cambridge 1986.



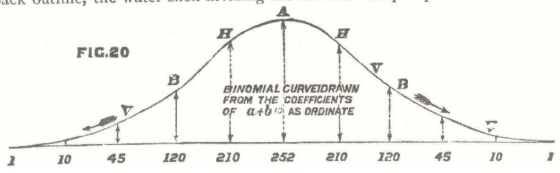
Obrázek 18: Queteletův souborný diagram (1827) poměrů r_M (horní lomená čára), r_D (střední plná čára), r_S (střední tečkovaná čára) a r_F (spodní plná čára). Odpovídající střední hodnoty jsou vyznačeny horizontálními čarami.



at present ; and they would, and might, give an irrigation to the desert of Egypt, which would be of immense agricultural value, by causing an earlier and larger overflow of the Nile. The opposite is also true, that rivers joined at the same slope traverse with greater velocity. If the rivers of the Rhine Delta were joined so as to have one mouth, the Rhine would have a deep navigable mouth, and the stream, by its velocity, would always keep 22 feet of water at the bar.

Fig. 20 is a diagram, the outline of which is a true binomial curve. This curve is shown in dotted line on Fig. 19, Mount Tabor, a type of a certain class of hills.*

Another common form is also shown in Fig. 21 ; the hills assume the fish-back outline, the water-shed dividing the hill into unequal parts.



Obrázek 19: Vrcholová eroze hory Tábor podle A. Tylora (1875).

cích veličin: počtu obyvatel n_O, počtu narozených dětí n_D, počtu uzavřených manželství n_S a počtu úmrtí n_M, z nichž pro každou oblast odhadl poměry $r_M = n_O/n_M, r_S = n_O/n_S, r_F = n_D/n_S$ a $r_D = n_O/n_D$ a oblasti srovnal za sebou tak, aby r_M bylo monotónní rostoucí. Výsledky jsou shrnuty ve známém Queteletově diagramu, který ukazuje poměrně velké rozdíly mezi hodnotami poměrů v jednotlivých oblastech a dále naznačuje, že mezi nimi je jen stěží nějaká korelace – Obr. 18. Odtud tedy vyplynula Queteletova ztráta důvěry v Laplaceův návrh výběrového sčítání.

Další Queteletova grafická práce se vztahuje k jeho koncepci „průměrného člověka“, jehož psychické i fyzické vlastnosti mají normální rozdělení (Quetelet však používal termíny *křivka možností*, *rozdělení možností*, *binomická křivka*). Přesvědčení, že každý homogenní soubor údajů musí mít normální rozdělení, považoval Quetelet za řešení de Kevenbergovy námitky o nemožnosti posoudit, zda data vytvářejí homogenní soubor či nikoliv. V řadě prací srovnával zjištěná data s normálním rozdělením, jež však nepoužíval v Gaussově integrálním tvaru, ale vycházel z binomického rozdělení $\mathrm{Bi}(999, 1/2)$, které podrobně propočítal s využitím vztahu $\mathrm{P}(X = k) = (n - k + 1)\mathrm{P}(X = k - 1)/k$, platného pro $\mathrm{Bi}(n, 1/2)$.

Obecnou popularitu normálního rozdělení dokumentuje článek A. Tylora[9] z roku 1875, v němž autor povýšil křivku normálního rozdělení na universální geologický standard (*binomická křivka* nebo-li *denudační křivka*) tvaru hor. Odchylky od ní jsou něj důkazem lokální eroze demonstrované na příkladu biblické hory Tábor – Obr. 19.

Luigi Perozzo vstoupil do historie grafického zobrazování prvním 3D grafem – Obr. 20, který nazval *stereogramem* a jenž využívá axonometrického promítání navrženého Gustavem Zeunerem v knize *Abhandlungen aus der mathematischen Statistik*, Leipzig (1969)[10]. 3D grafy byly často využívány pro znázornění vícerozměrných distribučních funkcí a hustot pravděpodobnosti.

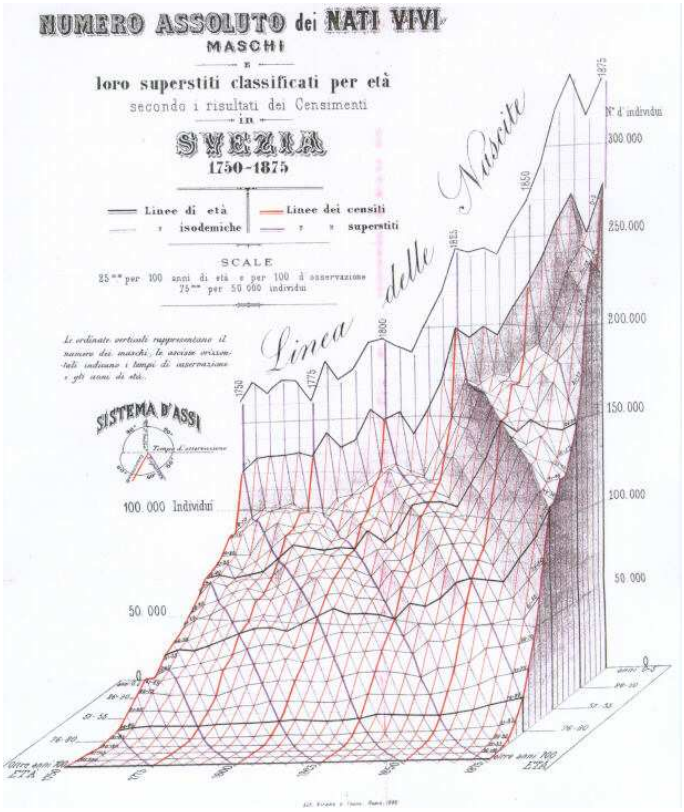
W. S. Jevons se v roce 1863 začal zabývat problémem kvantitativního popisu cenových změn vyvolaných událostmi obecného dosahu, konkrétně např. objevením australského a kalifornského zlata v roce 1849, jež mělo za následek dlouhodobý pokles ceny zlata. Ze sledovaných 118 produktů jich 84 zdražilo, ostatní zlevnily. Všechny změny Jevons zanesl do souborného semilogaritmického grafu a stanovil jejich *geometrické průměry* – Obr. 21. Tento typ výpočtu cenových změn je od té doby široce používán. Nebyl zdaleka první, avšak prosadil se ze dvou důvodů. Předně pro výrazně asymetrické rozdělení relativních cenových změn je geometrický průměr vhodnější, než do té doby používaný průměr aritmetický, jednak se ukázalo, že je vhodné sledovat velmi široký výběr produktů, což Jevonsovi předchůdci nedělali. Jevons pro svůj postup měl ovšem jen intuitivní důvody; zmiňoval např. alternativní možnost sledovat množství zboží, které lze po skokové změně zakoupit za stejnou cenu, cožby vedlo k průměru harmonickému, a svůj geometrický průměr vydával za střední cestu mezi oběma alternativami.

4 Závěr

Koncem XIX. století se začíná rozvíjet intenzivní zkoumání v oblasti lékařství a biologie v Anglii, do značné míry spojené s osobou Francise Galtona. Jeho zájem o aplikaci statistických přístupů včetně jejich grafické prezentace byl zcela mimořádný a vedl ke vzniku tzv. biometrické školy, jejímiž představiteli vedle Galtona byli Karl Pearson, Francis Weldon, Udna Yule a samozřejmě Ronald Aylmer Fisher. Grafická prezentace se v jejich pracích stala běžným prostředkem do té míry, že si dnes bez ní statistiku dovedeme jen stěží představit. Zhruba do konce XIX. století je možné vývoj grafického zobrazování alespoň v hrubých rysech sledovat v příspěvcích rozsahu srovnatelného s tímto textem. XX. století, zejména jeho druhá polovina ovlivněná rozvojem počítačové techniky a vstupem grafiky do všech medií, představuje pravý grafický

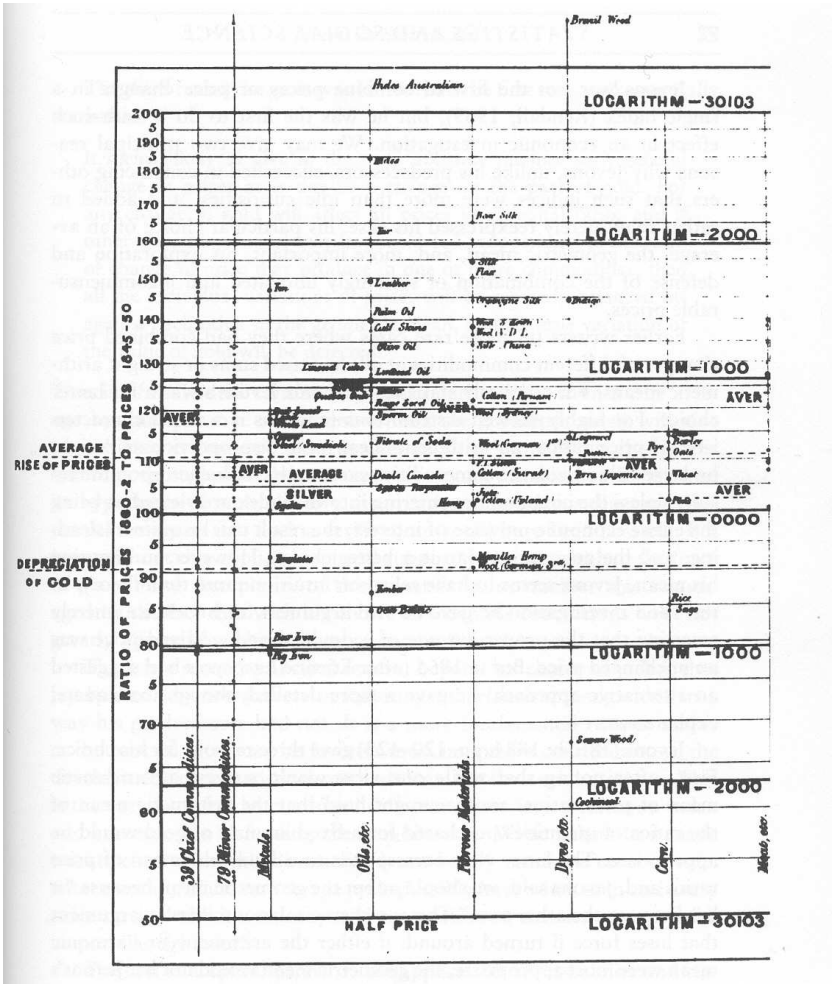
[9] A. Tylor: Action of denuding agencies. Geological Magazine (decade II) 2 (1875), 433–476 - viz S. M. Stigler: *The History of Statistics. The Measurement of Uncertainty before 1900.* Harvard University Press, Cambridge (Mass.) – London 1986.

[10] L. Perozzo: Della rappresentazione graphica di una collettività di individui nella successione del tempo. Annali di Statistica, 12 : 1–16.



Obrázek 20: Trojrozměrný graf švédské populace v letech 1750–1875 od L. Perozzo (1880).

výbuch. Jeho rozsah lze snad alespoň přibližně ocenit z odhadu E. Tufta, prezentovaného v osmdesátých letech minulého století: počet grafů vytvořených za rok se pohybuje mezi 9.10^{11} až 2.10^{12}; při počtu lidí v řádu 10^{10} připadá tedy 100 grafů na osobu za rok. Toto množství zdaleka neznamená, že grafy jsou kvalitní a že data jsou prezentována optimálním způsobem. Právě naopak; převážná většina grafů má za úkol zaujmout, obrátit pozornost k určité obchodní nebo politické problematice, „nakazit“ konzumenta názorem či záměrem svých autorů. Současná teorie grafického zobrazování dává přednost těm nejjednodušším formám snadno dešifrovatelné bodové a čárové reprezentace.



Obrázek 21: Jevonsův semilogaritmický diagram (1863) cenových změn po objevu australského a kalifornského zlata. Vyneseny jsou poměry průměrných cen v objevem zlata ovlivněných letech 1860 až 1862 k průměrným cenám ve srovnávacím období 1845 až 1850. Na levých dvou svislých přímkách jsou vyneseny všechny hlavní (39 položek) a vedlejší (89 položek) produkty a vyznačeny jim odpovídající průměrné změny, dále průměrné relativní zvýšení (cca 11 %) a jemu odpovídající relativní pokles ceny zlata (cca 9%). V jednotlivých sloupcích jsou většinou příbuzné produkty, jako železo, oceli a stříbro, různé druhy olejů, textilní látky, obilí atd.

Reference

Internetové stránky s početnými příklady grafik (většinou interaktivní):

[1] M. Friendly, D. J. Denis: *Milestones in the History of Thematic Cartography, Statistical Graphics, and Data Visualization.* `http://www.math.yorku.ca/SCS/Gallery/milestone/`[11]

[2] Cartographic images:
`http://www.henry-davis.com/MAPS/Ren/Ren1/carto.html`[11]

[3] The History of Cartography:
`http://feature.geography.wisc.edu/histcart/`[11]

[4] The History of Cartography:
`http://www-groups.dcs.st-and.ac.uk/~history/HistTopics/Cartography.html`[11]

[5] EIA Guidelines for Statistical Graphs:
`http://www.eia.doe.gov/neic/graphs/preface.htm`[11]

Ostatní odkazy:

[6] J. R. Beniger, D. L. Robyn: *Quantitative Graphics in Statistics*. The American Statistician 32 (1978), 1–11.

[7] M. Friendly: *Graphical Methods for Categorical Data*. SAS SUGI Conference, April 1992.
`http://www.math.yorku.ca/SCS/sugi/sugi17-paper.html`[11]

[8] H. G. Funkhauser: *Historical development of the graphical representation of statistical data.* Osiris 1 (1937), 269-405.

[9] R. W. Glenn: *Data Graphics*. Speech 250 (Adv. Public Speaking) Schedule, Spring 2004.
`http://web.utk.edu/~glenn/DataGraphics.html`[11]

[10] B. J. Glick: *Mortality Mapping.*
`http://zappa.nku.edu/~longa/geomed/modules/av/lab/`[11]

[11] T. L. Hankins: *Blood, Dirt, and Nomograms*. Isis 90 (1999), 50–80.

[12] G. Palsky: *Des Chiffres et des Cartes (La Carographie Quantitative au XIX^{e}) Siécle.* CTHS (Comité de travaux historiques et scientifiques), Paris 1996.

[13] E. Royston: *A Note on the History of the Graphical Presentation of Data*. Biometrika 43 (1956), 241–247.

[14] W. S. Cleveland: *A Model for Studying Display Methods of Statistical Graphics.* J. Comput. Graph. Stat. 2 (1993), 323–343.

[15] E. R. Tufte: *The Visual Display of Quantitative Information.* Graphic Press, Cheshire 1983.

[11] K datu 23. 10. 2004.

5 Dodatek

Anaximandros z Milétu (610 př. Kr. - po 546 př. Kr.), řecký filosof, podle Eratosthéna tvůrce řecké geografie, autor údajně první mapy světa vytvořené pod vlivem babylónské astronomie. Ta se sice nezachovala, bývá však rekonstruována podle Hérodotova popisu.

Abraham Cresques (1326? - 1387), významný představitel mallorské kartografické školy, která se proslavila kreslením námořních map. Nejznámější je Katalánský atlas vytvořený pro Karla V.

August Friedrich Wilhelm Crome (1753 - 1833), teolog, lektor zeměpisu a historie v Dessau, poté profesor politických věd v Gießenu od roku 1786 do smrti a diplomat. Autor řady knih, např. *Europens Produkte* (1782), *Über die Größe und Bevölkerung der europäischen Staaten* (1785), čtyřdílné *Geographisch-statistische Darstellung der Staatskräfte von den sämtlichen, zum deutschen Staatenbunde gehörigen Ländern* (1820 – 1828) a vynikající *Selbstbiographie* (1833). Jeho knihy i politické pamflety obsahují řadu tabelárně i graficky zpracovaných statistických údajů.

Pierre Charles François Dupin (1784 - 1873), francouzský matematik a ekonom, žák Mongeův, autor různých pamfletů s vědeckou tématikou. V práci *Carte de la France éclairée et de la France obscure* (1819) první použil různé barvy k zakreslení vývoje vzdělání v různých regionech.

Eratosthenés z Kýrény (275 př. Kr. - 195 př. Kr.), všestranný řecký vědec, etik a básník, především geograf. Jeho odhad obvodu Země (44 730 km) využívá měření délek stínů tyče ve dvou městech ležících na stejném poledníku (Samara a Aswan). Výpočet je pak založen na předpokladu kulatosti Země a výsledek je velmi blízký skutečnosti.

Joseph Fletcher (? - ?), autor článků s řadou kolorovaných map s údaji o zemědělství, průmyslu aj. („Moral and Educational Statistics of England and Wales“, Journ. London Stat. Soc. X (1847), 193-233 a XII (1849), 151-335).

Francis Galton (1822 - 1911), anglický genetik a eugenik, původně meteorolog (jako první popsal anticyklonu a zavedl mapy počasí založené na barometrických datech). Do genetiky zavedl biometrický přístup, do statistiky pojmy regrese a korelační koeficient. Do kriminologie přispěl zavedením otisků prstů jako identifikačního znaku.

Edmond Halley (1656 - 1742), přední anglický fyzik a astronom (objevitel komety po něm nazvané) ve vědeckých kruzích vzácného charakteru, autor významné práce z demografie a pojišťovnictví „An Estimate of the Degrees of the Mortality of Mankind“ (1693). Výrazně se zasloužil o publikování Newtonových *Principia Mathematica*. Posledních 22 let svého života byl královským astronomem se sídlem na hvězdárně v Greenwichi, kde také zemřel.

Wiliam Stanley Jevons (1835 - 1882), anglický ekonom a logik, autor knihy *Principles of Science* (1874), v níž hájí názor, že deduktivní poznatky a zákony jsou jen pravděpodobné, protože není možné rozebrat všechny možné příčiny a alternativy. V ekonomických pracích (*Theory of Political Economy*, 1871) vycházel ze statistické analýzy reálných dat, pro hodnocení cenových hladin zavedl dosud používaný index založený na *geometrických* průměrech cen širokého spektra komodit.

Léon Lalanne (1811 - 1892), francouzský stavební inženýr, vynálezce řady grafických postupů vypracovaných v souvislosti s výstavbou francouzské dopravní sítě, generální inspektor mostů a silnic, ředitel l'Ecole des Ponts et des Chausées atd.

Johann Heinrich Lambert (1728 - 1777), samouk, alsaský přírodovědec a filosof, známý pracemi z optiky (zvláště fotometrie), matematiky (teorie kuželoseček, hyperbolických a trigonometrických funkcí komplexní proměnné, důkaz iracionality π a e^x pro racionální $x \neq 0$, neukleidovská geometrie) a astronomie (teorie Vesmíru tvořeného galaxiemi a hvězdami, výpočet dráhy komet). Zabýval se též teorií pravděpodobnosti a navrhl grafickou formu metody maximální věrohodnosti. Člen Pruské akademie věd.

Michael Florent van Langren (1600-1675), holandský matematik, astronom a kartograf, pracoval ve službách španělského krále Filipa IV. Vyvinul metodu pro přesné určování zeměpisné délky pro potřeby lodní navigace, která se opírala o pozorování Měsíce a přivedla jej k autorství první mapy Měsíce.

Gerardus Mercator (1512 - 1594), nizozemský geograf, matematik a kartograf, vlastním jménem Gerhard Kremer. Autor Mercatorovy projekce zemského povrchu vhodné pro námořní plavbu. Mercatorovy mapy vydává i po jeho smrti početná rodina.

Charles Joseph Minard (1781-1870), význačný francouzský inženýr, specialista v oboru stavby mostů, silnic a kanálů, ředitel a profesor na l'Ecole des Ponts et des Chausées, po odchodu do důchodu (1839) se věnoval tématické kartografii.

Florence Nightingale (1820 - 1910), anglická statistička, dobrovolná zdravotní sestra v krymské válce, sestavovala časové tabulky úmrtí pacientů podle příčin a jimi prokázala nedostatečnost nemocniční hygieny. Grafickým znázorněním (radiální diagram) přesvědčila vojenské kruhy o nezbytnosti nápravy. Po návratu do Anglie měla značný podíl na celkovém zlepšení nemocniční péče.

Philbert Maurice d'Ocagne (1862 - ?), francouzský matematik, vynálezce nomogramu, profesor na l'Ecole des ponts et chaussées, systematicky se podílel na bibliografii matematických prací, též autor literárních esejí. Hlavní dílo je *Nomographie; les calcul usuel effectué aux moyen des abaques* (1891).

Charles Saunders Peirce (1839 - 1914), americký logik, matematik a filosof, původním povoláním a vzděláním chemik a geodet, mimořádně plodný v mnoha oborech. Obecně považován za nejpozoruhodnější intelektuální osobnost Spojených států v 19. století.

William Playfair (1759 - 1823), jako osoba téměř zapomenutý skotský vynálezce, pamfletista, novinář a vydavatel (z jeho pera pochází také popis pádu Bastilly, jehož se osobně zúčastnil), který kromě řady různých vynálezů první používal histogramy, kruhové a lineární grafy pro znázornění statistických dat (kniha Atlas se 44 různými grafy byla publikována poprve v r. 1786).

Louis-Ezechiel Pouchet (? - ?), výrobce bavlněných tkanin v Rouenu, autor několika publikací s problematikou jednoduchého grafického převodu jednotek, např. *Echelles graphiques des nouveaux poids, mesures et monnaies de France, comparées avec celles des pays les plus commerçantes de l'Europe* (1795).

Joseph Priestley (1733 - 1804), anglický fyzik (první objevil Coulombův zákon), chemik (vynálezce sodovky, objevitel fotosyntézy a dýchání rostlin), sociální filosof, teolog, podporovatel francouzské a americké revoluce, přítel Thomase Paina, pro svou podporu francouzské a americké revoluce donucen emigrovat do Ameriky (1794), zakládá tam unitářskou církev, přítel a spolupracovník T. Jeffersona.

Klaudios Ptolemaios (100 - 178?), autor souborného díla o matematice známého z arabského překladu z IX. stol. pod jménem Almagest. Základní kartografické dílo (do roku 1730 vyšlo na padesát různých vydání) je osmisvazkový $\Gamma\varepsilon o\gamma\rho\alpha\phi\iota\kappa\varepsilon\ \Sigma\upsilon\nu\tau\alpha\xi\iota\varsigma$ [Zeměpisný úvod] se souřadnicemi 8 000 obydlených míst a množstvím map, které se však nedochovaly a v následujících stoletích byly mnohokrát rekonstruovány. Další jeho díla jsou věnována hudbě a optice.

Lambert Adolphe Jacques Quetelet (1796-1874), belgický geometr, astronom (Queteletův kráter na Měsíci), meteorolog a geofyzik - především významný popularizátor ovlivněný pracemi Laplace, Poissona, Bernoulli, předseda belgické Comission centrale de statistique, organizátor statistických kongresů. Autor práce *Sur l'homme et le developpement de ses facultés, essai d'une physique sociale* (1835).

Valentine Seaman (1770 - 1817), americký lékař, roku 1795 při epidemii žluté horečky ve východním Manhattanu vypracoval mapu lokálního výskytu onemocnění a nákazu připsal hnijícímu odpadu v ulicích města (sice nesprávně, neboť ji šířili komáři druhu *Aedes Aegypti*, kteří se ve vlhkých nečistotách vyskytovali, nicméně úklidem nečistých zaplavovaných míst choroba pominula). Měl podíl na zavedení hromadném očkování proti černým neštovicím.

Aloys Senefelder (1771 - 1834), pražský rodák, herec a dramatik, kvůli tištění svých her experimentoval s tiskem a objevil litografii. Díky jejímu velkému úspěchu dosáhl všeobecného uznání.

John Snow (1813 - 1858), významný anglický lékař a vědec, autor souboru map Londýna (ještě stále překreslovaných a publikovaných) do nichž zakresloval lokální výskyt chorob. President Medical Society of London od roku 1855, spolupracovník osobních lékařů královny Viktorie (aplikace chloroformu při královských porodech), podle časopisu Hospital Doctor z roku 2003 nejvýznamnější lékař všech dob (Hypokratos až druhý).

Dugald Stewart (1753 - 1828), skotský filosof, profesor matematiky a filosofie na Universitě v Edinburgu, dlouhá léta vedoucí katedry etiky. Autor rozsáhlého literárního díla zahrnujícího i politickou filosofii a tehdy ještě málo rozšířenou politickou ekonomii. Ve své době velmi vážená osobnost, přitahující do Edinburgu studenty i z Evropy a Ameriky.

James Joseph Sylvester (1814 - 1897), anglický matematik a právník, autor řady prací z algebry, teorie čísel, řešení diofantických rovnic, teorie matic a geometrie. Autor knihy *Treatise on Elliptic Function* (1876), jako profesor na baltimorské universitě (1976 - 1883) měl velký podíl na rozvoji americké matematiky, psal též básně a byl autorem knihy *Laws of Verse* (1870).

Walter Frank Rafael Weldon (1860 - 1906), anglický zoolog, profesor na londýnské University College, jako první prováděl u živočichů, zvláště vodních, analogická statistická měření jako Quetelet a Galton u lidí. Spolu s Galtonem a Pearsonem založil v roce 1901 časopis Biometrika.

Christopher Wren (1632 - 1723), vynikající anglický architekt, stavitel londýnské katedrály sv. Pavla a řady dalších veřejných budov, autor nerealizovaných plánů na obnovu Londýna po Velkém požáru v roce 1666.

Gustav Zeuner (1828–1907), německý fyzik a pedagog, profesor technické mechaniky a strojírenství v Zürichu a Freibergu, poté dlouholetý ředitel drážďanské polytechniky, věnoval se zejména technické termodynamice a dopravnímu inženýrství.

Poděkování: Poděkování grantu GACR 201/03/0946 a výzkumnému záměru MSM 113200008.

Adresa: I. Saxl, L. Ilucová, Matematický ústav AV ČR, Žitná 25, 115 67 Praha 1

E-mail: saxl@math.cas.cz

A NOTE ON RANK-BASED TESTING FOR CONDITIONAL HETEROSKEDASTICITY

Miroslav Šiman

Keywords: Conditional heteroskedasticity, GARCH, test of independence, rank test, portmanteau test, autocorrelation, transformation.

Abstract: In this note, we will introduce a rank transformation that can lead to powerful rank-based tests for conditional heteroskedasticity. Its usefulness will be illustrated in a small simulation study dealing only with GARCH(1,1) alternatives.

1 Theoretical introduction

Throughout this article, we will assume that $\{z_t\} \sim IID(0,1)$ is a sequence of independent, identically and continuously distributed (i.i.d.) random variables with zero mean and unit variance.

Now let us consider a weakly stationary zero mean sequence of random variables $Y_1, \ldots, Y_T$ that we want to check for the presence of conditional heteroskedasticity. In other words, we want to test the null hypothesis that assumes $\{Y_t\}$ to be an i.i.d. sequence with finite variance,

$$H_0 : Y_t = z_t\sigma,\ \sigma > 0 \ \text{ is an unknown positive constant, } \ (t = 1, 2, \ldots, T),$$

against a weakly stationary alternative

$$H_1 : Y_t = z_t\sigma_t, \ \ (t = 1, 2, \ldots, T),$$

where σ_t's are driven by a more or less specified functional dependence on lagged random variables forming the time series of our interest. For example, GARCH, GARCH(p,q) or GARCH(1,1) alternatives widely used in practice are also of this type.

GARCH processes play a key role in modelling volatile time series and good results can be often achieved even with a simple GARCH(1,1) model (see [10]). Almost every GARCH process, say $\{\varepsilon_t\}$, is defined as follows:

$$\varepsilon_t = z_t\sigma_t$$

where

$$\sigma_t = f(\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots, \sigma_{t-1}, \sigma_{t-2}, \ldots)$$

and $\{z_t\} \sim IID(0,1)$. Each individual GARCH model only differs from the others by its specification of the conditional standard deviation σ_t and by the distribution of its innovations z_t's. For example, GARCH(p,q) models have their conditional variance given by the formula

$$\sigma_t^2 = c + a_1\varepsilon_{t-1}^2 + a_2\varepsilon_{t-2}^2 + \cdots + a_p\varepsilon_{t-p}^2 + b_1\sigma_{t-1}^2 + b_2\sigma_{t-2}^2 + \cdots + b_q\sigma_{t-q}^2,$$
$$c > 0,\ a_1, a_2, \ldots, a_{p-1} \geq 0, a_p > 0, b_1, b_2, \ldots, b_{q-1} \geq 0, b_q > 0.$$

These models were proposed by [6] and their statistical properties were recently summarized in [5]. Innovations of these processes are usually expected to have the standard normal distribution or more frequently, the standard Student distribution with three or slightly more degrees of freedom because often examined economic random variables are known usually to have finite variances but yet infinite fourth moments (see for instance [7] and references therein).

Testing for conditional heteroskedasticity is often performed by applying a test of independence to (possibly squared) absolute values of given observations $Y_1, \dots, Y_T$ because such data often exhibits serial autocorrelation if the null hypothesis does not hold. In case of any distributional uncertainty or any data distortion (caused e.g. by outliers), any test of independence based on the ranks $R_1, \dots, R_T$ of the absolute values $|Y_1|, \dots, |Y_T|$ can be used as a convenient alternative (see [9] for an example). These rank tests are very robust to outliers and heavy tails but ordinarily suffer from relatively low power in comparison with their analogues based on original observations. This drawback of theirs will disappear if we use suitably transformed ranks instead, as we are going to demonstrate now.

For such purpose, let us define the transformed ranks

$$P_i = -\log\left(1 - \frac{R_i}{T+1}\right) = F_{\exp}^{-1}\left(\frac{R_i}{T+1}\right), \quad i = 1, \dots, T,$$

where $F_{\exp}^{-1}$ stands for the quantile function of the exponential distribution with unit mean and variance. Under the alternatives of our interest, these transformed ranks P_i's usually exhibit much stronger autocorrelation pattern than mere ranks R_i's and that is why they may lead to the tests with substantially higher power. All this will be documented in our simulation study.

Abadir and Talmain [1, Example 1] showed an important case when a simple logarithmic transformation results in much more autocorrelated data. So it is quite natural to consider this transformation (or its modifications) even in our case. Unfortunately, the exact distribution of data ranks remains unknown under many volatile alternatives and so we cannot make any theoretical justification for just our choice of the transformation. We can only recommend it as a rule of thumb.

The validity of our null hypothesis (H_0) guarantees that $\{Y_i\}_{i=1}^T$, $\{R_i\}_{i=1}^T$ and $\{P_i\}_{i=1}^T$ are formed by exchangeable variables; thus a relevant theory (such as that developed in [8], [9] and related to sample autocorrelations) may be directly applied.

Let us do so. Let $\bar{Y}, \bar{R}$ and $\bar{P}$ be the sample means and let $\widehat{r}_1(k), \widehat{r}_2(k)$ and $\widehat{r}_3(k)$ $(k = 1, 2, \dots)$ be the sample autocorrelations of the sequences $\{Y_i\}_{i=1}^T$, $\{R_i\}_{i=1}^T$ and $\{P_i\}_{i=1}^T$, respectively. We employ the standard definitions, i.e.

$$\bar{Y} = \frac{1}{T}\sum_{i=1}^{T} Y_i \quad \text{and} \quad \widehat{r}_1(k) = \frac{\sum_{i=1}^{T-k}(Y_i - \bar{Y})(Y_{i+k} - \bar{Y})}{\sum_{i=1}^{T}(Y_i - \bar{Y})^2}, \quad k = 1, 2, \dots,$$

and so on. According to [8], [9], means and variances of these sample autocorrelations can be approximated in the following way ($k = 1, 2, \ldots,\ k < \frac{T}{2}$, $T > 3$):

$$\begin{aligned}
E\widehat{r}_1(k) &\sim 0, \\
E\widehat{r}_i(k) &= -\frac{T-k}{T(T-1)}, \quad i = 2, 3, \\
\operatorname{var}\widehat{r}_i(k) &= E\left(\widehat{r}_i(k)\right)^2 - \left(E\widehat{r}_i(k)\right)^2, \quad i = 1, 2, 3, \\
E\left(\widehat{r}_1(k)\right)^2 &\sim \frac{T-k}{T(T+2)}, \\
E\left(\widehat{r}_i(k)\right)^2 &= \frac{AC_i + B}{T(T-1)(T-2)(T-3)}, \quad i = 2, 3, \\
A &= -T^3 + (k+3)T^2 - k(T+6k), \\
B &= T^2(T-k-4) + 3(T-k) + 3k(T+k), \\
C_2 &= \frac{\sum_{i=1}^{T}(R_i - \bar{R})^4}{(\sum_{i=1}^{T}(R_i - \bar{R})^2)^2} = \frac{3}{5}\frac{(3T^2-7)}{T(T^2-1)}, \\
C_3 &= \frac{\sum_{i=1}^{T}(P_i - \bar{P})^4}{(\sum_{i=1}^{T}(P_i - \bar{P})^2)^2}.
\end{aligned}$$

These formulas hold exactly if and only if the equality sign "=" is used.

These approximations also figure in the three portmanteau statistics

$$S_i = \sum_{k=1}^{m} \frac{(\widehat{r}_i(k) - E\widehat{r}_i(k))^2}{\operatorname{var}(\widehat{r}_i(k))}, \quad i = 1, 2, 3,$$

that we will focus on in our Monte Carlo experiments. (In fact, any other test of independence based on sample autocorrelations could be taken into account alternatively in a similar way). The test corresponding to S_1 was mentioned as early as in 1970 by [4] and nowadays it belongs to the most widely employed tests of independence at all. The test statistic S_2 was proposed and examined in [9] for the first time. The third statistic S_3 exploits the newly introduced transformed ranks P_i's and will be investigated and compared with the other two in the next section.

It clearly results from [8], [9] that under the null of H_0 every statistic S_1, S_2 and S_3 has the asymptotic χ^2 distribution with m degrees of freedom.

Now we know enough to proceed with the simulation study.

2 Simulation study

We always generated $N = 10\,000$ data samples and considered a representative significance level of $\alpha = 0.10$. The test statistics S_1, S_2 and S_3 were then

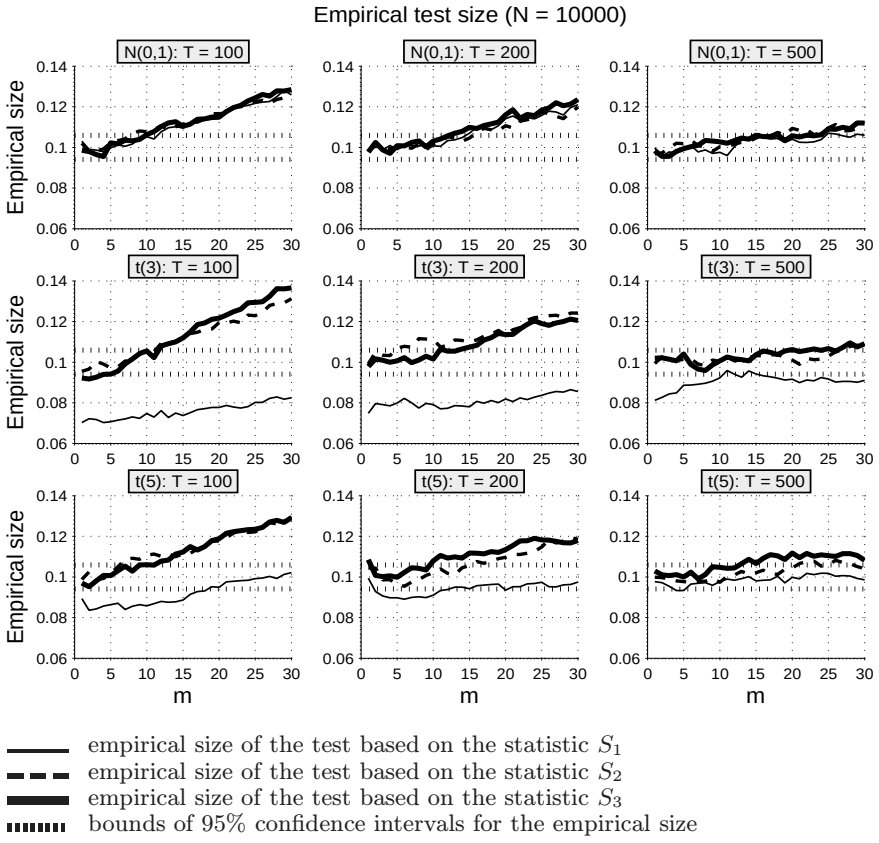


Figure 1: Empirical test size comparison taking into consideration both time series length and distribution of innovations z_t's ($N = 10\,000$, $\alpha = 0.10$).

applied to them with the threshold parameter m successively changing from 1 to 30. This parameter should not be set unnecessarily too high (see e.g. [2] or [3]) and good results can be often achieved even with $m \leq 10$.

When testing the size, we used N random samples with $T = 100, 200$ and 500 observations from the standard normal distribution and the standard Student distribution with three and five degrees of freedom. Our results are illustrated in Figure 1 and summarized in Table 1. This table contains observed empirical sizes (i.e. observed rejection frequencies of the null). Underlined figures differ from their expected values of α on the 5% significance level.

When testing the power, we concentrated on various GARCH(1,1) alternatives with innovations drawn from the same distributions that were considered in the test size study. All these alternatives have finite variances although the finiteness of their higher order moments depends on the choice of their parameters a_1, b_1 (see [5]). The remaining parameter c was fixed to

		actually observed significance level (empirical test size)											
		$T = 100$				$T = 200$				$T = 500$			
$m =$		5	10	15	20	5	10	15	20	5	10	15	20
S_1	N(0,1)	0.100	0.105	0.110	0.114	0.099	0.101	0.107	0.114	0.099	0.098	0.107	0.104
	t(3)	0.071	0.075	0.074	0.078	0.080	0.079	0.078	0.082	0.089	0.092	0.094	0.092
	t(5)	0.086	0.086	0.089	0.095	0.090	0.091	0.094	0.094	0.093	0.098	0.098	0.100
S_2	N(0,1)	0.102	0.108	0.111	0.117	0.100	0.102	0.105	0.111	0.102	0.100	0.105	0.109
	t(3)	0.097	0.106	0.112	0.120	0.107	0.111	0.110	0.113	0.101	0.100	0.103	0.101
	t(5)	0.102	0.111	0.111	0.118	0.096	0.102	0.106	0.110	0.098	0.097	0.102	0.104
S_3	N(0,1)	0.102	0.106	0.110	0.115	0.101	0.103	0.110	0.116	0.099	0.103	0.106	0.106
	t(3)	0.094	0.106	0.112	0.122	0.101	0.102	0.107	0.114	0.104	0.101	0.104	0.106
	t(5)	0.101	0.106	0.112	0.119	0.100	0.108	0.112	0.113	0.100	0.105	0.107	0.112

N ... number of replications α ... intended (asymptotic) significance level
T ... time series length m ... threshold parameter

Table 1: Empirical test size comparison taking into consideration both time series length and distribution of innovations z_t's ($N = 10\,000$, $\alpha = 0.10$).

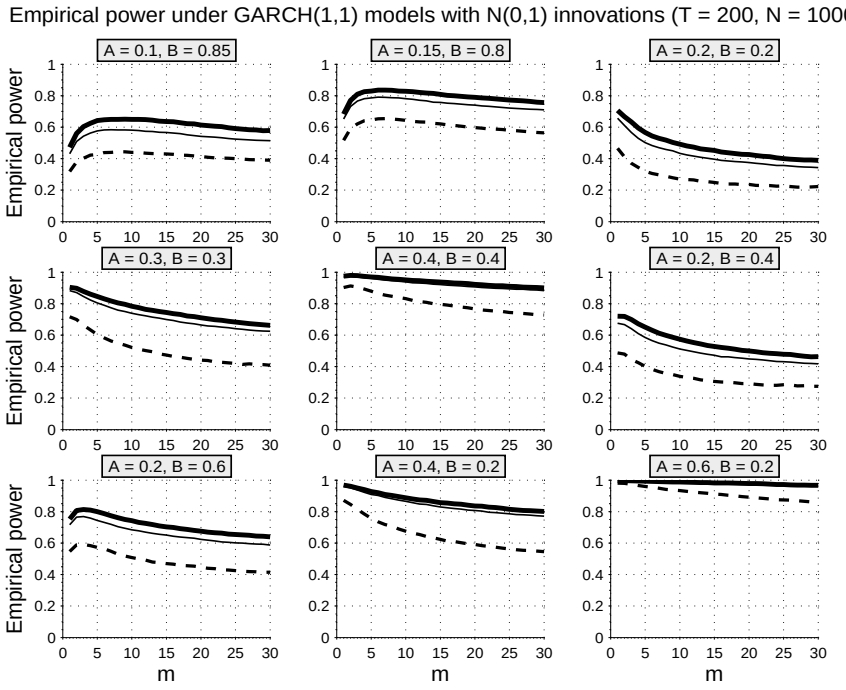
		empirical test power											
$a_1, b_1 =$		0.1, 0.85		0.2, 0.2		0.3, 0.3		0.4, 0.4		0.2, 0.6		0.4, 0.2	
$m =$		5	10	5	10	5	10	5	10	5	10	5	10
N(0,1)	S_1	0.576	0.581	0.501	0.434	0.805	0.739	0.960	0.939	0.744	0.685	0.909	0.866
	S_2	0.427	0.440	0.319	0.269	0.604	0.522	0.879	0.831	0.572	0.508	0.758	0.675
	S_3	0.642	0.649	0.566	0.492	0.846	0.784	0.970	0.951	0.799	0.742	0.927	0.888
t(3)	S_1	0.528	0.529	0.421	0.350	0.631	0.549	0.819	0.760	0.614	0.556	0.721	0.646
	S_2	0.418	0.422	0.231	0.199	0.398	0.336	0.651	0.572	0.425	0.376	0.498	0.420
	S_3	0.574	0.575	0.443	0.373	0.666	0.593	0.851	0.800	0.650	0.595	0.760	0.694
t(5)	S_1	0.608	0.607	0.485	0.413	0.750	0.677	0.927	0.891	0.716	0.657	0.852	0.797
	S_2	0.464	0.470	0.281	0.236	0.523	0.447	0.802	0.734	0.523	0.461	0.660	0.576
	S_3	0.642	0.643	0.510	0.434	0.774	0.705	0.938	0.907	0.749	0.690	0.871	0.819

N ... number of replications α ... intended (asymptotic) significance level
T ... time series length a_1, b_1 ... GARCH(1,1) parameters
m ... threshold parameter

Table 2: Empirical test power comparison under GARCH(1,1) alternatives with variously distributed innovations ($N = 10\,000$, $T = 200$, $\alpha = 0.10$).

1 throughout the whole study. Weakly stationary GARCH(1,1) models with persistent volatility, i.e. with parameters satisfying the conditions $a_1 + b_1 < 1$, $a_1 \sim 0.1$, $b_1 \sim 0.9$, stand in the centre of our attention because such alternatives are favoured in practice.

Selected results of our power test investigation are shown in Figure 2 and outlined in Table 2.



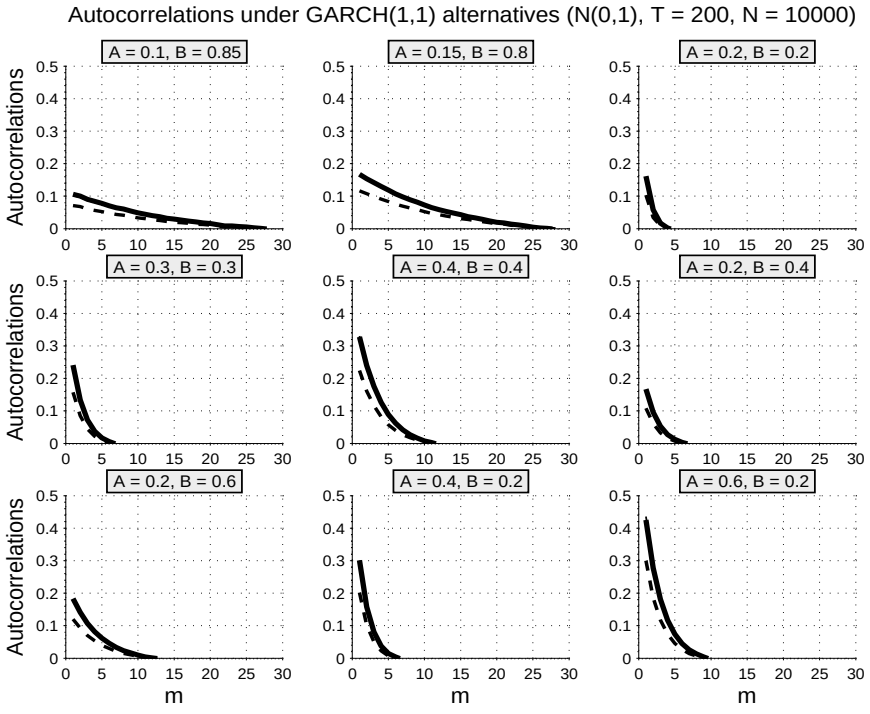
A, B GARCH(1,1) parameters a_1, b_1
empirical power of the test based on the statistic S_1
empirical power of the test based on the statistic S_2
empirical power of the test based on the statistic S_3

Figure 2: Empirical power test comparison under GARCH(1,1) alternatives with N(0,1) innovations ($N = 10\,000$, $T = 200$, $\alpha = 0.10$).

Sample autocorrelations $\widehat{r}_1(k), \widehat{r}_2(k)$ and $\widehat{r}_3(k)$ $(k = 1, \ldots, 30)$, coming from some considered alternatives are presented in Figure 3.

When only test size is taken into account, we can conclude that the test based on S_3 is comparable to that based on S_2 and better than the test corresponding to S_1. However, roughly speaking, the test based on S_3 outperforms the other two as for their power, uniformly in m. The difference between the powers of both rank-based tests may be even higher than 25 percentage points. The statistic S_3 leads to a test that is in all ways better than its non-rank analogue based on S_1. This is the main reason why it should not be ignored by practitioners any more.

Due to the limitation of the length of this article, it is not possible to include all outcomes of our simulations. But all of them seem to support the primacy of S_3 over S_1 and S_2.



N number of replications
T time series length
α intended (asymptotic) significance level
m threshold parameter
A, B GARCH(1,1) parameters a_1, b_1
——— sample autocorrelations $\widehat{r}_1(k),\ k = 1, \ldots, 30$, of $|Y_1|, \ldots, |Y_T|$
– – – sample autocorrelations $\widehat{r}_2(k),\ k = 1, \ldots, 30$, of $R_1, \ldots, R_T$
▬▬▬ sample autocorrelations $\widehat{r}_3(k),\ k = 1, \ldots, 30$, of $P_1, \ldots, P_T$
(lines ——— and ▬▬▬ visually coincide)

Figure 3: Sample autocorrelations computed under GARCH(1,1) alternatives with N(0,1) innovations ($N = 10\,000$, $T = 200$, $\alpha = 0.10$).

3 Concluding remarks

We have dealt with testing for conditional heteroskedasticity by using both the absolute values of given observations and their ranks. Under some common volatile alternatives we have shown that suitably transformed ranks P_i's are more autocorrelated than mere ranks R_i's and therefore they can lead to robust and powerful tests.

We hope that someone will find the transformed ranks P_i's or the test statistic S_3 useful and that our contribution will help the others to improve their testing (not only) for conditional heteroskedasticity.

References

[1] Abadir K.M., Talmain G. (2004). *Autocovariance functions of series and of their transforms.* Journal of Econometrics. Forthcoming.

[2] Battaglia F. (1990). *Approximate power of portmanteau tests for time series.* Statistics & Probability Letters **9**, 337 – 341.

[3] Burns P. (2002). *Robustness of the Ljung-box test and its rank equivalent.* Working Paper.

[4] Box G.E.P., Pierce D.A. (1970). *Distribution of residual autocorrelations in autoregressive-integrated moving average time series models.* Journal of the American Statistical Association **65**, 1509 – 1526.

[5] Berkes I., Horváth L., Kokoszka P. (2004). *Probabilistic and statistical properties of GARCH processes.* Fields Institute Communications. Forthcoming.

[6] Bollerslev T. (1986). *Generalized autoregressive conditional heteroskedasticity.* Journal of Econometrics **31**, 307 – 327.

[7] de Lima P.J.F. (1997). *On the robustness of nonlinearity tests to moment condition failure.* Journal of Econometrics **76**, 251 – 280.

[8] Dufour J.-M., Roy R. (1985). *Some robust exact results on sample autocorrelations and tests of randomness.* Journal of Econometrics **29**, 257 – 273. *Corrigendum.* Journal of Econometrics **41**, 279 – 281.

[9] Dufour J.-M., Roy R. (1986). *Generalized portmanteau statistics and tests of randomness.* Communications in Statistics, Theory and Methods **15**, 2953 – 2972.

[10] Hansen P.R., Lunde A. (2004). *A forecast comparison of volatility models: does anything beat a GARCH(1,1)?* Journal of Applied Econometrics. Forthcoming.

Acknowledgement: This contribution was supported by the MŠMT ČR grant MSM 113200008 and by the GA AV ČR grant 201/03/1027.

Address: M. Šiman, KPMS MFF UK, Sokolovská 83, 186 75 Praha 8

E-mail: siman@karlin.mff.cuni.cz

VYUŽITÍ POJMU HILBERTOVY BÁZE PRO OVĚŘOVÁNÍ HYPOTÉZY O SHODNOSTI STRUKTURÁLNÍCH A KOMBINATORICKÝCH IMSETŮ

Petr Šimeček, Milan Studený

Klíčová slova: Struktury podmíněné nezávislosti, celočíselná Hilbertova báze.

Abstrakt: Klíčovým problémem v metodě popisu struktur podmíněné nezávislosti (mezi N náhodnými veličinami) pomocí tzv. imsetů je otevřená otázka shodnosti dvou množin celočíselných vektorů, tzv. strukturálních a kombinatorických imsetů. Tato otázka svoji povahou spadá do oblasti celočíselného programování a souvisí úzce s úlohou nalezení tzv. minimální celočíselné Hilbertovy báze pro jistý racionální konvexní kužel. Tématem tohoto příspěvku jsou počítačové experimenty, jejichž cílem je potvrdit či vyvrátit hypotézu o shodnosti těchto dvou množin vektorů. S pomocí počítače se podařilo hypotézu ověřit pro $N \leq 4$, navíc byly dosaženy částečné výsledky pro $N = 5$ a nastíněny další možné směry postupu.

Tento příspěvek se věnuje klíčovému problému z oblasti popisu struktur podmíněné nezávislosti (mezi N náhodnými veličinami) pomocí tzv. imsetů a to hypotéze o shodnosti množin strukturálních a kombinatorických imsetů. Základní definice a většinu značení nalezne čtenář toužící po hlubším vhledu do problematiky v [1] a [2]. Zde se také nacházejí důkazy tvrzení, jež se nám zdály příliš zřejmé či naopak příliš obtížné na to, abychom je prezentovali v tomto příspěvku.

1 Základní pojmy

Zde si připomeňme alespoň základní pojmy. Nechť N je přirozené číslo (odpovídající počtu náhodných veličin).

Definice. *Imsetem rozumíme zobrazení potenční množiny $\mathcal{P}(\{1,2,\ldots,N\})$ do množiny celých čísel $\mathbb{Z}$. Hodnotu imsetu v pro $A \subseteq \{1,2,\ldots,N\}$ značíme $v(A)$.*

Imset lze chápat také jako celočíselný vektor v $\mathbb{R}^{2^N}$, jehož složky jsou indexovány podmnožinami $\{1,2,\ldots,N\}$.

Definice. *Elementárním imsetem odpovídajícím nezávislosti (nezávislostnímu vztahu) mezi náhodnými veličinami X_i a X_j dáno $\{X_k; k \in C\}$, kde $\{i\}$, $\{j\}$ a C jsou po dvou disjunktní podmnožiny $\{1,2,\ldots,N\}$, budeme rozumět imset $u_{\langle i,j|C\rangle}$ takový, že $u_{\langle i,j|C\rangle}(\{i,j\} \cup C) = u_{\langle i,j|C\rangle}(C) = 1$, $u_{\langle i,j|C\rangle}(\{i\} \cup C) = u_{\langle i,j|C\rangle}(\{j\} \cup C) = -1$ a zbylým prvkům potenční množiny přiřadí $u_{\langle i,j|C\rangle}$ nulu.*

Množinu všech elementárních imsetů budeme značit $\mathcal{E}_N$.

Povšimněme si, že $|\mathcal{E}_N| = \binom{N}{2} \cdot 2^{N-2}$.

Například pro $N = 3$ množinu všech elementárních imsetů $\mathcal{E}_3$ znázorňuje následující tabulka:

	$\emptyset$	**{1}**	**{2}**	**{3}**	**{1,2}**	**{1,3}**	**{2,3}**	**{1,2,3}**
$u_{\langle 1,2\mid\emptyset\rangle}$	1	−1	−1	0	1	0	0	0
$u_{\langle 1,3\mid\emptyset\rangle}$	1	−1	0	−1	0	1	0	0
$u_{\langle 2,3\mid\emptyset\rangle}$	1	0	−1	−1	0	0	1	0
$u_{\langle 2,3\mid\{1\}\rangle}$	0	1	0	0	−1	−1	0	1
$u_{\langle 1,3\mid\{2\}\rangle}$	0	0	1	0	−1	0	−1	1
$u_{\langle 1,2\mid\{3\}\rangle}$	0	0	0	1	0	−1	−1	1

Definice. *Množinou kombinatorických imsetů $\mathcal{C}_N$ budeme rozumět všechny možné součty konečně mnoha elementárních imsetů, neboli*

$$\mathcal{C}_N = \left\{ \sum_{i=1}^{k} \alpha_i u_i;\ \alpha_i \in \mathbb{N}, u_i \in \mathcal{E}_N, k \in \mathbb{N}_0 \right\}.$$

Rozklad kombinatorického imsetu na součet elementárních imsetů není jednoznačný, například platí

$$u_{\langle 1,2\mid\{3\}\rangle} + u_{\langle 1,3\mid\emptyset\rangle} = u_{\langle 1,3\mid\{2\}\rangle} + u_{\langle 1,2\mid\emptyset\rangle}.$$

Jednoznačný je však stupeň kombinatorického imsetu $v = \sum_{i=1}^{k} \alpha_i u_i$, který je definován jako $deg(v) = \sum_{i=1}^{k} \alpha_i$, neboť toto číslo – jak pozorný čtenář snadno nahlédne[1] – je rovno výrazu

$$deg(v) = \frac{1}{2} \sum_{A \subseteq \{1,\ldots,N\}} |A| \cdot (|A| - 1) \cdot v(A). \tag{1}$$

Povšimněme si, že stupeň kombinatorického imsetu je lineární forma. Navíc pomocí formule (1) můžeme rozšířit definici stupně $deg(v)$ na libovolný (nejen kombinatorický) imset v.

Dále můžeme nahlédnout, že jediný kombinatorický imset stupně nula je nulový imset a kombinatorické imsety stupně jedna jsou právě elementární imsety.

Definice. *Množinou strukturálních imsetů $\mathcal{S}_N$ budeme rozumět všechny imsety, jež lze vyjádřit jako nezápornou reálnou*[2] *kombinaci konečně mnoha elementárních imsetů, neboli*

[1] Stačí dokázat, rovnost (1) platí pro všechny elementární imsety, a ukázat, že takto definovaný stupeň je lineární forma.

[2] Lze ukázat, že pokud slovo „reálnou“ nahradíme slovem „racionální“, dostaneme ekvivalentní definici.

$$\mathcal{S}_N = \left\{ \sum_{i=1}^{k} \alpha_i u_i \in \mathbb{Z}^{\mathcal{P}(\{1,2,\ldots,N\})};\ \alpha_i \in \mathbb{R}^+, u_i \in \mathcal{E}_N, k \in \mathbb{N}_0 \right\}.$$

Stupeň strukturálního imsetu je celé číslo, jak je možno snadno nahlédnout z formule (1).

Povšimněme si, že u imsetu nízkého stupně je poměrně snadné rozhodnout, zda je kombinatorický. Na druhé straně otázka, zda-li je či není strukturální, je mnohem obtížnější. I z tohoto, avšak nejen z tohoto důvodu[3] je zajímavá otázka, zda náhodou neplatí, že pro všechna N nastává rovnost $\mathcal{C}_N = \mathcal{S}_N$.

Tento příspěvek si neklade za cíl tuto otázku teoreticky rozřešit. Pouze ji zodpovíme pro dostatečně nízká N a formulujeme tvrzení, jež mohou být základem dalšího bádání. Úhelným kamenem dalšího postupu bude pojem minimální celočíselné Hilbertovy báze.

2 Pojem minimální celočíselné Hilbertovy báze

Definice. *Každý konvexní kužel K v $\mathbb{R}^n$ kónicky generovaný konečnou množinou celočíselných vektorů obsahuje tzv. minimální celočíselnou Hilbertovu bázi, což je množina celočíselných vektorů $w_1, \ldots, w_m$ z K taková, že*

$$\forall x \in K \cap \mathbb{Z}^n\ \exists(\alpha_1, \ldots, \alpha_m) \in \mathbb{N}_0^m : \quad x = \sum_{i=1}^{m} \alpha_i w_i.$$

Tento pojem jsme přejali z knihy [4], kde je dokázáno, že tato definice je konzistentní, a tamtéž je na straně 233 v důkazu věty 16.4 ukázáno, že pokud $e_1, \ldots, e_l$ jsou generátorem výše zmíněného kužele, pak minimální celočíselnou Hilbertovu bázi stačí hledat v mnohostěnu $\mathcal{M}$ tvořeném body

$$\mathcal{M} = \left\{ \sum_{i=1}^{l} \lambda_i e_i;\ \lambda_i \in [0,1] \right\}.$$

Pakliže výše zmíněnou metodu aplikujeme na náš případ, zjistíme, že otázka, zda $\mathcal{C}_N = \mathcal{S}_N$, je ekvivalentní s otázkou, zda minimální celočíselná Hilbertova báze kužele generovaného $\mathcal{E}_N$ (jeho celočíselné body jsou $\mathcal{S}_N$) je rovna $\mathcal{E}_N$ (přičemž zjevně $\mathcal{E}_N$ obsahuje).

Protože rozhodnout, zda daný bod patří či nepatří do mnohostěnu $\mathcal{M}$ a tedy i nalézt všech body mnohostěnu $\mathcal{M}$ je v praxi obtížné, volíme postup pro nalezení Hilbertovy báze pro dané N následovně:

[3] Ověření této hypotézy by mělo zásadní vliv na počítačovou implementaci inferenčního mechanismu.

1. Zavedeme „vhodný“ obal $\mathcal{O}$ mnohostěnu $\mathcal{M}$. Vhodný v tomto kontextu znamená, že počítač dokáže rychle rozhodnout, zda daný imset do $\mathcal{O}$ patří či nikoli, a že dokáže v reálném čase prohledat všechny imsety v $\mathcal{O}$ obsažené.

2. Postupně volíme n od jedné do $|\mathcal{E}_N| = \binom{N}{2} \cdot 2^{N-2}$. Procházíme všechny imsety z $\mathcal{O}$ mající stupeň n, u každého z nich rozhodneme, zda jej lze zapsat jako součet nějakého imsetu stupně $n-1$ (ty už máme vyhledané v minulém kroku a uložené v paměti) a nějakého elementárního imsetu. Pokud ano, pokračujeme, pokud ne, nalezli jsme prvek $\mathcal{O}$, jehož zápis ve tvaru nezáporné celočíselné kombinace prvků $\mathcal{E}_N$ je „problematický“.

3. Pokud algoritmus skončí a žádný „problematický“ imset nenalezne, můžeme učinit závěr, že $\mathcal{C}_N = \mathcal{S}_N$. Pokud jej nalezne, nemusí být závěr jednoznačný, záleží na „obalení“ mnohostěnu $\mathcal{M}$ pomocí $\mathcal{O}$.

Obtíže nastanou již u prvního bodu výše uvedeného scénáře. Pokud jako obal použijeme kužel generovaný prvky $\mathcal{E}_N$, můžeme tento popsat jako průnik jistých poloprostorů[4]. Postup jejich hledání s využitím Fourier-Motzkinovy transformace za pomoci programu PORTA je popsán v [3]. Tento postup však díky jeho obrovské výpočetní složitosti můžeme použít jen pro $N \leq 5$, kdy pro N rovno třem, čtyřem a pěti potřebujeme po řadě 5, 37 a 117 978 nadrovin udávajících výše zmíněné poloprostory. Proto se nadále budeme soustředit pouze na případ $N \leq 5$.

Tento kužel zcela jistě obsahuje mnohostěn $\mathcal{M}$, přičemž můžeme jistě jako obal $\mathcal{O}$ brát pouze takové jeho imsety v, že $\forall A \subseteq \{1, \ldots, N\} : \ |v(A)| \leq deg(v)$.

3 Výsledky počítačových experimentů

Dále rozebereme výsledky našeho výzkumu pro různá N:

$N = 3$:

Použili jsme výše zmíněný postup a v několika málo sekundách sekundách se podařilo ověřit, že $\mathcal{C}_3 = \mathcal{S}_3$.

$N = 4$:

Zde již bylo nutné postupovat mnohem opatrněji. Za prvé si lze všimnout, že pro libovolný strukturální imset v platí

$$\sum_{A \subseteq \{1,\ldots,N\}} v(A) = 0,$$

a také že pro každé $i \in \{1, \ldots, N\}$ platí

[4] odvozených od jeho stěn nebo chcete-li od extremálních paprsků duálního kužele neboli tzv. „skeletonu“.

$$\sum_{A\subseteq\{1,\ldots,N\},\ i\in A} v(A) = 0.$$

Díky těmto dvěma vlastnostem stačí strukturální imset reprezentovat pomocí jeho hodnot pro $A \subseteq \{1,\ldots,N\} : |A| \geq 2$, neboť hodnoty pro $A \subseteq \{1,\ldots,N\} : |A| \leq 1$ jsou již těmito jednoznačně určeny. Tím se nám dimenze problému snižuje o $N+1$, přičemž se nijak nekomplikuje výpočet stupně.

Užitečné je též uvědomit si, že imsety z mnohostěnu $\mathcal{M}$ nabývají pro $A \subseteq \{1,\ldots,N\} : |A| = 2$ hodnot od -4 do 2, pro $|A| = 3$ hodnot od -3 do 3 a pro $|A| = 4$ hodnot od 0 do 6. Stačí se tedy omezit se na takovéto imsety. Tato změna je o to významnější, že nám umožní změnit meze do sebe vnořených $11 = 2^4 - (4+1)$ for–cyklů.

Dále je důležité vhodně zvolit datové typy, aby výpočet nebyl příliš náročný na paměť, a v bodě 2 výše zmíněného scénáře šikovně implementovat vyhledávání v imsetech stupně $n-1$ za pomoci jejich setřízení a hašovací tabulky, jinak úloha neskončí v reálném čase. Zdrojový kód pro GNU Pascal je možné nalézt na adrese:

`http://5r.matfyz.cz/ctyri.pas.`

Výpočet na stroji Artax s procesorem Intel Pentium 4 HT 2800 MHz a 1 GB paměti trval 12 minut, přičemž bylo využito 530 MB operační paměti.

$N = 5$:

Zde jsme sice využili další zmenšení obalu $\mathcal{O}$ založené na tom, že každý strukturální imset v musí zjevně splňovat

$$\sum_{A\subseteq\{1,\ldots,N\}} (v(A))^+ \leq 2\ deg(v), \qquad \sum_{A\subseteq\{1,\ldots,N\}} (v(A))^- \leq 2\ deg(v).$$

A také, že pro libovolnou $B \subset \{1,\ldots,N\}$ musí platit

$$\sum_{A:\ B\subseteq A\subseteq\{1,\ldots,N\}} v(A) \geq 0,$$

nicméně i při této redukci jsme byli schopni vyšetřit jen imsety stupně nejvýše čtyři, přičemž výpočet trval necelé tři dny. Program je k nahlédnutí na adrese

`http://5r.matfyz.cz/pet.tar.gz.`

4 Myšlenka fixovaného stupně

Nadějným směrem dalšího možného postupu je namísto „obalování“ celého mnohostěnu $\mathcal{M}$ aproximovat průniky $\mathcal{M}$ s množinou imsetů daného stupně n. Lze totiž ukázat:

Věta. *Nechť v je strukturální imset stupně $deg(v) = n$ z mnohostěnu*

$$\mathcal{M} = \left\{ \sum_{e_i \in \mathcal{E}_N} \lambda_i e_i;\ \lambda_i \in [0,1] \right\},$$

potom tento imset náleží i do mnohostěnu, jenž je konvexním obalem množiny všech součtů n různých elementárních imsetů neboli množiny

$$\mathcal{B}_n = \left\{ v = \sum_{e_i \in D}^{n} e_i;\ D \subseteq \mathcal{E}_N, |D| = n \right\}.$$

Důkaz. Dokážeme nejprve pomocné tvrzení, že každý bod r–rozměrné jednotkové krychle, jehož součet souřadnic je $n \in \mathbb{N}_0$, je konvexní kombinací bodů z krychle o souřadnicích $(s_1, \ldots, s_r)$ takových, že $\forall i : s_i \in \{0,1\}$ a $s_1 + s_2 + \cdots + s_r = n$. Pokud toto tvrzení platí, pak po dosazení $r = |\mathcal{E}_N|$, koeficienty λ_i použité v $\mathcal{M}$ definují příslušný bod krychle a prohozením příslušných sum snadno nahlédneme požadovaný závěr.

Pomocné tvrzení dokážeme indukcí podle r zároveň pro všechna přípustná n: pro $r \leq 2$ tvrzení evidentně platí. Předpokládáme-li platnost tvrzení pro všechna $r' < r$, pak jeho platnost pro r (a libovolné n) dokážeme nejprve pro body na stěnách krychle. Vezměme si tedy libovolnou stěnu krychle, bez újmy na obecnosti tedy třeba tu s pevnou první souřadnicí $s_1 = 0$, respektive $s_1 = 1$. Tedy má-li mít bod na této stěně součet souřadnic n, musí být součet druhé až r–té souřadnice roven n, respektive $n - 1$, a indukční předpoklad nám zaručuje, že takovýto bod již bude požadovanou konvexní kombinací.

Zbývá totéž dokázat o bodech uvnitř krychle. Ale každý takový bod je konvexní kombinací dvou bodů, pro něž platnost tvrzení již byla nahlédnuta. Vskutku ke každému bodu uvnitř krychle můžeme přičíst i odečíst vhodný násobek vektoru $\left(1, \frac{1}{r-1}, \ldots, \frac{1}{r-1}\right)$ tak, že dostaneme body ležící na stěnách krychle, z kterých lze tento bod nakombinovat. □

Problémem, na který opět narážíme, je obrovská výpočetní složitost. Ukazuje se, že asi nebude možné najít přesný popis konvexního uzávěru $\mathcal{B}_n$ ve tvaru průniku poloprostorů, ale spíše nějakou jeho vnější aproximaci.

5 Závěr

Podařilo se nám hypotézu, že $\mathcal{C}_N = \mathcal{S}_N$ ověřit pro $N \leq 4$.

Problém pro $N = 5$ nadále zůstává otevřený a uvítáme náměty či rady, jak postupovat při jeho řešení. Zatím je známo pouze to, že pakliže existuje prvek minimální celočíselné Hilbertovy báze $\mathcal{S}_5$ neobsažený v $\mathcal{E}_5$, pak má tento prvek stupeň ostře vyšší než 4.

Nadějné se zdá být hledání imsetů v mnohostěnech pro jednotlivé stupně, které však zatím naráží na příliš vysokou časovou náročnost.

Reference

[1] Studený M. (2001). *On mathematical description of probabilistic conditional independence structures*. Doktorská práce, ÚTIA AV ČR.

[2] Studený M. (2005). *Probabilistic conditional independence structures*. Springer-Verlag London.

[3] Studený M., Bouckaert R.R., Kočka T. (2000). *Extreme supermodular set functions over five variables*. Výzkumná zpráva číslo **1977**, ÚTIA AV ČR.

[4] Schrijver A., (1998). *Theory of linear and integer programming*. John Wiley.

Poděkování: Práci na tomto příspěvku byla poskytnuta podpora z grantu GA ČR 201/04/0393.

Adresa: P. Šimeček, M. Studený, ÚTIA AV ČR, Pod Vodárenskou věží 4, Praha 8

E-mail: `simecek@karlin.mff.cuni.cz, studeny@utia.cas.cz`

NEJMENŠÍ USEKNUTÉ ČTVERCE (LTS) JAKO DIAGNOSTICKÝ NÁSTROJ

Marie Šimečková

Klíčová slova: Nejmenší useknuté čtverce (LTS), diagnostika.

Abstrakt: Metoda nejmenších useknutých čtverců (LTS) je robustní variantou metody nejmenších čtverců. Příspěvek se zabývá odhadem LTS v lineárním modelu a je zaměřen především na jeho užití v případě, kdy závislá veličina je kromě sledovaných veličin ovlivněna ještě dalším faktorem. Metoda je předvedena na příkladě modelování docházky australských dětí do školy v závislosti na jejich věku, pohlaví, úspěšnosti ve škole a jejich etnickém původu. Je ukázáno, že i když do modelu závislost na etnickém původu nezahrneme, s užitím LTS zjistíme, že soubor dětí se rozpadá na dvě skupiny s různým poměrem dětí domorodého původu.

1 Úvod a definice LTS

V textu se budeme zabývat lineárním regresním modelem. Předpokládáme model

$$Y_i = \boldsymbol{X}_i\boldsymbol{\beta} + e_i, \quad i = 1, \ldots, n, \ n \in \mathbb{N}. \tag{1}$$

$\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ je vektor závisle proměnných, $\boldsymbol{X}_i$ je i-tý řádek matice vysvětlujících proměnných $\boldsymbol{X}$ o rozměrech $n \times p$, $p \in \mathbb{N}$ (první sloupec této matice může tvořit vektor jedniček), hodnost matice $\boldsymbol{X}$ je kladná a menší než n. Dále vektor chyb $\boldsymbol{e} = (e_1, \ldots, e_n)^T$ je vektor nezávislých stejně rozdělených náhodných veličin s nulovou střední hodnotou a konečným rozptylem a $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T \in \mathbb{R}^p$ je vektor parametrů.

Označme $r_i(\boldsymbol{\beta}) = Y_i - \boldsymbol{X}_i\boldsymbol{\beta}$, $i = 1, \ldots, n$ rezidua a $r^2_{(1)} \leq r^2_{(2)} \cdots \leq r^2_{(n)}$ vektor pořádkové statistiky jejich druhých mocnin.

Odhad metodou nejmenších čtverců (*least squares*) je takový vektor $\hat{\boldsymbol{\beta}}$, který minimalizuje výraz

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{n} r_i^2(\boldsymbol{\beta}).$$

Odhad metodou nejmenších useknutých čtverců (*least trimmed squares, LTS*) je takový vektor $\hat{\boldsymbol{\beta}}$, který minimalizuje výraz

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{h} r^2_{(i)}(\boldsymbol{\beta}),$$

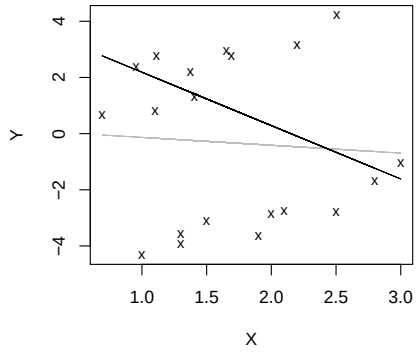
kde h je daná konstanta, $0 < h \leq n$. Nejvyšší bod selhání (*breakdown point*) má tento odhad pro h rovno $\lfloor \frac{n}{2} \rfloor + \lfloor \frac{p+1}{n} \rfloor$ (je roven $\frac{\lfloor \frac{n-p}{2} \rfloor + 1}{n}$). V dalším textu

budeme předpokládat tuto hodnotu konstanty h, nebude-li řečeno jinak. Více o vlastnostech tohoto odhadu v Rousseeuw, Leroy [1].

Dále se budeme zabývat případem, kdy model 1 není splněn a pozorování $(Y_1, \boldsymbol{X}_1), \ldots, (Y_n, \boldsymbol{X}_n)$ lze rozdělit do dvou skupin s různými závislostmi. Tzn.

$$Y_i = \boldsymbol{X}_i\boldsymbol{\beta} + e_i, \quad i \in \{i_1, \ldots, i_k\} \subset \{1, \ldots, n\},\ k < n$$
$$Y_i = \boldsymbol{X}_i\boldsymbol{\gamma} + e_i, \quad i \in \{1, \ldots, n\} \setminus \{i_1, \ldots, i_k\}.$$

Jednoduchý příklad můžeme vidět na obr 1. Na první pohled je vidět, že se jedná o směs dvou modelů se stejnou závislostí na veličině X, ale s různým absolutním členem. Jako velmi často v podobných situacích, většina pozorování, které odhad LTS nevyužije (jejich čtverce reziduí jsou velká), pochází z jednoho modelu, a proto odhad LTS je výrazně odlišný od odhadu metodou nejmenších čtverců.

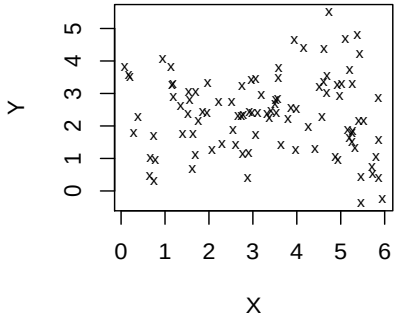


Obrázek 1: Odhad metodou nejmenších čtverců (šedá linka) a metodou LTS (černá linka) v lineární modelu závislosti náhodné veličiny Y na veličině X.

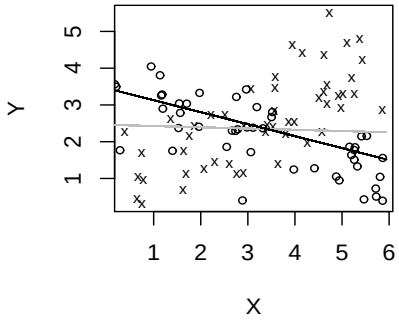
V tomto případě každý z obrázku vidí, že k vysvětlení veličiny Y nestačí veličina X a i grafy reziduí to jasně potvrzují. V případě více vysvětlujících proměnných nám ale obrázek nepomůže a i ve dvourozměrném případě může nastat situace na první pohled nejasná.

Podívejme se na obrázek 2. Zde žádné dvě odlišné skupiny vidět nejsou. Přesto se ale odhady metodou nejmenších čtverců a metodou LTS výrazně liší, jak je vidět na obrázku 3. Ve skutečnosti se jedná o směs dvou populací. Pro body označené křížkem platí model $Y = 1 + 0{,}5\ X + e$, pro body označené kroužkem platí model $Y = 3 - 0{,}3\ X + e$. (Chybové členy mají v obou případech normální rozdělení s nulovou střední hodnotou a rozptylem 0,5).

Tentokrát nás ani analýza reziduí odhadu metodou nejmenších čtverců neupozorní na možnost nesplnění předpokladů, viz obrázek 4, p-hodnota Shapirova - Wilkova testu normality je rovna 0,914.



Obrázek 2: Graf bodů $(X_1, Y_1), \ldots, (X_{100}, Y_{100})$.



Obrázek 3: Odhad metodou nejmenších čtverců (šedá linka) a metodou LTS (černá linka) v lineární modelu závislosti náhodné veličiny Y na veličině X.

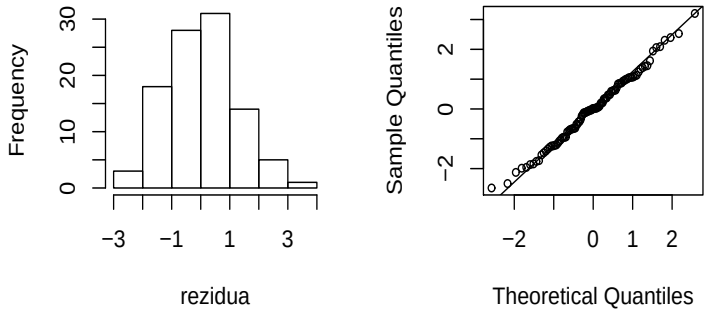
Odhad LTS rozdělil body (Y_i, X_i) na dvě části, první část k odhadu použil a druhou nepoužil. Tmavá linka na obrázku 3 tedy ukazuje odhad metodou nejmenších čtverců na základě 51 bodů „vybraných“ metodou LTS.

Po opakovaném použití odhadu LTS na tyto dvě skupiny a vyzkoušením odhadů LTS s různými hodnotami konstanty h bychom se dopracovali k ještě přesnějšímu rozdělení obou populací.

Celou analýzu si předvedeme na následujícím příkladě.

2 Docházka dětí do školy na australském venkově

Data pochází ze sociologické studie o australských dětech aborginského a běloského původu S. Quinové a jsou volně přístupná v prostředí R v knihovně `MASS` pod jménem `quine`.



Obrázek 4: Histogram a normální QQ-plot reziduí odhadu metodou nejmenších čtverců.

Do studie bylo zahrnuto 146 dětí ze čtyřech věkových skupin (poslední ročník základní školy a první tři ročníky střední školy) - tomu odpovídá veličina Age nabývající hodnot 1, 2, 3, 4. Dále je u dětí sledováno pohlaví (veličina Sex nabývající hodnot 0 pro dívku a 1 pro chlapce), etnický původ (veličina Eth nabývající hodnot A - Aborginec a N - jiný původ), úspěšnost ve škole (veličina Lrn nabývající hodnot 0 pro průměrného žáka a 1 pro slabého žáka, ve čtvrté věkové skupině není žádný pomalý žák) a kolik dní dítě během školního roku chybělo ve škole (veličina Days).

Studie se zúčastnilo 66 chlapců a 80 dívek, 69 Aborginců a 77 bílých dětí, 63 slabých a 83 průměrných žáků, ve věkových skupinách 1 až 4 bylo 17, 46, 40 a 30 dětí. Počet zameškaných dní se pohyboval v rozmezí 0 až 81, průměrná hodnota je 16,5 a medián 11.

Budeme se zabývat hledáním závislosti počtu zameškaných dnů na pohlaví, věku a úspěšnosti ve škole, tedy předpokládáme model

$$Days = \beta_0 + \beta_1 \cdot Sex + \beta_2 \cdot Age + \beta_3 \cdot Lrn. \tag{2}$$

Ukážeme, že odhad LTS rozdělí děti do dvou skupin s výrazně odlišným počtem dětí domorodého původu.

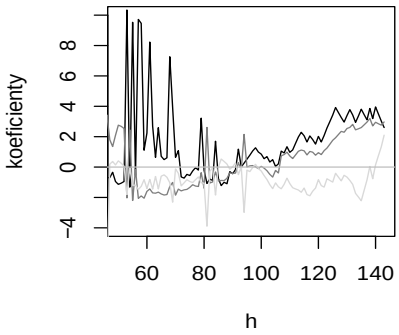
Odhady koeficientů v modelu 2 metodou nejmenších čtverců a metodou LTS (s konstantou h takovou, aby bod selhání byl maximální, tedy $h = 75$) jsou v tabulce 1.

	β_1	β_2	β_3
LS	3,63	3,30	3,70
LTS	-0,65	-1,34	-0,95

Tabulka 1: Odhady koeficientů v modelu 2, $h = 75$.

	chlapci	dívky	poměr	
dohromady	66	80	0,83	
1. skupina	29	46	0,63	
2. skupina	37	34	1,09	
	slabí	průměrní	poměr	
dohromady	63	83	0,76	
1. skupina	35	40	0,86	
2. skupina	28	43	0,65	
	min	maximum	průměr	medián
dohromady	1	4	2,54	2,5
1. skupina	1	4	2,33	2
2. skupina	1	4	2,76	3
	min	maximum	průměr	medián
dohromady	0	81	16,5	11
1. skupina	0	13	5,4	5
2. skupina	0	81	28,1	23

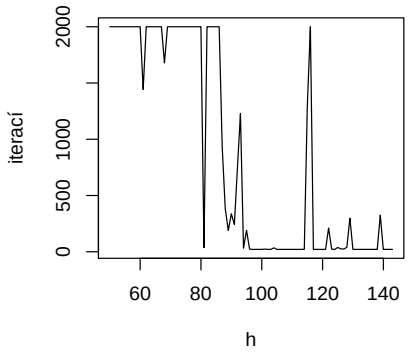
Tabulka 2: Charakteristiky pohlaví, úspěšnosti ve škole, věku a počtu zameškaných dní pro celý soubor, skupinu použitou k odhadu LTS ($h = 75$) a skupinu zbývajících dětí.



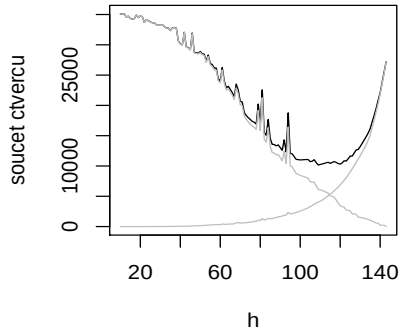
Obrázek 5: Změny odhadu koeficientů metodou LTS v závislosti na konstantě h (černě koef. β_1, světleji koef. β_2, nejsvětleji koef. β_3.

Závislost zjištěná pomocí obou metod je velmi různá, v prvním případě je kladná, v druhém záporná. Metoda LTS rozdělila soubor do dvou skupin, jejich charakteristiky viz tabulka 2.

Nyní zkusíme, co se stane, když použijeme různé hodnoty konstanty h. Na obrázcích 5 a 6 jsou pro h mezi 50 a 143 grafy změn koeficientů a počtů iterací nutných k nalezení odhadu LTS dle algoritmu popsaném ve Víšek [2]. Na obrázku 7 jsou součty čtverců reziduí lineárního modelu spočteného z dat



Obrázek 6: Závislost počtu iterací při určování LTS odhadu na konstantě h.



Obrázek 7: Závislost součtů čtverců reziduí lineárního modelu spočteného z dat použitých v odhadu LTS a ze zbývajících dat (šedé linky) a jejich součtu (černá linka) na konstantě h.

použitých v odhadu LTS, ze zbývajících dat a jejich součet, v závislosti na konstantě h.

Odhady koeficientů se ustálí pro h přibližně mezi 105 a 125, počet iterací je nízký pro h mezi 95 a 115 (2 000 je maximální počet iterací v algoritmu), součet součtu čtverců v obou skupinách je nejnižší pro h mezi 100 a 120. Dále budeme používat h rovné 110.

Pak odhady koeficientů jsou $\beta_1 = 0{,}97$, $\beta_2 = 0{,}73$ a $\beta_3 = -1{,}09$. Podíváme se na vlastnosti dvou souborů, na které děti rozdělila metoda LTS tentokrát, viz tabulka 3.

Metoda LTS rozdělila děti téměř přesně na děti s nižším a vyšším počtem zameškaných dní (hranice je 22 - 23 dní). U žádné z vysvětlujících veličin není statisticky významný rozdíl (na hladině 0,05) mezi první a druhou skupinou (testováno χ-kvadrát testem, resp. Wilcoxonovým dvouvýběrovým testem).

	chlapci	dívky	poměr	
dohromady	66	80	0,83	
1. skupina	47	63	0,75	
2. skupina	19	17	1,12	
	slabí	průměrní	poměr	
dohromady	63	83	0,76	
1. skupina	47	63	0,75	
2. skupina	16	20	0,80	
	min	maximum	průměr	medián
dohromady	1	4	2,54	2,5
1. skupina	1	4	2,46	2
2. skupina	1	4	2,81	3
	min	maximum	průměr	medián
dohromady	0	81	16,5	11
1. skupina	0	23	8,7	7
2. skupina	22	81	40,2	36,5

Tabulka 3: Charakteristiky pohlaví, úspěšnosti ve škole, věku a počtu zameškaných dní pro celý soubor, skupinu použitou k odhadu LTS ($h = 75$) a skupinu zbývajících dětí.

	domorodý	běloský	poměr
dohromady	69	77	0,90
1. skupina	45	65	0,69
2. skupina	24	12	2,00

Tabulka 4: Poměry počtu dětí aborginského a bělošského původu v celém souboru, ve skupině použité k odhadu LTS ($h = 110$) a ve skupině zbývajících dětí.

Podíváme se nyní na etnickou příslušnost dětí, viz tabulka 4.

Rozdíl v podílu aborginských dětí v 1. a 2. skupině je statisticky významný (p-hodnota χ-kvadrát testu je rovna 0,013). Protože obě etnika nebyla rozdělena přesně, je možné, že docházka do školy nezávisí na etnickém původu, ale spíše na něčem, co s ním souvisí, například na sociálním postavení rodiny nebo na vzdělání rodičů.

Reference

[1] Rousseeuw P. J., Leroy A. M. (1987). *Robust regression and outlier detection*. Wiley, New York.

[2] Víšek J. Á. (2000). *On the diversity of estimates*. Computational Statistics & Data Analysis **34**, 67 – 89.

Poděkování: Příspěvek vznikl s podporou grantu GA ČR 402/03/0084.

Adresa: Katedra pravděpodobnosti a matematické statistiky MFF UK, Sokolovská 83, Praha 8

E-mail: simecko@karlin.mff.cuni.cz

WEIGHTED GMM ESTIMATION

Jan Ámos Víšek

Keywords: Robustified GMM-estimation, robust regression, equivariance of the estimation, influential points, contamination, heteroscedasticity, instrumental variables, the least weighted squares..

Abstract: The paper presents an implementation of ideas of robust statistics in the *Generalized Method of Moments*. Robustification of GMM is proposed in the sense of the *Least Weighted Squares*. Heuristics of both, GMM as well as of LWS are briefly recalled. The examples are only sketched (due to limit of space).

1 Introduction and notations

In 1982 Lars Peter Hansen in the pioneering paper [8] proposed and discussed *Generalized Method of Moments* (GMM). One year later Peter Rousseeuw [12], [13], see also [6] defined the *Least Median of Squares* (LMS) and the *Least Trimmed Squares* (LTS), the first two feasible estimates with (possibly) 50% breakdown point. Both of them fulfill the Hampel's program of modern estimation [6] or even its enlargement, adding to the Hampel's the requirement of the equivariance, the existence of a reliable algorithm (possibly with an easy available implementation), the subsample stability[1] and, last but not least, a working heuristics, see [17], [19]; for algotritms see [1] and [14] or [18]. Nevertheless, in the case of dynamic framework, there is a serious reason why we should use another robust estimator of regression coefficients[2], namely LWS [17]. In the next section we shall recall philosophy of GMM. The third one is devoted to LWS, to remind the heuristics of them. Finally the last section offers a proposal how to profit from the "aliance" of both approaches.

Let N denote the set of all positive integers, R the real line, and R^p the p-dimensional Euclidean space and assume throughout the paper a fixed probability space $(\Omega, \mathcal{A}, P)$ probability space.

2 Generalized method of moments

Let us assume that we would like to estimate the "true" value θ^0 of a parameter of a parameterized family of models and there are two estimators, given as solution of $g^{(1)}(\theta, x_1, x_2, \ldots, x_n) = 0$ and $g^{(2)}(\theta, x_1, x_2, \ldots, x_n) = 0$, say. No of them is uniformly better then the other. A very first idea may be

[1] This point is fulfilled neither by LMS nor LTS. It is a consequence of the fact that both estimators too much rely on the selected observations while suppressing nearly completely the influence of the rest of data.

[2] The deletion of some observations by LMS or LTS represents evidently an unwelcome problem, since it may considerably modify the character of (possible) serial correlation of disturbances and/or explanatory variables.

to find a combination of them which can be uniformly better than both. Formally it may be done as follows. Select a symmetric positive definite matrix W and put

$$g(\theta, \mathbf{x}) = \left(g^{(1)}(\theta, x_1, x_2, \ldots, x_n), g^{(2)}(\theta, x_1, x_2, \ldots, x_n)\right)'$$

and

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \; g'(\theta) \cdot W \cdot g(\theta) \tag{1}$$

A nice example of estimating the mean of normal distribution, provided an additional information that $\mu = 3\sigma^2$ holds, can be found in [25].

The optimal choice of W is the inverse covariance matrix of $g(\mu, \mathbf{d})$, see [24]. Of course, we try to find a solution of the extremal problem (1) by finding a solution of equation

$$\left[\frac{\partial g(\theta)}{\partial \theta}\right]' \cdot W \cdot g(\theta) = 0 \tag{2}$$

where $\frac{\partial g(\theta)}{\partial \theta}$ denotes the matrix of partial derivatives $\left\{\frac{\partial g^{(j)}(\theta, \mathbf{x})}{\partial \theta_k}\right\}$, $k = 1, 2, \ldots, p$, $j = 1, 2$ (we have assumed that $\Theta \subset R^p$). (2) can be considered as a system of the *normal equations*.

Now, let us turn for a while to the *linear regression model* (say, in the dynamic framework and with random explanatory variables)

$$Y_t = \sum_{j=1}^{p} X_{tj}\beta_j^0 + \varepsilon_t = X_t'\beta^0 + \varepsilon_t, \; t = 1, 2, \ldots, T$$

the *Ordinary Least Squares* estimate is given as solution of the *normal equations*

$$X'(Y - X\beta) = 0. \text{ or equivalently } \frac{1}{T}X'(Y - X\beta) = 0 \tag{3}$$

which is an empirical counterpart of an *orthogonality condition*

$$I\!E\{X_1(Y_1 - X_1'\beta)\} = 0. \tag{4}$$

As we shall see later, for robust estimation of regression model (see (10) below), we have a similar system of *normal equations.*

So, in many cases we appear at a situation (when looking for an underlying model of data in question): *We assume existence of a vector of explanatory variables, say $X_t \in R^p$, and of matrix*[3] *of unobservable disturbances, say $e_t = H(X_t, \beta^0)$. Then to be able to establish a consistent estimator of β^0, we need the orthogonality condition*

$$I\!E\{e_1 \otimes X_1\} = 0.$$

However, there are many models exhibiting correlation of disturbances with explanatory variables[4]*. Then we employ (a collection of) instruments, say $z_t = G(X_t, \beta^0)$, such that*

[3]Generally, the disturbance for each observation can be multidimensional, if the "response variable" is not scalar.

[4]The most frequently recalled is probably model assuming the measurement of explanatory variables with a random error, see [15] or [16]. For many other examples see either [9] or [2].

$$I\!E\left\{e_1 \otimes z_1\right\} = I\!E\left\{H(X_1, \beta^0) \otimes G(X_1, \beta^0)\right\} = 0 \tag{5}$$

(and of course we assume that correlation of the explanatory variables X_t with the instruments z_t is (are) as high as possible).

The empirical counterpart of (5) is then

$$g_T(X, \beta) = \frac{1}{T}\sum_{t=1}^{T}\left[H(X_t, \beta) \otimes G(X_t, \beta)\right] = 0. \tag{6}$$

Unfortunately, especially due to substitution of the explanatory variables by the instruments, we need not be able to find any $\beta \in R^p$ fulfilling (6). Then we prefer to give the definition of the estimator in the following form.

Definice 2.1. *Let W be a symmetric positive definite matrix. Then the solution of the extremal problem*

$$\hat{\beta}^{(GMM,W)} = \underset{\beta \in R^p}{arg\,min}\ g'(X, \beta) \cdot W^{-1} \cdot g(X, \beta) \tag{7}$$

will be called the Estimator obtained by the Generalized Method of Moments, or GMM-estimator for short.

We have already mentioned that the optimal selection of W is the covariance matrix $C = I\!E\left[g_T(x, \beta) \cdot g'_T(x, \beta)\right]$ which is generally unknown. Hence, provided it is regular and consistently estimable, it is substituted by $\widehat{C}$, say. Of course, sometimes either no consistent estimate of W is available. Then we have to employ some heuristics for selecting some positive definite matrix as a substitute for the covariance one.

3 The least weighted squares

As we have already said LMS as well as LTS are disqualified for estimating the coefficients of dynamic regression model since we can't afford to delete some rows (without distorting the structure of (possible) serial correlation). On the other hand, we would like to depress a bit the influence of some observation either due to the suspicious values of explanatory variables or due to strange values of response. The *Least Weighted Squares* (LWS) is candidate which fulfills all points of enlarged Hampel's program, including the existence of the reliable algorithm and available implementation, see e.g. [10] or [11]. By the way, it means that, in distinction to the M-estimators, it is *scale-* and *regression-equivariant*[5], without estimating scale of disturbances

[5]Let us recall that having denoted $M(n, p)$ the set of all matrices of type $(n \times p)$ and recalling that the estimator $\hat{\beta}$ can be considered as a mapping

$$\hat{\beta}(Y, X):\ M(n, p+1) \to R^p,$$

the estimator $\hat{\beta}$ of β^0 is called *scale-equivariant*, if for any $c \in R^+, Y \in R^n$ and $X \in M(n, p)$ we have

$$\hat{\beta}(cY, X) = c\hat{\beta}(Y, X)$$

and *regression-equivariant* if for any $b \in R^p, Y \in R^n$ and $X \in M(n, p)$

$$\hat{\beta}(Y + Xb, X) = \hat{\beta}(Y, X) + b.$$

and performing studentization of residuals. Of course, the tax we pay for this advantage are rather complicated proofs of consistency of LWS, see again [10] or [11]. Let us recall the definition of LWS.

To this end, let for any $\beta \in R^p$ $r_t(\beta) = Y_t - X_t'\beta$ denote the t-th residual and $r^2_{(t)}(\beta)$ the t-th order statistic among the squared residuals. It means that $r^2_{(1)}(\beta) \leq r^2_{(2)}(\beta) \leq \ \ldots \ \leq r^2_{(T)}(\beta)$.

Definice 3.1. *Víšek [17] For a weight function* $w : [0,1] \to [0,1]$ *the solution of the extremal problem*

$$\hat{\beta}^{(LWS,T,w)} = \ \underset{\beta \in R^p}{arg\,min} \ \sum_{t=1}^{T} w\left(\frac{t-1}{T}\right) r^2_{(t)}(\beta) \tag{8}$$

will be called the Least Weighted Squares estimator[6] *(LWS).*

The weight function is usually assumed to fulfill:

C1 *Weight function* $w : [0,1] \to [0,1]$ *is absolutely continuous and nonincreasing, with the derivative* $w'(\alpha)$ *bounded from below by* L, $w(0) = 1$.

Now, for any $t \in \{1, 2, \ldots, T\}$ let us put $\pi(\beta, t) = j \in \{1, 2, \ldots, T\}$ so that $r^2_t(\beta) = r^2_{(j)}(\beta)$ (notice that $\pi(\beta, t)$ is r.v.; in analogy with the *theory of rank test*, we can call it the *rank of* t*-th observation*, see [7]). Then

$$\hat{\beta}^{(LWS,T,w)} = \ \underset{\beta \in R^p}{\arg\min} \ \sum_{t=1}^{T} w\left(\frac{\pi(\beta,t)-1}{T}\right) r^2_t(\beta). \tag{9}$$

Further, for any $T \in N$ by $\mathcal{P}_T$ let us denote the set of all permutations of the indices $\{1, 2, \ldots, T\}$ and by π_t the t-th coordinate of the vector $\pi \in \mathcal{P}_T$. Denoting $\pi(\beta) = (\pi(\beta, 1), \pi(\beta, 2), \ldots, \pi(\beta, T))'$, we have $\pi(\beta) \in \mathcal{P}_T$. Now, keeping in mind **C1**, (8) and (9), we easy verify that for any $\pi \in \mathcal{P}_T$

$$\sum_{t=1}^{T} w\left(\frac{\pi(\beta,t)-1}{T}\right) r^2_t(\hat{\beta}^{(LWS,T,w)}) \leq \sum_{t=1}^{T} w\left(\frac{\pi_t-1}{T}\right) r^2_t(\hat{\beta}^{(LWS,T,w)})$$

so that for any $\omega \in \Omega$ there is some $\pi \in \mathcal{P}_T$ such that for the vector of weights $w^* = (w(T^{-1}(\pi_1 - 1)), w(T^{-1}(\pi_2 - 1)), \ldots, w(T^{-1}(\pi_T - 1)))$ we have $\hat{\beta}^{(LWS,T,w)} = \hat{\beta}^{(WLS,T,w^*)}$, i.e. (in words) the *Least Weighted Squares* estimator is equal to the (classical) *Weighted Least Squares* estimator (with the weights w^*_t's) at given, fixed $\omega \in \Omega$. As $\hat{\beta}^{(WLS,T,w^*)}$ is the solution of normal equations (at given, fixed $\omega \in \Omega$)

$$\sum_{t=1}^{T} w^*_t X_t \left(Y_t - X_t'\beta\right) = 0$$

[6] See also [4] where the estimator is called the *Smoothed Least Trimmed Squares.*

taking into account successively all $\omega \in \Omega$, we conclude that $\hat{\beta}^{(LWS,T,w)}$ is one of solutions of *normal equations*

$$I\!N E_{X,T}(\beta) = \sum_{t=1}^{T} w\left(\frac{\pi(\beta,t)-1}{T}\right) X_t\left(Y_t - X_t'\beta\right) = 0. \tag{10}$$

For proposing a robustification of GMM another form of normal equations however will be more instructive.

Definice 3.2. *Denote by* $I\{X_t < x, e_t < v\}$ *the indicator of the set of all* ω*'s for which the explanatory variable* $X_t \in R^p$ *is smaller than* $x \in R^p$ *(coordinatewise) and the* t*-th disturbance is smaller than* $v \in R$*. Then define for any* $T \in N$ *the joint empirical distribution function of the explanatory variables and of disturbances as*

$$F_{X,e}^{(T)}(x,v) = F_{X,e}^{(T)}(x,v,\omega) = \frac{1}{T}\sum_{t=1}^{T} I\{X_t < x, e_t < v\}$$

Notice that $1/T\sum_{t=1}^{T} I\{X_t < x, e_t < v\} = 1/T\sum_{t=1}^{T} I\{X_t(\omega) < x, e_t(\omega) < v\}$.

Remark 3.1. *It seems at the first glance a bit strange to construct empirical distribution function when the random variables are not indentically distributed. Nevertheless, notice that* $I\!\!E I\{X_t < x, e_t < v\} = F_{X,e,t}(x,v)$ *and hence according to the (strong) law of large numbers for independently but non-identically distributed r.v.'s (see e.g. [3] or [5]) we have*

$$\frac{1}{T}\sum_{t=1}^{T}\left[I\{X_t < x, e_t < v\} - F_{X,e,t}(x,v)\right] \;\to\; 0 \; a.s.$$

or in an alternative notation

$$F_{X,e}^{(T)}(x,v) - \frac{1}{T}\sum_{t=1}^{T} F_{X,e,t}(x,v) \;\to\; 0 \; a.s.$$

Further, let us denote for any $\beta \in R^p$ the distribution of the absolute value of residual by $F_\beta(u)$. In other words,

$$\begin{aligned} F_\beta(u) = P(|Y_1 - X_1'\beta| < u) &= P(|e_1 - X_1'\left(\beta - \beta^0\right)| < u) \\ &= \int I\{|x'\left(\beta - \beta^0\right) - v| < u\}\mathrm{d}F_{X,e}(x,v). \end{aligned} \tag{11}$$

Moreover, for any $\beta \in R^p$ the empirical distribution of the absolute value of residual will be denoted $F_\beta^{(T)}(u)$. It means that we have

$$F_\beta^{(T)}(u) = \frac{1}{T}\sum_{t=1}^{T} I\left\{|r_t(\beta)| < u\right\} = \frac{1}{T}\sum_{t=1}^{T} I\left\{|e_t - X_t'\left(\beta - \beta^0\right)| < v\right\}. \tag{12}$$

It immediately gives

$$F_{\beta}^{(T)}(|r_t(\beta)|) = \frac{\pi(\beta,t)-1}{T}$$

and so (3.1) can be written as

$$\sum_{t=1}^{T} w\left(F_{\beta}^{(T)}(|r_t(\beta)|)\right) X_t \left(Y_t - X_t'\beta\right) = 0. \tag{13}$$

Remark 3.2. *Let us also recall that the Least Weighted Squares fulfils most of points of the enlarged Hampel's program of point estimation as given in previous. Of course, there are still only a few attempts for fulfilling the point about diagnostics, sensitivity and accompanying procedures, see e.g. [21], [22], [23].*

4 The weighted GMM estimation

In this section we are going to give a proposal how to "merge" the idea of robust estimation with the idea of generalized method of moments. Recalling that we have denoted by $e_t = H(X_t, \beta^0)$ unobservable vector of disturbances[7] and by $z_t = G(X_t, \beta^0)$ the vector of instrumental variables, let us enlarge the notation as follows. The empirical distribution function of the absolute values of residuals will be denoted by $F_{\beta}^{(T)}(|H(X_t,\beta)|)$. In what follows, we shall assume that

$$I\!E\left\{w\left(F_{\beta^0}^{(T)}(|H(X_t,\beta^0)|)\right)[u_t \otimes z_t]\right\} = 0 \text{ for } t = 1,2,\ldots,T.$$

Similarly as in previous, let us put for any $\beta \in R^p$

$$g_T^w(x,\beta) = \frac{1}{T}\sum_{t=1}^{T}\left[w\left(F_{\beta}^{(T)}(|H(X_t,\beta)|)\right)\cdot H(X_t,\beta)\otimes G(X_t,\beta)\right].$$

Definice 4.1. *Let W be a symmetric positive definite matrix. Then the solution of the extremal problem*

$$\hat{\beta}^{(GMM,\widehat{W})} = \underset{\beta\in R^p}{arg\,min}\ [g_T^w(x,\beta)]'\cdot\widehat{W}^{-1}\cdot g_T^w(x,\beta) \tag{14}$$

will be called the estimator obtained by Weighted Generalized Method of Moments, or WGMM-estimator for short.

Of course, analogously as in previous, if $I\!E\left[g_T^w(X,\beta^0)\left(g_T^w(X,\beta^0)\right)'\right]$ is regular matrix which is consistently estimable by a regular matrix $\widehat{W}$, we employ it in the role of W. As this is a very first attempt to employ the ideas of robust statistics in the estimation based on the *Generalized Method of Moments*. we have not yet at hand any optimality result about selection of W.

[7] Similarly as in previous, we shall denote for any $\beta \in R^p$ by $H(X_t,\beta)$ the residuals.

5 Conclusion

It is straightforward that the *Least Weighted Squares* are special case of the *WGMM-estimator*. Having put $G(X_t, \beta) = X_t$, $H(X_t, \beta) = Y_t - X_t'\beta$ and $W = I\!I$ and taking into account (13), we verify it (of course, the quadratic form in (14) turns to zero). On the other hand, (14) can give a reasonable (robust) estimate even for (much) more general cases, e. g. when we would not like to restrict ourselves to the one (robust) estimator of regression coefficients. The reason may be that we hesitate which of robust estimators can be the best "tailored" for our data (e. g. due to our only vague knowledge of level or of structure of contamination).

References

[1] Boček P., Lachout P. (1993). *Linear programming approach to LMS-estimation.* Memorial volume of Comput. Statist. & Data Analysis **19**, 129 – 134.

[2] Bowden R.J., Turkington D.A. (1984). *Instrumental variables.* Cambridge: Cambridge University Press.

[3] Breiman L. (1968). *Probability.* Addison-Wesley Publishing Company, London, 1968.

[4] Čížek P. (2002). *Robust estimation with discrete explanatory variables.* COMPSTAT 2002, Berlin, 509 – 514.

[5] Chung K.L. (1968). *A course in probability theory.* Harcourt, Brace and World, New York.

[6] Hampel F.R., Ronchetti E.M., Rousseeuw P.J., Stahel W.A. (1986). *Robust statistics – the approach based on influence functions.* J. Wiley & Sons, New York.

[7] Hájek J., Šidák Z. (1967). *Theory of rank test.* Academic Press, New York.

[8] Hansen L.P. (1982). *Large sample properties of generalized method of moments estimators.* Econometrica **50**, no 4, 1029 – 1054.

[9] Judge G.G., Griffiths W.E., Hill R.C., Lutkepohl H., Lee T.C. (1985). *The theory and practice of econometrics.* J. Wiley & Sons, New York.

[10] Mašíček L. (2004). *Diagnostics and sensitivity of robust estimators.* Dissertation, Charles University, Prague.

[11] Plát P. (2003). *The least weighted squares.* Diploma theses, the Czech Technical University, Prague, 2003.

[12] Rousseeuw P.J. (1983). *Multivariate estimation with high breakdown point.* Proceedings of the Fourth Pannonian Symposium on Mathematical Statistics, Bad Tatzmannsdorf, Austria, 283 – 297.

[13] Rousseeuw P.J. (1984). *Least median of square regression.* Journal of Amer. Statist. Association **79**, 871 – 880.

[14] Víšek J.Á. (1996). *On high breakdown point estimation.* Computational Statistics, Berlin **11**, 137 – 146.

[15] Víšek J.Á. (1997). *Statistická analýza dat, Statistical data analysis*, (in Czech). Vydavatelství Českého vysokého učení technického v Praze (the Czech Technical University Publishing House), 1997.

[16] Víšek J.Á. (1998). *Robust instruments.* Robust'98 (ed. Jaromír Antoch & Gejza Dohnal), published by Union of Czechoslovak Mathematicians and Physicists, 195 – 224.

[17] Víšek J.Á. (2000). *Regression with high breakdown point.* ROBUST 2000, 324 – 356, ISBN 80-7015-792-5.

[18] Víšek J.Á. (2000). *On the diversity of estimates.* Comput. Statist. and Data Analysis **34**, 67 – 89.

[19] Víšek J.Á. (2000). *A new paradigm of point estimation.* Proceedings of the Seminar: Data Analysis 2000/II, Modern Statistical Methods - Modelling, Regression, Classification and Data Mining, ISBN 80-238-6590-0, 195- 230.

[20] Víšek J.Á. (2002). *Sensitivity analysis of M-estimates of nonlinear regression model: Influence of data subsets.* Annals of the Institute of Statistical Mathematics **54**, No.2, 261- 290.

[21] Víšek J.Á. (2004). *Robustifying instrumental variables.* Proceedings of COMPSTAT'2004. Physica-Verlag/Springer, ISBN 3-7908-1554-3, 1947-1954.

[22] Víšek J.Á. (2004). *The least weighted squares for dynamic specification.* Celebrating Statistics: an International Conference in Honour of Sir David Cox in Occasion of his 80th Birthday.

[23] Víšek J.Á. (2004). *The least weighted squares for panel data.* 6th World Congress of the Bernoulli Society for Mathematical Statistics and Probability, 2004.

[24] Wooldridge J.M. (2001). *Econometric analysis of cross section and panel data.* MIT Press, Cambridge, Massachusetts.

[25] Wooldridge J.M. (2001). *Applications of generalized method of moments estimation.* J. of Economic Perspective **15**, no 4, 87 – 100.

Acknowledgement: Research was supported by grant of GA ČR number 402/03/0084.

Address: J.Á. Víšek, Department of Macroeconomics and Econometrics, Faculty of Social Sciences, Charles University Prague, Smetanovo nábřeží 6, 110 01 Praha 1 & Department of Stochastic Informatics, Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic

E-mail: visek@mbox.fsv.cuni.cz

ODHADY POČTU KOMPONENT MODELU

Petr Volf

Klíčová slova: Řád modelu, BIC, Bayes factor, MCMC.

Abstrakt: Příspěvek se zabývá kriterii BIC a Bayesův faktor, dále pak použitím „Reversible jump" algoritmu Metropolise Hastingse, a nakonec uplatněním zjednodušeného prohledávání a hledání (s pomocí simulovaného žíhání) maxima aposteriorního rozložení, zároveň pro parametry modelu a počet jeho komponent.

1 Úvod

V mnoha případech analýzy dat je jedním z úkolů také výběr optimální složitosti modelu, nejčastěji související se stanovením počtu komponent, řádu modelu, např. v úlohách autoregrese, regresních modelů, i modelování pomocí směsí distribucí. Existuje řada kriterií odvozených exaktně pro určité případy (např. Akaikeho IC), varianty se pak používají i v případech jiných, na základě určité zkušenosti. Samozřejmě vznikla idea použít bayesovský pohled na počet komponent jako na neznámý parametr, na tomto základě lze odvodit BIC G. Schwarze. Mimochodem, tento pohled, kdy řád modelu vyplynul z aposteriorní pravděpodobnosti, je od 70. let uplatňován pro řízení procesů (AR typu) v ÚTIA, původně skupinou kolem Ing. V. Peterky, na kterou plynule navázalo oddělení adaptivních systémů (M. Kárný a další). Začínali s rodinou konjugovaných rozděleni exponenciálního typu. V 90. letech rozvoj MCMC metod umožnil neomezovat se při Bayesově analýze na takovéto typy distribucí a vedl k značnému rozšíření použitelnosti bayesovských metod. Na druhé straně, začlenění řádu modelu do MCMC generování v sobě nese přeskoky mezi parametry (a odpovídajícími prostory) různé dimenze. V praxi sice při použití např. algoritmu Metropolise-Hastingse na to byl zpravidla brán zřetel (viz i P. Volf v Robustu 1998, optimální počet splinů pro neparametrickou regresi), ale bylo třeba i teoreticky zformulovat metody dávající záruky pro reversibilnost skoků v takovém algoritmu. To učinil P. Green [3]. Po počátečním boomu používání MCMC metod nyní snad již nastal stav, kdy se rozlišuje, zda je jejich početně náročné použití účelné. Protože nám jde často pouze o nalezení optimální konfigurace (modelu), je zcela dostatečné a jednodušší používat jen prvky MCMC k prohledávání prostoru a kombinovat je s přímou optimalizací tam, kde to umíme. Důsledné MCMC generování má pak za cíl především dospět k reprezentaci cílových (v našem případě aposteriorních) distribucí.

2 Některé metody určení počtu komponent

Nejčastější situací, kdy se s tímto problémem setkáváme, jsou lineární modely (tj. modely složené lineárně z určitých jednotek, komponent, např. polynomiální regrese, ARMA posloupnosti). Bylo navrženo poměrně dost kritérií, velká část z nich více méně penalizuje vlastní kriterium fitu modelu (nejčastěji maximum věrohodnosti) penaltou závislou na počtu v modelu použitých komponent. Přehled některých kritérií je třeba v T. Cipra, Robust 84 (pro určení řádu ARMA modelů), a poměrně nesystematicky i v různých monpografiích zabývajících se aspekty statistického modelování (např. [4]). My se zde zastavíme u kritérií vycházejících sice z Bayesova přístupu, ale vlastně navržených pro použití v „nebayesovské“ statistice.

2.1 Bayesovské informační kriterium

Představme si situaci, v níž máme data $\boldsymbol{x}$, a sadu možných modelů s různým počtem komponent k, $f(\boldsymbol{x}|\alpha, k), \alpha = \alpha_k$. Uvažujme i příslušné Bayesovo schéma, tj. priory pro α a k: $g_0(\alpha|k)$, $P_0(k)$ na $1, 2, \ldots, K = K_{\max}$.

Pak tedy aposteriorní rozdělení pro α, při daném k je

$$g(\alpha_k|\boldsymbol{x}, k) = f(\boldsymbol{x}|\alpha_k, k) \cdot g_0(\alpha_k|k) \,/\, c_1(\boldsymbol{x}|k), \tag{1}$$

kde $c_1(\boldsymbol{x}|k) = \int_{\alpha_k} f(\boldsymbol{x}|\alpha_k, k) \cdot g_0(\alpha_k|k) \mathrm{d}\alpha_k$.

Bayesův odhad parametru je nejčastěji definován jako maximizer aposteriorního rozdělení (MAP), zde tedy $\hat{\alpha} = \arg\max_\alpha \, g(\alpha|\boldsymbol{x}, k)$, člen $c_1(\boldsymbol{x}|k)$ k jeho určení znát nepotřebujeme. Společný posterior pro α a k je

$$f(\boldsymbol{x}|\alpha_k, k) \cdot g_0(\alpha_k|k) \cdot P_0(k) \,/\, c_2(\boldsymbol{x}), \tag{2}$$

zatímco marginální posterior pro α_k dostaneme z (2) součtem přes $k = 1, \ldots, K_{\max}$. Výraz (2) by byl ten výraz, který bychom měli maximizovat při důsledném bayesovském hledání řešení (k tomu se dostaneme v další části). Zde $c_2(\boldsymbol{x})$ je součet přes všechna k z $c_1(\boldsymbol{x}|k) \cdot P_0(k)$.

Konečně aposteriorní rozdělení pro počet komponent k je

$$P(k|\boldsymbol{x}) \;=\; c_1(\boldsymbol{x}|k) \cdot P_0(k) \,/\, c_2(\boldsymbol{x}), \tag{3}$$

Představme si nyní úlohu najít k maximalizující (3). Přímo to jde opět například v případě normálního rozdělení, s konjugovanými priory (normální pro μ a gamma pro $1/\sigma$). G Schwarz [6] toto řešil trochu obecněji (ale že to není problém, ukazuje jeho článek, který má pouhé 3 strany), s tím, že zároveň uvažoval počet dat $n \to \infty$, a došel ke kriteriu BIC, které také nese jeho jméno. Přesněji, uvažoval náhodný výběr rozsahu n z rozdělení (s neznámým k) exponenciálního typu, tj. $f(x_i|\alpha_k, k) = \exp(\alpha_k \cdot y(x_i) - b(\alpha_k))$, tedy

$$f(\boldsymbol{x}|\alpha_k, k) = \exp(n(Y \cdot \alpha_k - b(\alpha_k)),$$

kde $Y = \sum_{i=1}^{n} y(x_i)/n$ je postačující statistika pro α_k. Dále předpokládal regularitu, včetně toho, že druhé derivace $b(\alpha_k)$ jsou pozitivně definitní, pak tedy existuje (jeden) MVO α_k^* maximizující $Y \cdot \alpha_k - b(\alpha_k)$. Nechť dále počet možných modelů je konečný (tj.$K_{\max} < \infty$) a apriorní hustoty $g_0(\alpha_k|k)$ jsou ohraničené i odražené od 0. Úloha je pak najít v souladu s (3) k maximizující

$$S(Y,n,k) = \log\left\{P_0(k)\int_{\alpha_k}\exp(n(Y\cdot\alpha_k - b(\alpha_k))\cdot g_0(\alpha_k|k)\mathrm{d}\alpha_k\right\}. \qquad (4)$$

Základ důkazu je v Taylorově rozvoji $Y \cdot \alpha_k - b(\alpha_k)$ v okolí α_k^*, kde

$$Y\cdot\alpha_k - b(\alpha_k) \sim Y\cdot\alpha_k^* - b(\alpha_k^*) - a||\alpha_k - \alpha_k^*||^2 + r(Y,k,n),$$

s $a > 0$ a s r vhodně malým podle volby okolí. Dále v (4) dostaneme

$$S(Y,n,k) \sim \log P_0(k) + n(Y\cdot\alpha_k^* - b(\alpha_k^*)) +$$

$$+\log\int_{\alpha_k}\exp(-na||\alpha_k - \alpha_k^*||^2))\cdot g_0(\alpha_k|k)\mathrm{d}\alpha_k + R(Y,k,n),$$

kde $R(Y,k,n)$ jsou již ohraničené (v n). Pak už je třeba jen $\exp(-na||\alpha_k - \alpha_k^*||^2))$ doplnit na normální hustotu vynásobením a vydělením členem $(\frac{na}{\pi})^{k/2}$ a dostaneme, že

$$S(Y,n,k) \sim n(Y\cdot\alpha_k^* - b(\alpha_k^*)) - \frac{k}{2}\log n + R_1(Y,k,n),$$

kde R_1 je opět ohraničené. Takže vidíme výsledný tvar BIC kriteria (tak jak se většinou uvádí), které říká, že je třeba minimalizovat přes k

$$-2\cdot\max_{\alpha} loglik(\boldsymbol{x},k,\alpha) + \log n\cdot k. \qquad (5)$$

2.2 Bayesův faktor

Bayesův faktor je kriterium k porovnání dvou možných modelů a k rozhodnutí, který model preferovat. Z toho hlediska jde vlastně zobecnění pojmu věrohodnostní poměr, a to ve dvojím smyslu: jednak do „bayesovského světa“, jednak zde se nepožaduje, aby druhý model byl rozšířením prvého, což je potřeba pro použití věrohodnostního poměru (vnořené či „vhnízděné“, nested, modely). Zůstaneme u našeho značení, mějme tedy data $\boldsymbol{x}$, dva modely charakterizované čísly k_1, k_2, také jejich apriorní pravděpodobnosti $P_0(k_j)$, a příslušné parametry α_j s jejich obory hodnot a priory $g_0(\alpha_j|k_j)$, $j = 1,2$. Nyní udělejme poměr aposteriorních pravděpodobností obou modelů, z (3), přičemž se funkce $c_2(\boldsymbol{x})$ zkrátí,

$$\frac{P(k_1|\boldsymbol{x})}{P(k_2|\boldsymbol{x})} = \frac{P_0(k_1)}{P_0(k_2)}\cdot\frac{\int_{\alpha_1} f(\boldsymbol{x}|\alpha_1,k_1)\cdot g_0(\alpha_1|k_1)\mathrm{d}\alpha_1}{\int_{\alpha_2} f(\boldsymbol{x}|\alpha_2,k_2)\cdot g_0(\alpha_2|k_2)\mathrm{d}\alpha_2}.$$

Zde integrály v 2. části lze také považovat za pravděpodobnosti (hustoty) $p(\boldsymbol{x}|k_j)$ a právě jejich poměr, tj. kdoví proč bez poměru priorů pro k_j, se označuje pojmem Bayesův faktor [2]:

$$B_{12} = \frac{p(\boldsymbol{x}|k_1)}{p(\boldsymbol{x}|k_2)}.$$

Je pravda (podobně jako u BIC), že při rostoucím rozsahu dat ($n \to \infty$) vliv apriorních pravděpodobností $P_0(k_j)$ klesá, a působnost BF je srovnatelná s BIC. Konkrétně, pokud označíme příslušné MVO α_j^* a $S_{12} = \log f(\boldsymbol{x}|\alpha_1^*, k_1) - \log f(\boldsymbol{x}|\alpha_2^*, k_2) - (k_1 - k_2)\log(n)/2$, tak

$$\frac{S_{12} - \log B_{12}}{\log B_{12}} \longrightarrow 0. \tag{6}$$

Pokud neumíme přímo spočítat $p(\boldsymbol{x}|k)$, tak jednou cestou je generování hodnot $\alpha_{(i)}, i = 1, \ldots, N$, z jejich prioru $g_0(\alpha|k)$, pak je tedy odhad $\hat{p}(\boldsymbol{x}|k) = \sum_i f(\boldsymbol{x}|\alpha_{(i)}, k)/N$.

2.3 Použití standardních statistických testů

Standardní statistické testy jsou samozřejmě také nástrojem sloužícím k nalezení nejvhodnějšího modelu. Pokud zůstaneme ve „věrohodnostních“ modelech splňujících podmínky regularity, tak máme k dispozici především testy významnosti jednotlivých parametrů, založené na asymptotické normalitě jejich MV odhadů. V těchto případech jde skutečně o porovnání vhnízděných modelů, nejčastěji testujeme přidání 1 komponenty k modelu stávajícímu. Místo penalty vyskytující se v předešlých kritériích (a vznikající tam z apriorních rozdělení parametrů) zde podobným způsobem působí kritická hodnota testu.

Zkusme teď porovnat **test pomocí věrohodnostního poměru** s BIC. Mějme model M_0 a model M_1, který je rozšířením M_0 o 1 komponentu. Naše hypotéza H_0 je, že ona komponenta je navíc, že model M_0 je správný. Kriterium pomocí poměru věrohodností je založené na MVO v obou modelech a porovnání dosažených maxim věrohodnostních funkcí, řekněme V_0^* a V_1^*. Konkrétně se užívá

$$LVP_{10} = 2\log\left(\frac{V_1^*}{V_0^*}\right) = 2L_1^* - 2L_0^*,$$

kde $L_j^* = \log V_j^*$. Při platnosti H_0 má tato veličina limitní rozdělení chikvadrát s 1 stupněm volnosti. Takže H_0 zamítneme (pokud zvolíme hladinu testu např. 1%), až pokud $LVP_{10} > \chi^2_{(1)}(0.99) \approx 6.635$.

Kdybychom stejné rozhodování dělali pomocí BIC kriteria, ke stejnému závěru (prefernci M_1 před M_0) bychom došli, pokud by bylo $2L_1^* - 2L_0^* > \log n$. Takže až zhruba pro $n \sim \exp(6.635) \approx 760$ bychom dostali stejnou kritickou hodnotu. Pro menší n by kritická hodnota BIC byla menší, a naopak.

Takže se zdá, že test poměrem věrohodnosti je pro střední rozsahy výběrů dost „ostrý“. Ale musíme vzít v úvahu, že tato kriteria jsou formulována pro dost obecné případy, a teprve podrobná analýza (a simulace) v konkrétních případech ukazují, jak si která kriteria vedou. Podobné porovnání můžeme dostat i pro BF, když z (6) vyvodíme, že

$$2\log B_{10} \sim LVP_{10} - (k_1 - k_0)\log n.$$

O určité vágnosti takto obecně pojatého kritéria svědčí i doporučení obsažené např. v [2] v následující tabulce:

$2\log B_{10}$:	B_{10}:	Evidence against H_0
0 to 2	1 to 3	Not worth than a bare mention
2 to 6	3 to 20	Positive
6 to 10	20 to 150	Strong
> 10	> 150	Very strong

Poznámka: V praxi se při konstrukci modelů (např. regresních modelů z bázových funkcí) osvědčuje takový postup, že nejprve vybereme řád modelu pomocí BIC či podobného kriteria s penalizací, a pak tento model dále zjednodušujeme již statistickými testy významnosti jednotlivých parametrů. Navíc, na základě zkušeností se penalta v BIC často násobí vhodnou konstantou, která může být chápána jako parametr řídící optimalizaci modelu.

3 Maximalizace sdruženého aposteriorního rozdělení

Hledáme nyní takovou konfiguraci α_k a k, která maximalizuje sdružené aposteriorní rozdělení (2), tj. hledáme $arg\ \max_{\alpha_k,k}$ pro

$$f(\boldsymbol{x}|\alpha_k, k) \cdot g_0(\alpha_k|k) \cdot P_0(k). \tag{7}$$

Nejjednodušší případ je, že jsme schopni pro každé k spočítat MAP odhad $\hat{\alpha}_k$, ten dosadit do (7) a porovnat pro různá k. Pokud to neumíme, je třeba se k optimu dostat nějak postupně. Jednou možností je MCMC metoda. O těchto metodách bylo na Robustech již mnoho napsáno (Janžura, i Volf, 1998). Nyní popíšeme, jak by taková vhodná metoda mohla vypadat. Jak jsme již řekli, důsledné použití MCMC procedur (Gibbsův sampler, Metropolis-Hastings algoritmus) vede k generování reprezentace aposteriorního rozdělení, což v tomto případě vlastně nepotřebujeme. Je pravda, že když místo bodového odhadu máme celou reprezentaci, je to značně bohatší informace. Za tu cenu, že generování je většinou časově náročnější, a navíc, každá taková procedura má určité „volné“ parametry (dodávané analytikem).

Poznámka 2: Už i volba apriorní distribuce je vlastně jeden z podstatných do jisté míry „volných parametrů“ celé analýzy. I proto se v rámci bayesovské statistiky rozvíjí téma zvané „bayesovská robustnost“, zkoumající mimo jiné právě citlivost výsledku na vnesené aproirní informaci.

Poznámka 3: Jsou ovšem typické situace, kdy je výhodné mít reprezentaci aposteriorního rozdělení, např. pro predikci dalšího chování systému. Na základě dat a apriorní informace odvozujeme vlastně distribuci možných konfigurací (stavů) systému.

Formálně, MH algoritmus by postupoval tak, že k danému stavu α, k by navrhoval pomocí nějaké zvolené distribuce $q_0(\alpha^*|\alpha, k, k^*)Q(k^*|k)$ nové hodnoty α^*, k^* a přijímal je (jako další člen vytvářené Markovovy posloupnosti) s pravděpodobností $\min(\pi, 1)$, kde

$$\pi = \frac{f(\boldsymbol{x}|\alpha^*, k^*)g_0(\alpha^*|k^*)P_0(k^*)}{f(\boldsymbol{x}|\alpha, k)g_0(\alpha|k)P_0(k)} \cdot \frac{q_0(\alpha|k)Q(k|k^*)}{q_0(\alpha^*|k^*)Q(k^*|k)}.$$

Vidíme, že v posledním zlomku je ve jmenovateli popsán krok, který navrhujeme, zatímco v čitateli je krok opačný, od navrhované konfigurace k původní. A právě tady vzniká problém, jakmile se v takovém kroku mění i k, tj. i dimenze α. Jeden možný způsob, jak tento problém překlenout, navrhl Green [3], viz i Richardson a Green [5].

3.1 RJMH algoritmus

Tato zkratka znamená „Reversible Jump Metropolis-Hastings“ a označuje právě onu Greenem navrženou „trans-dimensional“ variantu MH algoritmu. Představme si zjednodušeně popsanou situaci: Máme jen 2 různé prostory možných konfigurací parametrů, řekněme A, B, na nich apriorní distribuce parametrů $g_A(\alpha), g_B(\beta)$ a taky apriorní pravděpodobnosti $P(A), P(B)$. A také předpokládáme, že lze vždy oba parametry doplnit nějakými veličinami u, v z nějakých prostorů U, V tak, aby (α, u) už bylo stejné dimenze jako (β, v) a bylo možné vybrat vzájemně jednoznačné a diferencovatelné zobrazení, které by zároveň bylo použitelné pro navrhování přechodů, $(\beta, v) = h(\alpha, u)$. Přitom u a v generujeme náhodně a nezávisle na stávající konfiguraci z nějakých rozdělení $r(u), s(v)$.

Poznámka 4: Při MCMC (i jiných prohledávacích) algoritmech jde i o to, aby přechod něco přinesl z hlediska cíle generování, tj. aby se navrhovaly spíš „dobré“ konfigurace, které budou mít větší šanci na přijetí, aby prohledávání prostoru možných konfigurací bylo „inteligentní“ – něco jako v řeči genetických algoritmů, navrhovat i mutace a kombinace dobrých konfigurací. Např. pokud $A \subset R_k$ a $B \subset R_{k+1}$, tak vezmeme nejspíš $U \subset R_1$, $V = \emptyset$. Ale například v případě, kdy parametry mají význam uzlů regresních splinů nebo centrů shluků, spíš než abychom k danému α přigenerovali 1 komponentu, je

z hlediska efektivity hledání řešení lepší jeden z dosavadních centrů rozštěpit na 2, tak, že k němu přičteme a odečteme náhodně navržené u.

Jak tedy nyní vypadá onen poměr navrhovacích pravděpodobností, tj. ta 2. část v přijímací pravděpodobnosti MH algoritmu (7)? Ve jmenovateli je popsán navržený přechod, v čitateli přechod opačný, dostaneme tedy

$$\frac{s(v)}{r(u)}\frac{q_1(\alpha,u|\beta,v)}{q_2(\beta,v|\alpha,u)}\frac{P_{BA}}{P_{AB}},$$

kde P_{AB}, P_{BA} jsou pravděpodobnosti, s jakými ony přechody mezi prostory navrhujeme (tj. zbývající možnosti, při nichž jde o standardni kroky MH algoritmu, jsou setrvání v daném prostoru, které se navrhují s pravděpodobnostmi P_{AA}, a P_{BB}). Ale protože $q_1(\alpha,u|\beta,v) = I[(\beta,v) = h(\alpha,u]/q_3(\alpha,u)$ a obdobně $q_2(\beta,v|\alpha,u) = I[(\beta,v) = h(\alpha,u]/q_4(\beta,v)$, tak prostřední poměr je vlastně jen Jacobián zobrazení h:

$$\frac{q_1}{q_2} = \frac{q_4(\beta,v)}{q_3(\alpha,u)} = |\frac{\partial(\beta,v)}{\partial(\alpha,u)}|.$$

Je dobré připomenout, že pokud netrváme na vygenerování reprezentace aposteriorního rozdělení, je lépe se takovýmto algoritmům vyhnout. V praxi se dále komplikují i tím, že apriorní rozdělení mají také své parametry, které se opět mohou generovat z jejich apriorních rozdělení, která mají zase nějaké parametry ... (ale dále se většinou již nejde, tyto „metaparametry“ se již nějak zvolí a zachází se s nimi také jako s řídícími parametry procedury. To je tzv. hiearchické schema v bayesovské analýze).

3.2 Návrh praktické metody

Praktická metoda by měla mít dvě důležité složky, a to navrhování vhodných konfigurací a jejich vyhodnocování (a přijímání dobrých konfigurací). Navrhování by mělo mít ony vlastnosti zmíněné shora (pamatovat si dobrá řešení), ale zároveň by mělo občas přijít s odskokem do jiné oblasti konfigurací (mělo by vytvářet nerozložitelný řetězec). Přijímání by pak mělo používat postup typu simulované žíhání. Většinou jsme v situaci, že část parametrů můžeme pro každou navrženou konfiguraci spočítat přímo (zde tedy jako Bayesův odhad), část musíme hledat. Například v regresním modelu s regresními spliny lineární koeficienty odhadneme přímo, uzly splinů získáme prohledávám (viz i Volf, 1998). Postupovat budeme tedy nejspíš tak, že navržené konfigurace budeme přijímat s pravděpodobností $\min(\pi(m), 1)$, kde nyní m je počet provedených iterací (resp. skupin iterací, sweeps) a

$$\pi(m) \;=\; \{\frac{f(\boldsymbol{x}|\alpha^*,k^*)g_0(\alpha^*|k^*)P_0(k^*)}{f(\boldsymbol{x}|\alpha,k)g_0(\alpha|k)P_0(k)}\}^{1/temp(m)},$$

kde $temp(m) \to 0$ vhodnou rychlostí, řekněme o něco rychleji než $1/\log m$. O zkušenostech s tímto postupem viz Janžura, Robusty 1990, 2004.

4 Závěr, poznámky o shlukové analýze

Tento příspěvek měl být původně o shlukové analýze, či o souvisejících modelech směsí distribucí. Ale pro nedostatek místa odkazuji na některý z budoucích článků. Směsový model shlukové analýzy je

$$x_i \sim \sum_{j=1}^{k} \pi_j f_j(x_i|\alpha_j),$$

kde k je pro nás neznámý počet shluků, π jsou jejich váhy, a máme n dat $\boldsymbol{x} = (x_1, \ldots, x_n)$. Praktičtější je použít tento popis: nechť $\boldsymbol{z} = (z_1, \ldots, z_n)$ jsou indikátory příslušnosti bodu x_i do shluku j, tj. $z_i \in \{1, \ldots, k\}$. Úkolem shlukové analýzy je odhadnout parametry α_j, indikátory $\boldsymbol{z}$ a také počet shluků k. Při daném k je úloha poměrně snadno početně řešitelná modifikací EM algoritmu: Při daných z_i, tj daném zařazení dat x_i k jednotlivým distribucím, spočteme MVO parametrů α_j, a naopak, při daných parametrech, tj. známých distribucích $f_j(.|\alpha_j)$, volíme $z_i = argmax_j\, f_j(x_i|\alpha_j)$. Ovšem problém určení počtu komponent není uspokojivě řešen. Metody popsané v 1. části této práce se samozřejmě používají, ale protože tato úloha nyní nesplňuje podmínky regularity, nelze je považovat za konzistentní. Metody náhodného prohledávání jsou v tomto případě tedy vhodnou možností. O mixture models existuje několik monografií a mnoho článků, ale upozorňuji na přehled výsledků (i vyzkoušení RJMH postupu pro jednoduchý příklad shlukování) v diplomové práci J. Němečka [1].

Reference

[1] Němeček J. (2004). *Klasifikace a rozpoznávání.* Diplomová práce KPMS MFF UK Praha.

[2] Kass R.E., Raftery A.E. (1995). *Bayes factors.* J.A.S.A. **90**, 773 – 795.

[3] Green P.J. (1995). *Reversible jump MCMC computation and Bayesian model determination.* Biometrika **82**, 711 – 732.

[4] Ripley B.D. (1996). *Pattern recognition and neural networks.* Cambridge Univ. Press.

[5] Richardson S., Green P.J. (1997). *On Bayesian analysis of mixtures with an unknown numbers of components* (with discussion). J.R.S.S., ser. B **59**, 731 – 792.

[6] Schwarz G. (1978). *Estimating the dimensionof a model.* Annals Statist **6**, 461 – 464.

Poděkování: Práce byla podpořena grantem GA ČR č. 201/02/0049.

Adresa: P. Volf, ÚTIA AV ČR, Pod vodárenskou věží 4, 182 08 Praha 8

E-mail: volf@utia.cas.cz

KONFIDENČNÉ INTERVALY PRE EFEKT OŠETRENIA V KLINICKÝCH POKUSOCH

Gejza Wimmer, Viktor Witkovský

Kľúčové slová: Meta-analýza klinických pokusov, zmiešaný lineárny model, metóda Kenwarda-Rogera.

Abstrakt: V príspevku je navrhnutá metóda konštrukcie konfidenčného intervalu pre spoločný efekt ošetrenia, ktorý je hlavným parametrom záujmu v klinických štúdiach, resp. v meta-analýze klinických štúdii, ktoré sú založené na nezávislých pokusoch v k zdravotníckych zariadeniach, alebo klinických štúdiach. Metóda je založená na spojitej normálnej aproximácii rozdelenia výberových pravdepodobností v jednotlivých zariadeniach, ktorá vedie k modelu jednoduchého triedenia s náhodným efektom vplyvu zdravotníckeho zariadenia a s nerovnakými rozptylmi vo vnútri jednotlivých zariadení. Takýto model sa používa tiež na modelovanie medzilaboratórnych porovnávacích štúdii a je špeciálnym prípadom zmiešaného lineárneho modelu. Na konštrukciu približného $(1-\alpha)$-konfidenčného intervalu pre parameter spoločného efektu ošetrenia využívame metódu navrhnutú v práci [6] a pre prípad medzilaboratórnych porovnávacích štúdii podrobne študovanú v práci [14] pre prípad spoločnej strednej hodnoty v medzilaboratórnych porovnávacích štúdiach. Simulačná štúdia ukazuje vlastnosti (resp. ohraničenia) navrhnutého konfidenčného intervalu.

1 Úvod

Uvažujme určitý klinický pokus, napr. liečbu istej choroby, špecifikovaný operačný zákrok, podávanie určitého lieku, a pod. Ten istý pokus sa realizuje v k zdravotníckych zariadeniach (k nemocniciach, vykonáva ho k lekárov, podáva sa liek k homogénnym skupinám pacientov, atď.). Medzi základné otázky lekárov, farmaceutov, riadiacich pracovníkov v zdravotníctve, resp. v zdravotných poisťovniach, výrobcov liekov, pacientov, je: *„Aká je skutočná úspešnosť ošetrenia (operačného zákroku, terapie liekom, atď.)."* Z matematicko-štatistického pohľadu je to problém návrhu „dostatočne dobrého", čo možno najjednoduchšieho modelu vyššie spomenutých klinických pokusov a návrh vhodného bodového a intervalového odhadu úspešnosti ošetrenia v klinických pokusoch.

Klinické pokusy sa ukazujú akousi diskrétnou paralelou medzilaboratórnych kľúčových porovnávaní, (pozri napr. [2], [7], [3], [9], [8], [4], [10], [14], [12], [13]).

V tomto príspevku využijeme výsledky dosiahnuté v modeloch medzilaboratórnych kľúčových porovnávaní (hlavne aplikáciu postupu navrhnutú

v práci [6]) pre získanie odpovedí na vyššie uvedené matematicko-štatistické problémy. Výsledky dosiahnuté v modeloch medzilaboratórnych kľúčových porovnávaní modifikujeme na spojitú aproximáciu diskrétneho modelu klinických pokusov.

Špecifikovaný klinický pokus sa realizuje v k zdravotníckych zariadeniach. Počet realizovaných pokusov v i-tom zariadení je n_i, $i = 1, \ldots, k$. Nech pravdepodobnosť úspechu (v pokuse) v i-tom zariadení je p_i, $i = 1, \ldots, k$. Všetky pokusy v jednom zdravotnom zariadení považujeme za navzájom nezávislé. Označme X_i náhodnú premennú – počet úspešných pokusov v i-tom zdravotníckom zariadení. Preto $X_i \sim Bi(n_i, p_i)$, $i = 1, \ldots, k$. ($X_i \sim Bi(n_i, p_i)$ znamená, že X_i má binomické rozdelenie s parametrami n_i a p_i). $X_1, \ldots, X_k$ považujeme za stochasticky nezávislé náhodné premenné. V ďalšom budeme pracovať s náhodnými premennými Y_i, $i = 1, \ldots, k$, pričom $Y_i = \frac{X_i}{n_i}$. Dostávame teda model

$$Y_i = \frac{X_i}{n_i}, \quad i = 1, \ldots, k. \tag{1}$$

2 Model klinických pokusov a bodový odhad úspešnosti ošetrenia

Predpokladajme najskôr, že vo všetkých uvažovaných zdravotníckych zariadeniach v danom klinickom pokuse je na ošetrenie (liečbu) použitá rovnaká metóda, ktorá je aplikovaná na skupinu „homogénnych“ pacientov. V centre nášho záujmu je v tomto prípade hodnota pravdepodobnosti úspešnosti daného ošetrenia (liečby), ktorá, ako predpokladáme, je rovnaká v každom zdravotníckom zariadení. V praxi však možno často pozorovať porušenie tohto prepokladu. Skutočná pravdepodobnosť úspešnej liečby v i-tom zariadení, p_i, môže náhodne kolísať okolo spoločnej pravdepodobnosti úspešnej liečby, povedzme p, ktorá je v centre nášho záujmu.

Na modelovanie tejto skutočnosti budeme uvažovať „spojitú aproximáciu“ modelu (1), normálnym modelom jednoduchého triedenia s náhodnými efektmi. Presnejšie, budeme predpokladať, že pre dostatočne veľké počty n_i jednotlivých ošetrení (pokusov) v zdravotníckych zariadeniach platí (aspoň približne) model

$$Y_i = p + b_i + \varepsilon_i, \quad i = 1, \ldots, k, \tag{2}$$

kde p reprezentuje spoločnú pravdepodobnosť úspešnej liečby, b_i je náhodný efekt i-tého zdravotníckeho zariadenia. Teda pravdepodobnosť úspešnej ličeby v i-tom zariadení je v tomto modeli $p_i = p + b_i$. Napokon ε_i je nezávislá aditívna chyba, pričom predpokladáme, že pre $i = 1, \ldots, k$ je $b_i \sim N(0, \sigma_0^2)$, $\varepsilon_i \sim N(0, \sigma_i^2/n_i)$ a teda

$$Y_i \sim N(p, \sigma_0^2 + \sigma_i^2/n_i), \quad i = 1, \ldots, k.$$

Náhodný (observačný) vektor $Y = (Y_1, \ldots, Y_k)'$ modelujeme teda zmiešaným lineárnym modelom

$$Y \sim N\left(\mathit{1}_{k,1}p, \Sigma_{k,k} = \sum_{i=0}^{k} \sigma_i^2 G_i\right). \tag{3}$$

Teda, observačný vektor Y je normálne rozdelený so strednou hodnotou $\mathcal{E}(Y) = \mathit{1}p$ a kovariančnou maticou $cov(Y) = \Sigma_{k,k} = \sum_{i=0}^{k} \sigma_i^2 G_i$, pričom $G_0 = I_{k,k}$ a $G_i = \text{ diag}(0, \ldots, 0, \frac{1}{n_i}, 0, \ldots, 0)$, $\mathit{1} = (1, \ldots, 1)'$. Ak by sme poznali variančné komponenty σ_0^2 a σ_i^2, $i = 1, \ldots, k$, optimálnym odhadom úspešnosti liečby p (spoločnej pravdepodobnosti úspešnej liečby) by bol

$$\hat{p} = (\mathit{1}'\Sigma^{-1}\mathit{1})^{-1}\mathit{1}'\Sigma^{-1}Y. \tag{4}$$

Z (1) vyplýva, že disperzia

$$Var(Y_i|p_i) = \frac{p_i(1-p_i)}{n_i},$$

odkiaľ dostávame $\sigma_i^2 = p_i(1-p_i)$, $i = 1, \ldots, k$. Ako prirodzený odhad parametra σ_i^2 budeme uvažovať odhad $\tilde{\sigma}_i^2 = \tilde{p}_i(1-\tilde{p}_i)$, kde za odhad pravdepodobnosti úspešnosti liečby v i-tom zariadení voľme odhad navrhnutý v práci [1], a síce $\tilde{p}_i = \frac{X_i+2}{n_i+4}$, ktorý, napriek svojej jednoduchosti, má mnohé dobré štatistické vlastnosti už pre stredné rozsahy výberov z binomického rozdelenia. Dostávame

$$\tilde{\sigma}_i^2 = \frac{X_i+2}{n_i+4}\left(1 - \frac{X_i+2}{n_i+4}\right).$$

Ako vhodný odhad parametra σ_0^2 sa núka odhad $\tilde{\sigma}_0^2$ navrhnutý v práci [7]. Získa sa iteratívne z rovníc

$$\hat{\mu}^{(MP)} = \frac{\sum_{i=1}^{k} \frac{X_i}{n_i\tilde{\sigma}_0^2+\tilde{\sigma}_i^2}}{\sum_{i=1}^{k} \frac{n_i}{n_i\tilde{\sigma}_0^2+\tilde{\sigma}_i^2}},$$

$$\sum_{i=1}^{k} \frac{(X_i - n_i\hat{\mu}^{(MP)})^2}{n_i\tilde{\sigma}_0^2 + \tilde{\sigma}_i^2} = k - 1.$$

(Poznamenávame len, že ľavá strana v poslednej rovnici je s pravdepodobnosťou jedna monotónne klesajúca funkcia, pozri [7], [9] a [5].) Ak v (4) nahradíme skutočné (neznáme) hodnoty σ_0^2 a σ_i^2, $i = 1, \ldots, k$ odhadmi $\tilde{\sigma}_0^2$ a $\tilde{\sigma}_i^2$, $i = 1, \ldots, k$, dostávame bodový odhad úspešnosti ošetrenia

$$\tilde{p} = \frac{\sum_{i=1}^{k} \frac{X_i}{n_i\tilde{\sigma}_0^2+\tilde{\sigma}_i^2}}{\sum_{i=1}^{k} \frac{n_i}{n_i\tilde{\sigma}_0^2+\tilde{\sigma}_i^2}}. \tag{5}$$

3 Intervalový odhad úspešnosti ošetrenia

Keď v modeli (3) poznáme σ_0^2 a σ_i^2, $i = 1, \ldots, k$, tak disperzia odhadu $\hat{p}$ z (4) je

$$\Phi(\sigma_0^2, \sigma_1^2, \ldots, \sigma_k^2) = (1'\Sigma^{-1}1)^{-1}.$$

Kenward & Roger [6] navrhujú korigovaný odhad $\hat{\Phi}_A$ disperzie Φ_A a aproximáciu náhodnej premennej

$$F = \lambda \frac{(\tilde{p} - p)^2}{\hat{\Phi}_A} \tag{6}$$

$F_{1,m}$ rozdelením s 1 a m stupňami voľnosti. Postupujúc úplne analogicky ako Kenward & Roger [6] dostávame

$$\begin{aligned}
P_0 &= -\sum_{i=1}^{k} \left(\frac{n_i}{n_i\sigma_0^2 + \sigma_i^2} \right)^2, \\
P_i &= -\frac{n_i}{(n_i\sigma_0^2 + \sigma_i^2)^2}, \quad i = 1, \ldots, k, \\
Q_{0,0} &= \sum_{i=1}^{k} \left(\frac{n_i}{n_i\sigma_0^2 + \sigma_i^2} \right)^3, \\
Q_{0,i} &= Q_{i,0} = \frac{n_i^2}{(n_i\sigma_0^2 + \sigma_i^2)^3}, \quad i = 1, \ldots, k, \\
Q_{ij} &= \begin{cases} 0, & i \neq j, \\ \dfrac{n_i}{(n_i\sigma_0^2 + \sigma_i^2)^3}, & i = j, \end{cases}, \quad i = 1, \ldots, k.
\end{aligned}$$

Ako kovariančnú maticu odhadov variančných komponentov $\tilde{\sigma}_0^2, \tilde{\sigma}_1^2, \ldots, \tilde{\sigma}_k^2$, použjeme $I_F^{-1}(\sigma_0^2, \sigma_1^2, \ldots, \sigma_k^2)$ - inverziu Fisherovej informačnej matice REML (Restricted Maximum Likelihood) odhadov variančných komponentov σ_0^2, σ_1^2, $\ldots, \sigma_k^2$. Tento krok podporuje aj výsledok z práce [8], podľa ktorého možno Mandel-Pauleho odhady za platnosti predpokladu o normalite rozdelenia Y považovať za zjednodušenú verziu REML odhadov.

Prvky matice I_F sú dané vzťahmi:

$$\begin{aligned}
\{I_F\}_{i,j} &= \frac{1}{2}\left[tr\left(\frac{\partial\Sigma^{-1}}{\partial\sigma_i^2}\Sigma\frac{\partial\Sigma^{-1}}{\partial\sigma_j^2}\Sigma \right) - tr(2\Phi Q_{ij} - \Phi P_i \Phi P_j) \right] \\
&= \frac{1}{2}\left[\{S\}_{ij} - \{R\}_{ij} \right], \quad i, j = 0, 1, \ldots, k,
\end{aligned}$$

pričom prvky matíc R a S sú:

$$
\begin{aligned}
\{R\}_{ij} &= \Phi(2Q_{ij} - P_i\Phi P_j), \quad i,j = 0,1,\ldots,k,\\
\{S\}_{0,0} &= -P_0 = \sum_{i=1}^{k}\left(\frac{n_i}{n_i\sigma_0^2+\sigma_i^2}\right)^2,\\
\{S\}_{0,i} &= \{S\}_{i,0} = -P_i = \frac{n_i}{(n_i\sigma_0^2+\sigma_i^2)^2}, \quad i = 1,\ldots,k,\\
\{S\}_{ij} &= \begin{cases} 0, & i \neq j,\\ \dfrac{1}{\sigma_i^4}\left(n_i - \dfrac{2\sigma_0^2 n_i}{n_i\sigma_0^2+\sigma_i^2} + \dfrac{n_i\sigma_0^4}{(n_i\sigma_0^2+\sigma_i^2)^2}\right), & i = j, \end{cases}, \; i = 1,\ldots,k,
\end{aligned}
$$

pozri tiež [11] a [14].

Teda kovariančná matica W odhadov variančných komponentov je

$$W = W(\sigma_0^2,\sigma_1^2,\ldots,\sigma_k^2) = I_F^{-1}(\sigma_0^2,\sigma_1^2,\ldots,\sigma_k^2).$$

Podľa Kenward & Roger[6] ďalej platí

$$\hat{\Phi}_A = \hat{\Phi} + 2\hat{\Lambda}, \tag{7}$$

kde

$$\hat{\Lambda} = \hat{\Phi}^2 \sum_{i=0}^{k}\sum_{i=0}^{k} \hat{W}_{ij}(\hat{Q}_{ij} - \hat{P}_i\hat{\Phi}\hat{P}_j),$$

pričom $\hat{\Phi}$, $\hat{W}_{ij}$, $\hat{Q}_{ij}$ a $\hat{P}_i$, sú odhady Φ, W_{ij}, Q_{ij} a P_i, $i,j = 0,\ldots,k$, ktoré dostaneme nahradením neznámych variančných komponentov $\sigma_0^2,\sigma_1^2,\ldots,\sigma_k^2$ vo výrazoch Φ, W_{ij}, Q_{ij} a P_i, $i,j = 0,\ldots,k$, ich odhadmi $\tilde{\sigma}_0^2,\tilde{\sigma}_1^2,\ldots,\tilde{\sigma}_k^2$.

Ďalej

$$\lambda = 1, \quad m = \frac{2}{\hat{\Phi}^2(\hat{P}'\hat{W}\hat{P})}, \tag{8}$$

kde $\hat{P} = (\hat{P}_0,\hat{P}_1,\ldots,\hat{P}_k)'$.

Hľadaný $(1-\alpha)$-konfidenčný interval pre úspešnosť ošetrenia p je

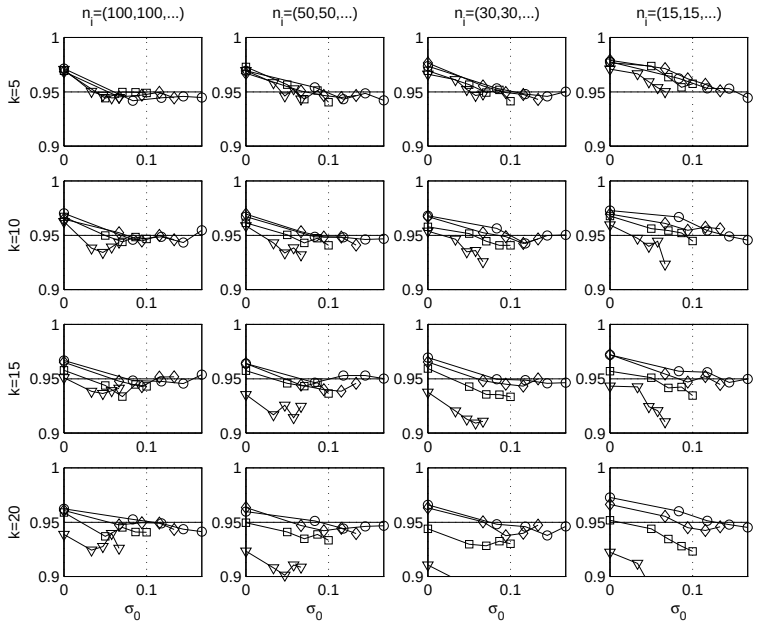
$$\left\langle \tilde{p} - \sqrt{\hat{\Phi}_A F_{1,m}(\alpha)}, \tilde{p} + \sqrt{\hat{\Phi}_A F_{1,m}(\alpha)} \right\rangle \cap \langle 0,1\rangle, \tag{9}$$

pričom $F_{1,m}(\alpha)$ je α-kritická hodnota $F_{1,m}$ rozdelenia, $\tilde{p}$ je dané v (5), $\hat{\Phi}_A$ je dané v (7) a m je dané v (8).

4 Simulačná štúdia

V simulačnej štúdii sme overovali empirické pravdepodobnosti pokrytia úspešnosti ošetrenia p $(1-\alpha)$-konfidenčným intervalom (9) pre rôzne hodnoty parametrov k, n_i, $i = 1,\ldots,k$, p a σ_0^2 modelu (3).

Počet zdravotníckych zariadení $k \in \{5,10,15,20\}$.



Obrázok 1: Empirické pravdepodobnosti pokrytia úspešnosti ošetrenia p 95%-konfidenčným intervalom (9) pre rôzne hodnoty parametrov k, n_i, $i = 1, \ldots, k$, p a σ_0^2 vo *vyváženom* modeli (3). Symbol $\bigtriangledown$ označuje návrhy experimentu s parametrom $p = 0.2$, $\square$ označuje návrhy experimentu s parametrom $p = 0.3$, $\diamond$ označuje návrhy experimentu s parametrom $p = 0.4$, $\bigcirc$ označuje návrhy experimentu s parametrom $p = 0.5$.

Počet pokusov v jednotlivých zariadeniach pri vyvážených pokusoch: $n_i = 100$, $i = 1, \ldots, k$; $n_i = 50$, $i = 1, \ldots, k$; $n_i = 30$, $i = 1, \ldots, k$; $n_i = 15$, $i = 1, \ldots, k$. Počet pokusov v jednotlivých zariadeniach pri nevyvážených pokusoch: $n_i = \{15, 50, 15, 50, \ldots\}$; $n_i = \{30, 100, 30, 100, \ldots\}$; $n_i = \{15, 100, 15, 100, \ldots\}$; $n_i = \{15, 30, 50, 100, 15, 30, 50, 100, \ldots\}$.

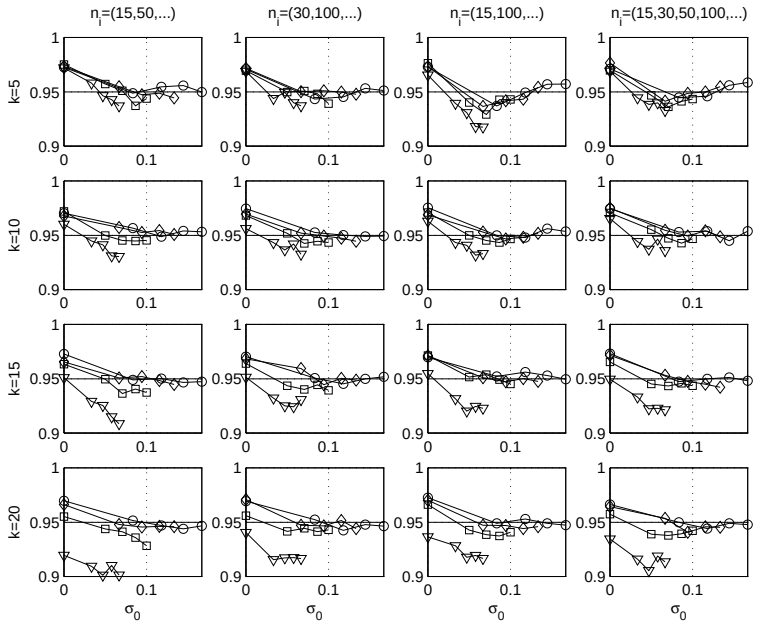
Spoločná pravdepodobnosť úspešnosti liečby $p \in \{0.2, 0.3, 0.4, 0.5\}$.

Disperzia náhodných efektov zdravotníckych zariadení $\sigma_0^2 \in \{0, \frac{1}{4}(\frac{p}{3})^2, \frac{1}{2}(\frac{p}{3})^2, \frac{3}{4}(\frac{p}{3})^2, (\frac{p}{3})^2\}$, kde p je spoločná pravdepodobnosť úspešnosti liečby v danej sérii pokusov.

Za nominálnu hladinu významnosti α sme zvolili $\alpha = 0.05$. Pre každú kombináciu parametrov sme realizovali 5000 opakovaní (série) pokusov.

5 Diskusia

V práci sa navrhla spojitá normálna aproximácia modelu 1) efektu ošetrenia v klinických pokusoch zmiešaným lineárnym modelom (3).



Obrázok 2: Empirické pravdepodobnosti pokrytia úspešnosti ošetrenia p 95%-konfidenčným intervalom (9) pre rôzne hodnoty parametrov k, n_i, $i = 1, \ldots, k$, p a σ_0^2 v *nevyváženom* modeli (3). Symbol $\bigtriangledown$ označuje návrhy experimentu s parametrom $p = 0.2$, $\square$ označuje návrhy experimentu s parametrom $p = 0.3$, $\diamond$ označuje návrhy experimentu s parametrom $p = 0.4$, $\bigcirc$ označuje návrhy experimentu s parametrom $p = 0.5$.

Využijúc ďalej poznatky získané pri modelovaní medzilaboratórnych porovnávacích štúdií sa navrhol v (5) bodový odhad spoločnej úspešnosti ošetrenia p a intervalový odhad (9) založený na metóde Kenwarda a Rogera.

Simulačná štúdia ukázala, že pre $0.2 < p < 0.8$, pri všetkých návrhoch (ktoré sa blížili reálne sa vyskytujúcim situáciam v praxi), je empirická hladina významnosti navrhnutého konfidenčného intervalu (9) veľmi blízka nominálnej hodnote, čo ukazuje na to, že uvažovaná spojitá aproximácia modelu klinických pokusov a navrhnuté odhady (bodový a intervalový) spoločnej úspešnosti ošetrenia sú vhodné pre praktické použitie.

Reference

[1] Agresti A., Caffo B. (2000). *Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures.* The American Statistician **54**, 280 – 288.

[2] Cochran W.G. (1937). *Problems arising in the analysis of a series of similar experiments.* Journal of the Royal Statistical Society, Supplement **4**, 102 – 118.

[3] DerSimonian R., Laird N.M. (1982). *Meta-analysis in clinical trials.* Controlled Clinical Trials **7**, 177 – 188.

[4] Hartung J., Böckenhoff A., Knapp G. (2003). *Generalized Cochran-Wald statistics in combining of experiments.* Journal of Statistical Planning and Inference **113**, 215 – 237.

[5] Iyer H.K., Wang C.M., Mathew T. (2002). *Models and confidence intervals for true values in interlaboratory trials.* Submitted for publication.

[6] Kenward M.G., Roger J.H. (1997). *Small sample inference for fixed effects from restricted maximum likelihood.* Biometrics **53**, 983 – 997.

[7] Paule R.C., Mandel J. (1982). *Consensus values and weighting factors.* Journal of Research of the National Bureau of Standards **87**, 377 – 385.

[8] Rukhin A.L., Biggerstaff B.J., Vangel M.G. (2000). *Restricted maximum likelihood estimation of a common mean and the Mandel-Paule algorithm.* Journal of Statistical Planning and Inference **83**, 319 – 330.

[9] Rukhin A.L., Vangel M.G. (1998). *Estimation of a common mean and weighted means statistics.* Journal of the American Statistical Association **93**, 303 – 308.

[10] Savin A., Wimmer G., Witkovský V. (2003). *On Kenward-Roger confidence intervals for common mean in interlaboratory trials.* Measurement Science Review **3**, Section 1, 53 – 56. `http://www.measurement.sk/`.

[11] Searle S.R., Casella G., McCulloch C.E. (1992). *Variance components.* J. Wiley & Sons, New York.

[12] Wimmer G., Witkovský, V. (2003). *Between group variance component interval estimation for the unbalanced heteroscedastic one-way random effects model.* Journal of Statistical Computation and Simulation **73**, 333 – 346.

[13] Wimmer G., Witkovský V. (2003). *Consensus mean and interval estimators for the common mean.* Tatra Mountains Mathematical Publications **26**, 183 – 194.

[14] Witkovský V., Savin A., Wimmer G. (2003). *On small sample inference for common mean in heteroscedastic one-way model.* Discussiones Mathematicae: Probability and Statistics **23**, 123 – 145.

Poďakovanie: Podporené grantom VEGA 1/0264/03 a VEGA 2/4026/04.

Adresa: G. Wimmer, Inštitút matematiky a informatiky, Banská Bystrica, Slovenská republika, a Přírodovědecká fakulta, Masarykova univerzita, Brno, Česká republika
V. Witkovský, Ústav merania SAV, Dúbravská cesta 9, 841 04 Bratislava, Slovenská republika

E-mail: `wimmer@mat.savba.sk, witkovsky@savba.sk`

GRAFICKÉ MODELY V ANALÝZE FINANČNÍCH DAT

Jitka Zichová

Klíčová slova: Grafický model, podmíněná nezávislost.

Abstrakt: Grafické modely jsou jedním z nástrojů mnohorozměrné statistické analýzy. Umožňují popis a přehledné znázornění struktury vzájemných závislostí v dané množině proměnných. V poslední době se uplatňují i v oblasti financí, o čemž svědčí například publikace [1], [2], [3]. Článek shrnuje některé aplikace zpracované s použitím českých i zahraničních finančních dat diplomanty oboru Finanční a pojistná matematika na MFF UK v Praze pod vedením autorky příspěvku.

1 Grafický model

Uvažujme sloupcový náhodný vektor $X = (X_1, X_2, \ldots, X_k)^T$, indexovou množinu $K = \{1, 2, \ldots, k\}$ a graf $G = (K, E)$, v němž množina vrcholů je K a E označuje množinu hran. Nechť chybějící hrana (i, j) indikuje podmíněnou nezávislost náhodných veličin X_i a X_j při pevných hodnotách ostatních složek vektoru X, což značíme $X_i \perp X_j | \{X_r; r \neq i, j\}$. Znamená to, že pro podmíněné hustoty veličin $X_i,\ X_j$ a vektoru $(X_i, X_j)^T$ platí

$$f_{X_i, X_j | \{X_r; r \neq i,j\}} = f_{X_i | \{X_r; r \neq i,j\}} f_{X_j | \{X_r; r \neq i,j\}}.$$

Nechť $K = A \cup B \cup C$. Označme X_a podvektor vektoru X obsahující složky X_i, $i \in A$ a analogicky podvektory X_b a X_c se složkami s indexy z B respektive z C. Množina vrcholů C separuje množiny A a B, když všechny cesty z některého vrcholu $i \in A$ do některého vrcholu $j \in B$ obsahují alespoň jeden vrchol z C. Separaci interpretujeme tak, že náhodné vektory X_a a X_b jsou podmíněně nezávislé při pevné hodnotě vektoru X_c, t.j. $X_a \perp X_b | X_c$.

Úplný graf má všechny dvojice vrcholů spojené hranou. Klika je maximální úplný podgraf, jejím rozšířením o další vrcholy vznikne podgraf, který již není úplný. Řetězový graf má vrcholy uspořádané do bloků, takže $K = b_1 \cup b_2 \cup \cdots \cup b_m$ pro nějaké přirozené $m < k$. Nechť $r(j)$ je index bloku obsahujícího vrchol j. Na množině vrcholů existuje částečné uspořádání definované předpisem $i < j$ když $r(i) < r(j)$, $i \leq j$ když $r(i) = r(j)$. Hrany spojující vrcholy z téhož bloku jsou neorientované zatímco hrany, jež spojují vrcholy z různých bloků, jsou orientované od bloku s nižším indexem k bloku s indexem vyšším. Nechť $K(j) = b_1 \cup b_2 \cup \cdots \cup b_{r(j)}$. Chybějící hrana (i, j), $i \leq j$ znamená, že $X_i \perp X_j | \{X_r; r \in K(j), r \neq i, j\}$.

Grafický model s grafem G je systém pravděpodobnostních rozdělení náhodného vektoru X splňujících podmíněné nezávislosti dané grafem G. Speciálním případem je saturovaný model s úplným grafem.

V praxi se používají systémy normálních rozdělení pro analýzu spojitých dat a systémy rozdělení určených mnohorozměrnou kontingenční tabulkou pro zpracování dat diskrétních. Zkoumání podmíněných nezávislostí v množině proměnných umožňují modely s neorientovanými grafy. Chceme-li vyšetřovat příčinné souvislosti, to jest vztahy mezi soubory závisle a nezávisle proměnných, používáme modely s řetězovými grafy.

2 Selekce modelu

Předpokládejme nadále, že máme k dispozici data ve formě n realizací k-rozměrného náhodného vektoru X. Naším cílem je popsat strukturu podmíněných nezávislostí složek vektoru X vhodným grafickým modelem. K tomu účelu byly vypracovány různé selekční algoritmy v rámci věrohodnostního a bayesovského přístupu.

Omezíme-li se na věrohodnostní přístup, je základním nástrojem selekčních algoritmů deviance. Pro grafický model s grafem G ji definujeme předpisem $\mathrm{dev}(G) = 2(l_S - l_G)$, kde l_S je maximum logaritmické věrohodnostní funkce v saturovaném modelu a l_G je maximum logaritmické věrohodnostní funkce v modelu s grafem G. Deviance má asymptoticky chí-kvadrát rozdělení, počet stupňů volnosti f závisí na rozdělení dat a zmíníme jej později. Je testovou statistikou pro test modelu s grafem G proti alternativě saturovaného modelu.

Selekční algoritmy pracují v krocích spočívajících v postupném ubírání hran počínaje saturovaným modelem s úplným grafem (typ backward) nebo naopak v postupném přidávání hran počínaje grafem bez hran (typ forward). Zřejmě je tedy třeba umět testovat model s grafem G_2 proti alternativě modelu s grafem G_1 obsahujícím oproti G_2 navíc jednu nebo více hran. Testovou statistikou je v takových případech diference deviancí $\mathrm{dev}(G_2) - \mathrm{dev}(G_1)$ s asymptotickým chí-kvadrát rozdělením o $f_2 - f_1$ stupních volnosti, kde f_2 jsou stupně volnosti pro $\mathrm{dev}(G_2)$ a f_1 jsou stupně volnosti pro $\mathrm{dev}(G_1)$. Překročí-li deviance respektive diference deviancí kritickou hodnotu příslušného chí-kvadrát rozdělení, zamítáme testovaný model ve prospěch alternativního modelu s grafem s více hranami. Podrobný popis selekčních algoritmů nalezneme v knize [4] a v citovaých diplomových pracích.

3 Gaussovské grafické modely

Předpokládejme, že náhodný vektor X má mnohorozměrné normální rozdělení s nulovou střední hodnotou a varianční maticí V. Označme $D = V^{-1}$ inverzní varianční matici a $d_{ij}, i, j = 1, 2, \ldots, k$ její prvky. Lze dokázat, že $X_i \perp X_j | \{X_r; r \neq i, j\}$ právě tehdy, když $d_{ij} = 0$. Deviance modelu s grafem G má tvar

$$\mathrm{dev}(G) = n\{tr(S\hat{D}) - \ln[\det(S\hat{D})] - k\},$$

kde $\hat{D} = \hat{V}^{-1}$ a $\hat{V}$ je maximálně věrohodný odhad varianční matice V v modelu s grafem G. Tento odhad se počítá iteračně aplikací tzv. IPF algoritmu

(Iterative Proportional Fitting), který je popsán např. v [4]. Výběrová varianční matice S je maximálně věrohodným odhadem pro V v saturovaném modelu. Počet stupňů volnosti pro chí-kvadrát rozdělení deviance modelu s grafem G je roven počtu chybějících hran v G.

Následující příklad byl řešen v diplomové práci [6] s pomocí programu napsaného autorem práce v systému Mathematica.

Příklad 1. Analýza vzájemných vztahů českých burzovních indexů.

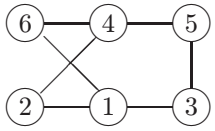
Databáze byla tvořena časovými řadami měsíčních pozorování uzávěrkových kursů odvětvových indexů Burzy cenných papírů Praha z let 1994-2001. Zaměřili jsme se na šestici odvětví, a to výroba nápojů a tabáku (X_1), textilní průmysl (X_2), hutnictví (X_3), elektroprůmysl (X_4), služby (X_5) a investiční fondy (X_6). Data byla transformována diferencemi logaritmů, které splnily předpoklady normality a nezávislosti pozorovaných realizací.

Podívejme se nejprve na korelační matici indexů pro sledovaná odvětví.

Nápoje Textil Hutnictví Elektro Služby Fondy

1	**0.33**	**0.43**	0.37	0.33	**0.42**
	1	0.31	**0.48**	0.33	0.38
		1	0.31	**0.41**	0.32
			1	**0.42**	**0.44**
				1	0.35
					1

Naprogramovaný backward algoritmus vybral pro popis vzájemných souvislostí v datech graf



Korelace dvojic odvětví spojených v grafu hranami jsou vytištěny tučně. Z grafu lze číst, že chování indexu investičních fondů X_6 výrazně ovlivňují z uvažovaných odvětví výroba nápojů a tabáku X_1 a elektroprůmysl X_4, čemuž odpovídají dvě nejvyšší korelace v posledním sloupci korelační matice. U normálně rozdělených dat však hrany v grafu znamenají nenulové hodnoty parciálních korelací. Vidíme například, že vrcholy 2 a 6 nejsou spojeny hranou, tudíž proměnné X_2 a X_6 mají nevýznamnou parciální korelaci. Jejich relativně vysoká korelace 0.38 je způsobena vlivem ostatních proměnných.

V Příkladu 2 řešeném v práci [8] ukážeme aplikaci modelu s řetězovým grafem na data podobného charakteru.

Příklad 2. Chování indexu PX50 v závislosti na odvětvových indexech.

Opět máme k dispozici časové řady odvětvových burzovních indexů pro vybraná odvětví a navíc řadu hodnot průřezového indexu PX50 za stejné období. Byla sledována odvětví výroba nápojů a tabáku (X_1), textilní průmysl (X_2), chemický průmysl (X_3), elektroprůmysl (X_4), energetika (X_5), doprava a spoje (X_6), služby (X_7), sklářský průmysl (X_8), investiční fondy (X_9) a index ostatních odvětví (X_{10}). Závisle proměnnou (Y) je index PX50. Stejně jako v Příkladu 1 byly zpracovány diference logaritmů všech proměnných, a to třístupňovým algoritmem pro selekci řetězového grafu popsaným v knize [4] a realizovaným diplomantkou v systému Mathematica.

Uveďme nejprve parciální korelace indexu PX50 s oborovými indexy. Statisticky významné parciální korelace na pětiprocentní hladině jsou vytištěny tučně.

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
0.15	-0.11	0.14	0.13	**0.78**	**0.87**	0.09	0.13	**0.31**	**0.21**

Blok nezávisle proměnných b_1 je tvořen veličinami $X_1, X_2, \ldots, X_{10}$ a blok b_2 obsahuje jedinou závisle proměnnou Y reprezentující index PX50. V grafickém modelu pro popis dat nás zajímají především hrany spojující vrchol 11 veličiny Y s vrcholy z bloku b_1. Selekční algoritmus navrhl model s grafem obsahujícím orientované hrany pro dvojice $(X_5, Y), (X_6, Y), (X_9, Y), (X_{10}, Y)$. To znamená, že index PX50 je významně ovlivňován indexy energetiky, dopravy a spojů, investičních fondů a indexem ostatních odvětví. Odpovídá to statisticky významným parciálním korelacím. Ze získaného výsledku lze vyjít například při modelování regresní závislosti PX50 na odvětvových indexech. Vybraný grafický model nám oproti regresnímu modelování poskytuje navíc informaci o vzájemných souvislostech v množině nezávisle proměnných, a to prostřednictvím neorientovaných hran spojujících vrcholy z bloku b_1. Podgraf pro blok b_1 lze popsat maticí sousednosti, která má 1 na místě (i, j), spojuje-li vrcholy příslušející proměnným (X_i, X_j) hrana, a 0 v případě nepřítomnosti hrany.

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
0	0	**1**	0	0	**1**	0	**1**	**1**	0
	0	0	**1**	0	0	0	0	**1**	0
		0	0	1	0	1	0	0	0
			0	0	0	**1**	**1**	0	**1**
				0	1	0	0	**1**	0
					0	0	0	0	1
						0	0	0	0
							0	0	0
								0	**1**
									0

Největší provázanost s ostatními odvětvími vykazují indexy výroby nápojů a tabáku X_1, elektroprůmyslu X_4 a investičních fondů X_9, jak ukazují tučně vytištěné 1 ve výše uvedené matici.

Další příklad byl řešen v práci [5] a byl věnován studiu závislosti mezi několika bloky proměnných.

Příklad 3. Analýza odvětvových indexů a indexu IBIX prostřednictvím blokové struktury.

Odvětvové indexy byly rozděleny do bloků b_1, b_2, b_3, u nichž lze usuzovat, že proměnné z bloku s nižším indexem mohou ovlivňovat chování proměnných z bloků s vyšším indexem. Vstup představovaly časové řady diferencí logaritmů denních pozorování indexů z let 1993-1994. Blok b_1 obsahoval indexy zemědělství (X_1), dřevozpracujícího průmyslu (X_2), chemického průmyslu (X_3) a hutnictví (X_4). V bloku b_2 byla zastoupena odvětví potravinářství (X_5), textilní průmysl (X_6), stavebnictví (X_7) a strojírenský průmysl (X_8). Blok b_3 zahrnoval elektroprůmysl (X_9) a obchod (X_{10}). Jediná proměnná X_{11} v bloku b_4 reprezentovala průřezový index IBIX, jenž byl sestavován Investiční a Poštovní bankou. Autor práce naprogramoval zobecněný třístupňový algoritmus pro selekci grafického modelu s řetězovým grafem zahrnujícím více bloků proměnných. Tímto algoritmem byl pro vyšetřovaná data navržen graf, jehož strukturu zde opět naznačíme v maticové formě.

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}
0	0	1	1	*	0	*	0	0	0	0
	0	1	1	*	*	*	0	0	0	0
		0	1	*	0	*	*	0	0	*
			0	*	*	0	*	*	0	0
				0	1	1	1	0	0	*
					0	1	1	*	0	0
						0	1	0	*	*
							0	*	0	*
								0	0	0
									0	*
										0

Hvězdička na místě (i, j) představuje orientovanou hranu vedoucí z vrcholu i do vrcholu j v bloku s vyšším indexem, jednička na místě (i, j) pak neorientovanou hranu spojující dva vrcholy téhož bloku. Například vrchol 5 proměnné potravinářství je spojen orientovanými hranami s vrcholy 1, 2, 3, 4 všech odvětví bloku b_1 a neorientovanými hranami s vrcholy 6, 7 a 8. Vysoká

provázanost s ostatními odvětvími způsobuje i vliv proměnné potravinářství X_5 na index IBIX. Vrchol 10 proměnné obchod není spojen s žádným z vrcholů bloku b_1. Vidíme, že index obchodu je prostřednictvím orientované hrany ovlivněn pouze chováním indexu stavebnictví s vrcholem 7 v bloku b_2. Dále lze konstatovat, že hodnota indexu IBIX je ovlivněna indexy chemického, potravinářského a strojírenského průmyslu, stavebnictví a obchodu, a že všechny čtyři bloky proměnných spolu souvisejí. Největší počet orientovaných hran je mezi bloky b_1 a b_2.

4 Grafické modely pro kategoriální data

Nechť nyní $X = (X_1, X_2, \ldots, X_k)^T$ představuje náhodný vektor měřených znaků na určitém subjektu, přičemž i-tý znak nabývá hodnot $0, 1, 2, \ldots, r_i$, $i = 1, 2, \ldots, k$. Označíme-li symbolem x konkrétní kombinaci sledovaných k znaků, je rozdělení vektoru X dáno k-rozměrnou tabulkou pravděpodobností $P(X = x)$ všech možných kombinací. Databáze je v tomto případě tvořena n subjekty, z nichž každý je popsán k znaky. Četnost kombinace x v datech označíme $n(x)$, přičemž $\sum_x n(x) = n$. Deviance modelu s grafem G je

$$\mathrm{dev}(G) = 2 \sum_x n(x) \ln \frac{n(x)}{n\hat{p}(x)},$$

kde $\hat{p}(x)$ je maximálně věrohodný odhad pravděpodobnosti $p(x)$ v modelu s grafem G a relativní četnost $n(x)/n$ je maximálně věrohodný odhad pro $p(x)$ v saturovaném modelu. Odhady $\hat{p}(x)$ se opět počítají iteračně pomocí IPF algoritmu.

Logaritmicko-lineární rozvoj hustoty lze psát ve tvaru

$$\ln p(x) = \sum_{a \subset K} u_a(x_a),$$

kde sčítáme přes všechny podmnožiny a množiny vrcholů K a $u_a(x_a)$ jsou tzv. u-členy, pro něž platí $u_a(x_a) = u_a(x_i; i \in a)$ a $u_a(x_a) = 0$, existuje-li takové $i \in a$, že $x_i = 0$. Počet stupňů volnosti pro devianci je roven počtu chybějících u-členů s nenulovými argumenty v logaritmicko-lineárním rozvoji $p(x)$, neboť $X_i \perp X_j | \{X_r; r \neq i, j\}$ právě tehdy, když $u_a(x_a) = 0$ pro všechna $a \subset K$ taková, že $i, j \in a$.

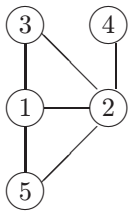
V práci [7] byl řešen problém z oblasti credit scoringu, to jest posuzování bonity žadatelů o úvěry. K dispozici byla databáze klientů jisté německé banky z doby před zavedením Eura. Základním sledovaným znakem je to, zda klientovi byl či nebyl bankou poskytnut úvěr, dalšími znaky je například pohlaví klienta, výše požadovaného úvěru apod.

Příklad 4. Stanovení faktorů ovlivňujících přidělení úvěru.

Uvažujme následující kategoriální proměnné zaznamenávané u žadatelů o bankovní úvěr: úvěr (X_1) nabývající hodnot 0 (neposkytnut) a 1 (poskytnut), výše úvěru (X_2) s hodnotami 0 (<1500 DEM), 1 (1500 až 5000 DEM)

a 2 (>5000 DEM), úspory (X_3) s hodnotami 0 (<100 DEM), 1 (100 až 1000 DEM) a 2 (>1000 DEM), pohlaví (X_4), kde 0 kóduje muže a 1 ženu, a jiný úvěr (X_5) s hodnotami 0 (ano) a 1 (ne). Poslední proměnná indikuje, zda žadatel již má přidělen jiný úvěr.

Selekční algoritmus naprogramovaný diplomantkou v Mathematice navrhl model s grafem



Graf nás informuje, o tom, že přidělení úvěru X_1 ovlivňují kromě pohlaví X_4 všechny sledované znaky. Vrcholy 1, 2, a 3 odpovídající znakům úvěr, výše úvěru, úspory tvoří kliku dokumentující vzájemnou provázanost v této trojici proměnných. Další klika je tvořena vrcholy 1, 2, a 5, jež představují znaky úvěr, výše úvěru a jiný úvěr. Pohlaví X_4 souvisí pouze s požadovanou výší úvěru X_2. Podíváme-li se na procentní podíl žen žádajících o úvěr v dané databázi, zjistíme, že se skutečně liší podle výše úvěru:

< 1500 DEM	*1500 až 5000 DEM*	*>5000 DEM*
54 procent	36 procent	30 procent

Grafické modely v databázích uvedeného typu mohou například poskytnout bankám informaci o tom, které znaky je důležité u klientů evidovat a které nikoli.

Reference

[1] Giudici P. (2001). *Bayesian data mining with application to benchmarking and credit scoring.* Applied Stochastic Models in Business and Industry **17**, 69 – 81.

[2] Hand D.J., Mc Conway K.J., Stanghellini E. (1997). *Graphical models of applicants for credit.* IMA Journal of Mathematics Applied in Business and Industry **8**, 143 – 155.

[3] Stanghellini E., Mc Conway K.J., Hand D.J. (1999). *A discrete variable chain graph for applicants for credit.* Applied Statistics **48**, Part 2, 239 – 251.

[4] Whittaker J. (1990). *Graphical models in applied multivariate statistics.* Wiley, New York.

[5] Ambrož Z. (2004). *Regresní modely pro analýzu výnosu portfolia.* Diplomová práce, KPMS MFF UK, Praha.

[6] Chýna V. (2002). *Grafické modely pro analýzu spojitých finančních dat.* Diplomová práce, KPMS MFF UK, Praha.
[7] Svobodová B. (2003). *Analýza kategoriálních finančních dat.* Diplomová práce, KPMS MFF UK, Praha.
[8] Zelinková J. (2003). *Regrese a grafické modely pro finanční analýzu.* Diplomová práce, KPMS MFF UK, Praha.

Poděkování: Tato práce je podporována výzkumným záměrem MSM 113200008.

Adresa: J. Zichová, KPMS MFF UK, Sokolovská 83, 186 75 Praha 8
E-mail: zichova@karlin.mff.cuni.cz

ROBUSTNOST V MODELU RŮSTOVÝCH KŘIVEK

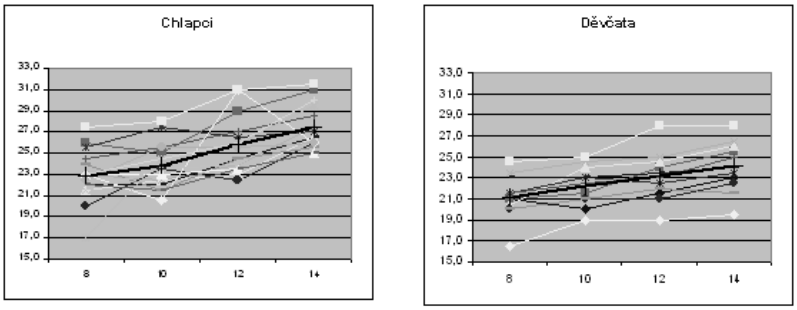
Ivan Žežula, Daniel Klein

Klíčová slova: Mnohorozměrný lineární model, simulace.

Abstrakt: Příspěvek se zabývá dvěmi variantami modelu růstových křivek – replikovaným modelem a modelem se speciálními variančními strukturami – a zkoumá chování odhadů parametrů v případě porušení předpokladů, zejména normality rozdělení.

1 Motivace a popis modelu

Model růstových křivek vznikl při analýze anatomických dat: na jisté zubní klinice byl sledován vývoj vzdálenosti středu hypofýzy od pterygomaxilární brázdy u chlapců a děvčat. Otázkou bylo, zda tato vzdálenost je u děvčat menší než u chlapců a zda rychlost růstu je stejná. Získaná data jsou na obrázku. Silnou čarou je vyznačen průměr; jeho průběh ukazuje, že růst má zhruba lineární trend.



Na první pohled to vypadá jako dvě nezávislé regrese. Potthoff a Roy si však uvědomili, že oba soubory – vzhledem k okolnostem vzniku – mají stejnou varianční matici. Proto navrhli společný model, který tuto skutečnost zohledňoval:

$$\mathrm{E}\, Y = XBZ \qquad \mathrm{var}(\mathrm{vec}\, Y) = \Sigma \otimes I$$

Tento model přirozeným způsobem spojuje regresní analýzu s analýzou rozptylu – matice X je maticí analýzy rozptylu, Z maticí regresních konstant. Neznámé parametry jsou v maticích B (regresní koeficienty pro jednotlivé skupiny) a Σ (variance a kovariance pro jednotlivé časy pozorování). Později se tento model dočkal mnoha dalších aplikací i rozšíření. Z těchto aplikací vyplynuly různé struktury varianční matice Σ:

- obecná p.d. matice
- $\Sigma = \sum \theta_i V_i$, V_i známé
- rovnoměrná varianční struktura, t.j.

$$\Sigma = \sigma^2 \left((1-\rho) I + \rho \mathbf{1}\mathbf{1}' \right)$$

- seriální varianční struktura, t.j.

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho & \dots & \rho^{p-1} \\ \rho & 1 & \dots & \rho^{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{p-1} & \rho^{p-2} & \dots & 1 \end{pmatrix}$$

Všimněme si, že rozklad $\Sigma = \sum \theta_i V_i$ je zcela obecný a zahrnuje všechny ostatní případy. Např. obecnou matici Σ lze napsat ve tvaru $\Sigma = \sum_i \sigma_{ii} e_i e_i' + \sum_{i<j} \sigma_{ij} \left(e_i e_j' + e_j e_i' \right)$, kde e_i jsou jednotkové vektory (t.j. obsahující 1 na i-tém místě a 0 jinde).

Při praktickém použití modelu se však setkáváme s některými problémy. K odhadu B totiž potřebujeme znalost Σ:

$$\hat{B} = \left(X'X \right)^{-1} X'Y\Sigma^{-1}Z' \left(Z\Sigma^{-1}Z' \right)^{-1} .$$

To přináší dva okruhy otázek: Jednak, jak odhadnout Σ resp. její komponenty? A následně, jak se změní jeho vlastnosti odhadu B, jestliže Σ odhadneme? Specielně, o kolik vzroste jeho rozptyl?

V tomto směru již bylo dosaženo mnoha výsledků. Především je známo, že v obecné situaci, při rozkladu Σ na varianční komponenty, existují stejnoměrně nejlepší odhady jen v triviálních případech. Tedy ve většině případů můžeme použít buď jen maximálně věrohodné odhady (pro něž většinou neexistují explicitní vzorce, počítají se pouze numericky a známe jen asymptotické vlastnosti) anebo lokálně nejlepší odhady (u nichž sice známe explicitní vzorce, ale zase musíme dobře vědět, v kterém bodě parametrického prostoru je počítat). Stejnoměrně nejlepší odhad existuje v jednom důležitém případě: když Σ je zcela neznámá. Částečná znalost varianční struktury situaci komplikuje. Lokálně nejlepší odhady variančních komponent θ mohou být navíc závislé na hodnotě B – dochází pak ke kruhové závislosti. Proto důležitosti nabývají odhady invariantní, které na B nezávisí; i pro ně jsou známé explicitní vzorce, odhady však nemusí vždy existovat. V případě odhadování pouze σ^2 a ρ známe nestranné odhadové rovnice pro oba parametry. Ve všech těchto případech je odhad B asymptoticky nestranný. Podrobněji o těchto modelech viz např. [1], [3].

2 Replikovaný model

Tento model byl vytvořen k oslabení závislosti lokálně nejlepších odhadů na volbě počátečního bodu. Obsahuje nezávislá opakování měření ze základního modelu:

$$Y_j = XBZ + e_j, j = 1, \ldots, s$$

Varianční struktura přitom zůstává stejná.

Odhad B je také prakticky stejný, jako v základním modelu:

$$\hat{B} = (X'X)^{-1} X'\bar{Y}\Sigma^{-1}Z' \left(Z\Sigma^{-1}Z'\right)^{-1} .$$

Odhady variančních komponent v tomto modelu vždy existují. Výsledky z tohoto modelu jsou shrnuty v [4]. Dřívější simulace (viz [5]) ukázaly rychlou konvergenci odhadů k známému asymptotickému rozdělení.

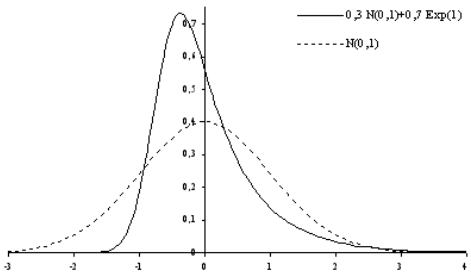
Nás zajímal problém robustnosti odhadů v tomto modelu. Všechny známé výsledky jsou totiž založeny na předpokladu normality. Je tedy přirozené se ptát, co se stane, když rozdělení chyb není normální? Podobnou otázkou je: co se stane s odhady, když chyby nemají nulovou střední hodnotu, t.j. když pozorování obsahují systematickou chybu?

K tomuto účelu jsme provedli rozsáhlou simulační studii. Uvažovali jsme přitom různě složité modely střední hodnoty (polynomy 1. až 3. stupně) a různě složité modely varianční struktury:

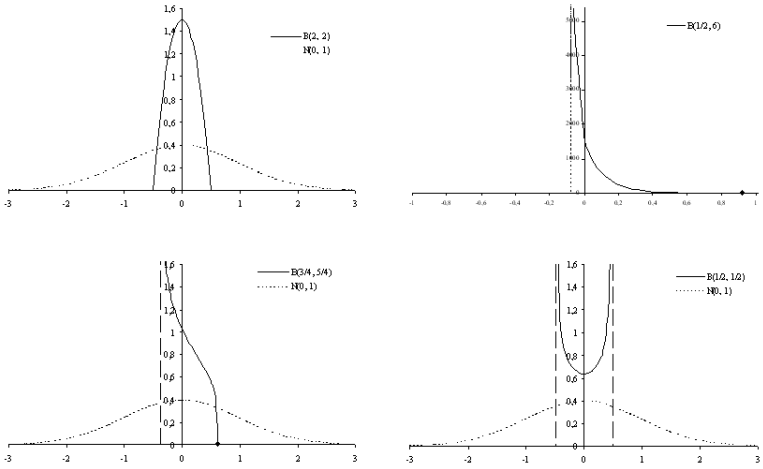
1. Σ obecná p.d. matice
2. Σ s konstantními diagonálami
3. Σ se dvěmi komponentami ($\Sigma = \theta_1 V + \theta_2 I$)

Uvažovaná chybová rozdělení byla následující:

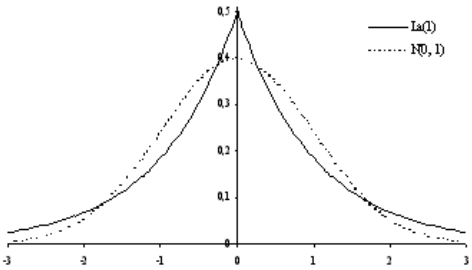
1. normální, ale se systematickou chybou ($\mu \neq 0$)
2. směs normálního a exponenciálního



3. různé formy beta rozdělení



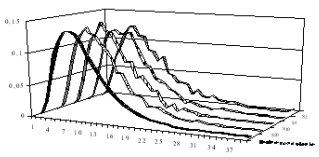
4. Laplaceovo rozdělení



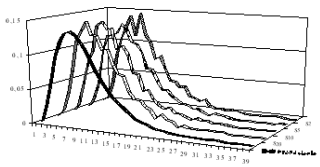
Všechna rozdělení s výjimkou 1) byla posunuta tak, aby měla nulovou střední hodnotu.

Výsledky simulací:

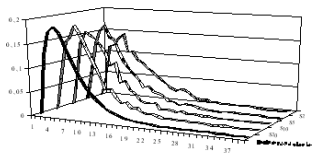
Následující vybrané grafy zobrazují empirické rozdělení statistiky χ^2 (která měří celkovou vzdálenost od odhadu $\widehat{\widehat{B}}$ k jeho skutečné hodnotě) pro různé hodnoty s. Počet replikací roste zezadu dopředu, zcela vpředu je asymptotické rozdělení.



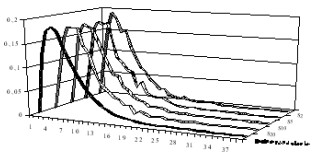
Kvadratická závislost, Σ obecná, normální rozdělení



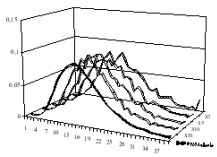
Kvadratická závislost, Σ se 2 komponentami, normální rozdělení



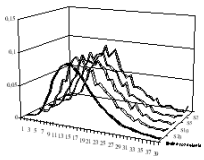
Lineární závislost, Σ obecná, rozdělení B(0.5,6)



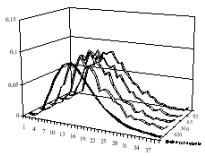
Lineární závislost, Σ se 2 komponentami, rozdělení B(0.5,6)



Kubická závislost, Σ obecná, rozdělení B(2,2)



Kubická závislost, Σ s konstatními diagonálami, rozdělení B(2,2)



Kubická závislost, Σ se 2 komponentami, rozdělení B(2,2)

V případě nenulovosti střední hodnoty chyby – t.j. existence systematické chyby – grafy neuvádíme, jelikož odhady ležely daleko od skutečných hodnot a s rostoucím počtem replikací divergovaly do nekonečna. Lze tedy udělat následující závěry:

- Model je silně citlivý na přítomnost systematické chyby
- Ve všech ostatních případech je konvergence k asymptotickému rozdělení velmi rychlá
- Rychlost konvergence je nepřímo úměrná počtu odhadovaných parametrů

3 Speciální varianční struktury

Dále jsme zkoumali robustnost základního modelu s rovnoměrnou a seriální strukturou (t.j. bez replikací). Odhady $\hat{\sigma}^2$ a $\hat{\rho}$ pro oba modely jsou uvedeny v [6]; porovnej také s [2]. V obou modelech nás především zajímalo, jak se projeví různá chybová rozdělení na střední kvadratické chybě (MSE) odhadu sledovaných parametrů σ^2 a ρ.

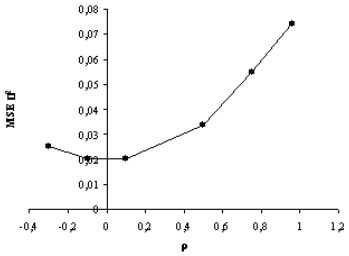
V modelu s rovnoměrnou strukturou lze pro námi uvažované odhady odvodit, že platí:

$$\begin{aligned} \mathrm{MSE}\,\hat{\sigma}^2 &= \mathrm{var}\,\hat{\sigma}^2 = \frac{2\sigma^4}{n-r(X)} \cdot \frac{1+(p-1)\rho^2}{p} \\ \mathrm{MSE}\,\hat{\rho} &= \frac{2}{n-r(X)} \cdot \frac{(1-\rho)^2(1+(p-1)\rho)^2}{p\,(p-1)} + o\left(n^{-1}\right) \end{aligned}$$

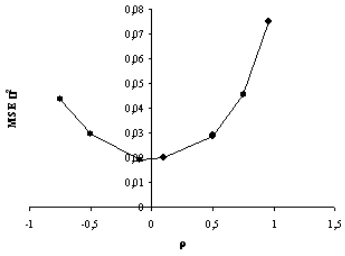
Přirozenou otázkou bylo, jestli lze odhad MSE, který vznikne dosazením odhadnutých hodnot $\hat{\sigma}^2$ a $\hat{\rho}$ do tohoto teoretického vzorce, je prakticky použitelný, t.j. jestli se příliš neliší od skutečnosti.

V obou modelech simulace ukázaly, že odhady parametrů σ^2 i ρ jsou dostatečně robustní a téměř nezávisí na chybovém rozdělení. Největší vliv na MSE těchto odhadů měla hodnota ρ. Zjistili jsme, že

1. chyba odhadu $\hat{\sigma}^2$ roste s $|\rho|$

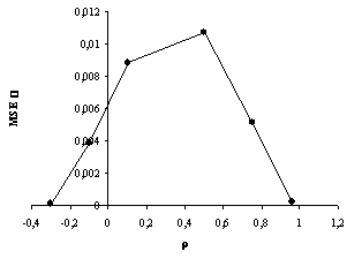


Model s rovnoměrnou strukturou

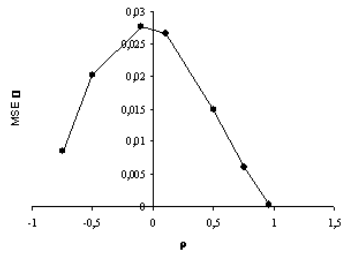


Model se seriální strukturou

2. chyba odhadu $\hat{\rho}$ klesá s $|\rho|$



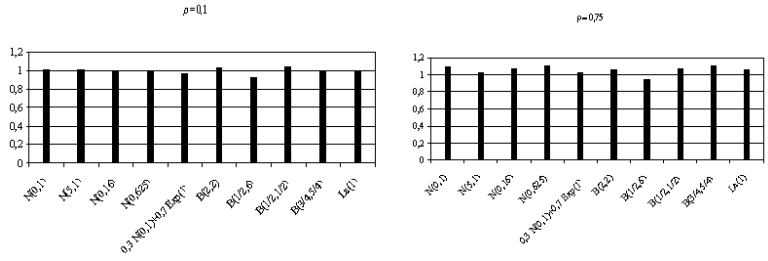
Model s rovnoměrnou strukturou



Model se seriální strukturou

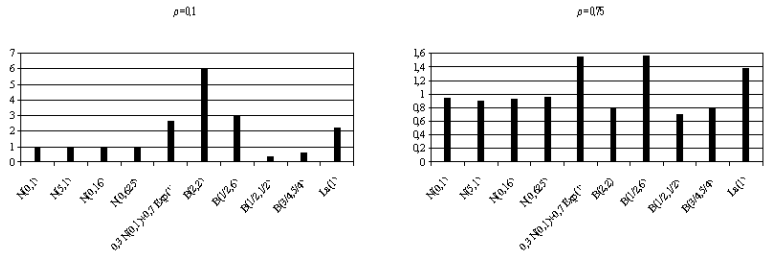
Trochu jiná situace je u aproximace MSE pomocí odhadnutých hodnot parametrů. Zde jsme zjistili, že

1. aproximace MSE $\hat{\rho}$ je robustní



ρ malé ρ velké

2. aproximace MSE $\hat{\sigma}^2$ je citlivá na změnu rozdělení, hlavně pro malá ρ



ρ malé ρ velké

Reference

[1] von Rosen D. (1989). *Maximum likelihood estimators in multivariate linear normal models.* Journal of Multivariate Analysis **31**, 187 – 200.

[2] Lee J. C. (1988). *Prediction and estimation of growth curves with special covariance structures.* JASA **83**, No. 402, 432 – 440.

[3] Žežula I. (1993). *Covariance components estimation in the growth curve model.* Statistics **24**, 321 – 330.

[4] Žežula I. (1997). *Asymptotic properties of the growth curve model with covariance components.* Applications of Mathematics **42**, No. 1, 57 – 69.

[5] Žežula I. (1996). *Simulation study in the growth curve model.* Tatra Mountains Mathematical Publications **7**, 183 – 188.

[6] Žežula I. (2004). *Special variance structures in the growth curve model.* To appear.

Adresa: I. Žežula, D. Klein, UMV PF UPJŠ, Jesenná 5, 041 54 Košice, SR

E-mail: `zezula@kosice.upjs.sk, klein@science.upjs.sk`

PŘÍSPĚVEK K ANALÝZE ROZDĚLENÍ PŘÍJMŮ DOMÁCNOSTÍ V ČR

Jitka Bartošová

Klíčová slova: Příjmové rozdělení, teoretický model, test odlehlosti.

Abstrakt: Pro správné ohodnocení příjmové stránky životní úrovně obyvatelstva i pro správné rozhodování ohledně opatření v této oblasti je nezbytné znát úplné rozdělení příjmů daného období, tj. znát obsazení ve všech příjmových skupinách. Vzhledem k probíhající transformaci hospodářství z plánované formy na tržní dochází ke změnám ve složení příjmů obyvatelstva. Aktuálním úkolem současnosti je ověření platnosti dosud používaného statistického modelu rozdělení ročních příjmů domácností v ČR. Tento příspěvek se zabývá ověřováním platnosti dosud používaného statistického modelu rozdělení příjmů domácností získaných z výběrového šetření ČSÚ Mikrocensus 1996.

1 Důvody zkoumání příjmových rozdělení

Modely příjmových rozdělení umožňují zhodnocení životní úrovně všech obyvatel státu bez rozdílu, stejně jako srovnání životní úrovně příslušníků různých společenských skupin nebo obyvatel různých regionů. Jsou rovněž ukazatelem relativní životní úrovně obyvatelstva vybraného státu ve srovnání s dalšími státy. Pro správnou kvantifikaci té složky životní úrovně obyvatelstva, která přímo závisí na příjmech, je potřeba vystihnout úroveň, strukturu a vývojový trend příjmů obyvatelstva komplexně, tj. nalézt vhodné statistické modely příjmových rozdělení pro jednotlivé sociální skupiny i pro obyvatelstvo jako celek, bez ohledu na sociální skupinu.

2 Statistický model příjmových rozdělení

2.1 Volba teoretického modelu

Základním úkolem při konstrukci statického modelu rozdělení ročních příjmů domácností je nalezení takové teoretické distribuční funkce, která maximálně odpovídá empirickému rozdělení četností. Dosud používaným statistickým modelem příjmových rozdělení bylo logaritmicko – normální rozdělení se dvěma (popřípadě se třemi parametry) $LN(\mu, \sigma^2)$ (popřípadě $LN(\mu, \sigma^2, \gamma)$, kde parametr γ je teoretické minimum náhodné veličiny X). Logaritmicko - normální rozdělení (především jeho varianta se třemi parametry) zatím představovalo dobrou aproximaci příjmových rozdělení pro většinu sociálních skupin. Různost zdrojů, ze kterých příjmy pocházejí, a současný proces diferenciace mezd, který probíhá v některých skupinách velmi bouřlivě, může mít za následek jednak nesourodost příjmových rozdělení jednotlivých sociálních skupin a jednak vysokou variabilitu uvnitř těchto skupin. Empirické

rozdělení četností příjmů v některých sociálních skupinách by proto mohlo být lépe vystiženo některým jiným modelem (např. normálním, Weibullovým, nebo Γ-rozdělením atd.). Odchylky empirického rozdělení ročních příjmů domácností od předpokládaného teoretického modelu mohou být zapříčiněny rovněž přítomností odlehlých hodnot, popřípadě heterogenitou dat, odpovídající např. směsi několika navzájem posunutých logaritmicko - normálních křivek, popřípadě směsi logaritmicko - normálního rozdělení s některým jiným rozdělením, např. normálním apod. S tímto problémem se můžeme setkat nejenom u rozdělení příjmů všech obyvatel bez rozdílu sociální skupiny, ale i u rozdělení příjmů v některých jednotlivých sociálních skupinách.

Testování shody empirické a teoretické distribuční funkce $F_E(x)$ a $F_T(x)$ můžeme provádět buď početně, pomocí testovacích statistik, nebo graficky. Často používanou početní metodou pro testování nulové hypotézy o shodě výběrového příjmového rozdělení s předpokládaným teoretickým modelem je χ^2-test dobré shody ([1]), který pracuje se statistikou $\chi^2 = \sum_{i=1}^{k} \frac{(n_i - n\pi_i)^2}{n\pi_i}$, kde n_i a $n\pi_i$ (π_i) jsou absolutní empirické a absolutní (relativní) teoretické četnosti, k je počet tříd, n je rozsah výběru. Dalšími měrami shody empirického a teoretického rozdělení jsou např. suma čtverců (popřípadě suma absolutních hodnot) odchylek empirického rozdělení od teoretického, reprezentovaná statistikou $MSE = \sum_{i=1}^{k}(p_i - \pi_i)^2$ (popřípadě $MAE = \sum_{i=1}^{k} |p_i - \pi_i|$), kde $p_i = \frac{n_i}{n}$ a π_i jsou relativní četnosti empirického a teoretického rozdělení.

Na hodnoty empirické distribuční funkce $F_E(x)$, a tedy i na hodnoty uvedené testovací statistiky χ^2, má vliv volba velikosti třídních intervalů h při seskupování dat, která určuje počet tříd $k \approx \frac{x_{max} - x_{min}}{h}$. Odhad počtu tříd podle Sturgessova pravidla, který je dán vztahem $\hat{k} \approx 1 + 3{,}3 \log_{10} n$, kde n je rozsah výběru, je vhodný pro pouze menší výběry. Pro velké výběry je toto dělení příliš „hrubé“. V takovém případě je vhodnější použít k odhadu šířky intervalů h a tedy i k určení počtu tříd k Scottovo pravidlo ([10]), popřípadě robustní Freedmanovo-Diaconisovo pravidlo.

2.2 Identifikace odlehlých hodnot

Problém vzrůstu variability příjmů celkem i uvnitř jednotlivých sociálních skupin je způsoben především vznikem skupin obyvatel s extrémně nízkými popřípadě extrémně vysokými příjmy. Tyto hodnoty příjmů, které můžeme z hlediska zvoleného modelu považovat za odlehlá pozorování, způsobují narušení vybraného teoretického modelu a snižují jeho shodu s empirickým rozdělením četností. Proto v případě, že v příjmech některé sociální skupiny byla detekována odlehlá pozorování, je vhodné omezit vliv těchto hodnot na odhad parametrů modelu buď jejich úplným vyloučením nebo použitím některé z robustních metod odhadu ([2], [8], [4], [5]). Tímto způsobem můžeme v dosáhnout výrazného zvýšení shody teoretického modelu s empirickým rozdělením příjmů u většiny sociálních skupin.

Identifikaci odlehlých hodnot lze realizovat opět buď graficky (např. s vy-

užitím krabicových diagramů) nebo početně (pomocí vhodných testů odlehlosti). Testové metody identifikace odlehlých pozorování jsou propracovány především pro soubory s normálním rozdělením, dále pak pro soubory s exponenciálním a rovnoměrným rozdělením. Pokud je rozdělení souboru jiného typu, lze v mnoha případech vhodnou transformaci docílit toho, aby transformovaná data $y = f(x)$ měla některé z výše uvedených rozdělení. (Např. transformace na normální rozdělení je pro data s logaritmicko - normálním rozdělením dána vztahy $y = ln(x)$, $y = ln(x - \gamma)$ nebo $y = ln\frac{x-\gamma}{\delta-x}$, kde γ, δ jsou hodnoty teoretického minima a maxima, atd.) Vzhledem k tomu, že předpokládaným statistickým modelem příjmových rozdělení je ve většině případů logaritmicko - normální rozdělení, můžeme po transformaci dat použít některou z metod identifikace odlehlých pozorování založené na předpokladu normality výběru.

K nejčastěji používaným inkluzivním testům existence jednoho nebo dvou odlehlých pozorování v datovém souboru s normálním rozdělením patří test založený na modifikovaném studentizovaném reziduu. Např. test odlehlosti maxima $x_{(n)}$ pracuje se statistikou $T_1 = \frac{x_{(n)}-\bar{x}_1}{\hat{\sigma}_1}$, kde $\bar{x}_1$ a $\hat{\sigma}_1$ jsou odhady průměru a směrodatné odchylky získané z redukovaného výběru, tj. z výběru, který vznikne vypuštěním hodnoty $x_{(n)}$, kde $x_{(n)}$ je n-tá pořádková statistika. H_0 zamítáme na hladině významnosti α, pokud je splněna nerovnost $T_1 > \sqrt{\frac{n}{n-2}} \cdot t_{1-\frac{\alpha}{n}}(n-2)$, kde $t_{1-\frac{\alpha}{n}}(n-2)$ je $100(1-\frac{\alpha}{n})\%$-ní kvantil Studentova rozdělení s $(n-2)$ stupni volnosti. S uvedeným testem úzce souvisí test založený na klasickém studentizovaném reziduu a exkluzivní test Grubbsův. Velmi dobré vlastnosti mají také Dixonovy r-statistiky, které jsou založeny na porovnávání různých vzdáleností mezi pořádkovými statistikami. Potřebné kvantily lze pro uvedené testovací statistiky nalézt např. v [3].

Pro maximalizaci shody empirického rozdělení s teoretickým v případě kontaminovaného modelu je důležitý odhad stupně kontaminace $\hat{\varepsilon}$. Vzhledem k tomu, že rozdělení příjmů je téměř ve všech skupinách asymetrické, lze očekávat, že i kontaminace bude mít asymetrický charakter a hodnoty optimálních useknutí dat $\hat{\alpha}_d$ a $\hat{\alpha}_h$, které odpovídají stupni kontaminace zdola $\hat{\varepsilon}_d$ a shora $\hat{\varepsilon}_k$, budou mít různou velikost. K odhadu stupně kontaminace je nezbytné použít některou z metod detekce většího počtu odlehlých pozorování. K tomuto problému můžeme přistupovat dvěma způsoby. Buď můžeme testovat hypotézu H_0: „Ve výběru neexistují odlehlá pozorování“ proti alternativě H_1: „Ve výběru je právě r odlehlých pozorování“ - tj. provádět tzv. blokové testy odlehlosti, nebo můžeme testovat hypotézu H_0: „Ve výběru je méně než k odlehlých pozorování“ proti alternativě H_1: „Ve výběru je právě k odlehlých pozorování“, kde k nabývá postupně hodnot $r, r-1, r-2, \ldots, 1$ a testovací statistiky se určují z příslušných podmnožin výběru – tj. provádět tzv. sekvenční testy odlehlosti. V případě správného určení hodnoty r mají blokové testy optimální vlastnosti. V praxi jsou však častěji používány sekvenční testy, které nevyžadují velkou přesnost při odhadu předpokládaného počtu odlehlých pozorování r.

Mezi nejznámější sekvenční testy patří ESD test, který pracuje s tzv. extrémní studentizovanou odchylkou. Při určování příslušné testovací statistiky vycházíme z posloupnosti podmnožin výběru $\{A_0, A_1, \ldots, A_{r-1}\}$, kde první člen posloupnosti je tvořen celým výběrem, tj. $A_0 = \{x_{(1)}, x_{(2)}, \ldots, x_{(n)}\}$, a každý následující člen posloupnosti je dán rekurzivně vztahem $A_{i+1} = A_i - \{x(A_i)\}$, kde $x(A_i)$ je nejvzdálenější prvek od průměru $\bar{x}(A_i)$ na množině A_i. To znamená, že pro tuto hodnotu musí platit

$$|x(A_i) - \bar{x}(A_i)| = \max_{x_j \in A_i} |x_j - \bar{x}(A_i)|,$$

kde $\bar{x}(A_i)$ je průměr na podmnožině A_i, $i = 0, 1, \ldots, r-1$.
Extrémní studentizovaná odchylka na podmnožině A_i, tj. veličina ESD_{i+1}, je dána vztahem

$$ESD_{i+1} = \frac{\max_{x_j \in A_i} |x_j - \bar{x}(A_i)|}{s(A_i)},$$

kde $s(A_i)$ je směrodatná odchylka na podmnožině A_i, $i = 0, 1, \ldots, r-1$. H_0 zamítáme na hladině významnosti α, pokud je splněna nerovnost $ESD_{i+1} > L_{i+1}$, kde příslušné kvantily jsou stanovené aproximativně v [9]. Při detekci odlehlých pozorování sekvenční metodou postupujeme iterativně, tzv. „zpětným krokováním“. To znamená, že porovnání vypočtených hodnot provádíme od poslední (nejmenší) vytvořené podmnožiny A_i, $i = r$. Pokud $ESD_{i+1} > L_{i+1}$, pak $i = r$, to znamená, že ve výběru bylo identifikováno r odlehlých pozorování. Sekvenční identifikační procesy jsou vhodné pro počítačové zpracování například pomocí softwarových produktů Matlab, Matematika, MS-Excel apod.

Další skupinu testů, užívaných k detekci většího počtu odlehlých hodnot, tvoří tzv. jednokrokové procedury. Jedná se o postup, kdy procházíme celým souborem (krok za krokem) a testujeme postupně všechny jeho prvky. K rozhodnutí o odlehlosti přitom používáme některou ze statistik vhodných pro identifikaci jedné odlehlé hodnoty.

Pokud je model příjmového rozdělení kontaminován odlehlými pozorováními, projeví se vliv těchto výrazně odlišných hodnot snížením shody empirického rozdělení s předpokládaným teoretickým logaritmicko - normálním modelem. Opětného zvyšování této shody můžeme docílit v tomto případě vhodným useknutím výběrového souboru. Odhady stupňů useknutí dat zdola $\hat{\alpha}_d$ a shora $\hat{\alpha}_h$ by měly zároveň tvořit horní hranice pro odhadnuté stupně kontaminace $\hat{\varepsilon}_d$ a $\hat{\varepsilon}_h$, aby byly splněny nerovnosti $(\alpha_d \geq \varepsilon_d) \wedge (\alpha_h \geq \varepsilon_h)$. Odhad horní hranice kontaminace příjmových rozdělení může být tedy v jednotlivých sociálních skupinách získán odhadem optimální hodnoty useknutí dat. Takovýto odhad lze realizovat např. prostřednictvím dvojrozměrné numerické maximalizace shody empirického rozdělení ročních příjmů domácností s teoretickým modelem. K určení aktuální useknuté hodnoty může být v iteračním kroku použita výše popsaná metoda identifikace nejvzdálenějšího prvku od průměru. Odhad optimálních hodnot useknutí $\hat{\alpha}_d$ a $\hat{\alpha}_h$,

odpovídající maximální dosažitelné shodě empirického a teoretického rozdělení, lze provést např. pomocí numerické minimalizace testovací statistiky $\chi^2 = \sum_{i=1}^{k} \frac{(n_i - n\pi_i)^2}{n\pi_i}$, popřípadě $MSE = \sum_{i=1}^{k}(p_i - \pi_i)^2$ apod. Odhadnuté hodnoty optimálních useknutí budou vždy závislé nejenom na konkrétním výběrovém souboru příjmů a na jeho rozdělení do tříd, ale také na volbě teoretického modelu, na počtu parametrů modelu a na metodě použité k jejich odhadu.

3 Některé dílčí výsledky analýzy příjmových rozdělení

3.1 Použité metody a dosažené výsledky

Zkoumaný datový soubor příjmů domácností pochází z celostátního statistického šetření Mikrocensus 1996. Obsahuje jednak hodnoty ročních příjmů domácností, počty členů domácností a zařazení domácnosti do sociální skupiny podle typu zaměstnání osoby v čele domácnosti. K účelům zkoumání rozdělení ročních příjmů obyvatelstva byly vybrány následující ukazatele: (a) Sociální skupina osoby v čele domácnosti (1 - dělník, 2 - samostatně činný (mimo zemědělství), 3 - zaměstnanec, 4 - samostatně hospodařící rolník, 5 - družstevní rolník, 6 - důchodce v domácnosti s ekonomicky aktivními členy, 7 - důchodce v domácnosti bez ekonomicky aktivních členů, 8 - nezaměstnaný, 0 - ostatní), (b) Počet členů domácnosti, (c) Čistý peněžní příjem domácnosti (v Kč za rok).

Bez újmy na obecnosti se zde můžeme soustředit např. pouze na analýzu souborů dat ročních peněžních příjmů na domácnost. Vizualizací datových souborů (pomocí histogramů a P-P grafů) byly vytipovány „problémové“ soubory, které vykazovaly odlišnosti od předpokládaného teoretického modelu logaritmicko - normální rozdělení se dvěma parametry $LN(\mu, \sigma^2)$ (popřípadě se třemi parametry $LN(\mu, \sigma^2, \gamma)$. Jedná se o výběrové soubory ročních příjmů domácností bez ohledu na sociální skupinu a skupin 1, 3 a 7 ([6]).

Po grafickém průzkumu byl v celém datovém souboru i v každé sociální skupině proveden χ^2 test shody empirického rozdělní s příslušným logaritmicko - normálním modelem. Vzhledem k tomu, že shoda empirického rozdělení četností příjmů na domácnost s teoretickým model byla prokázány pouze u sociálních skupin 5, 6, 8 a 0, bylo provedeno useknutí datových souborů odpovídající stupni kontaminace odhadnutému pomocí jednokrokové procedury využívající klasický t-test standardizovaných reziduí na 5%-ní hladině významnosti (viz tab. 1). K významnému zvýšení shody empirického rozdělení s teoretickým modelem po useknutí došlo pouze u skupin 0, 2 a 4. Naproti tomu u skupin 6 a 8 došlo ke snížení oproti původní situaci a u skupiny č. 5 došlo ke zvýšení pouze u dat seskupených do tříd podle Sturgesova pravidla. Skutečnost, že podle dalších použitých pravidel seskupování došlo ke snížení, může být zapříčiněna změnou počtu tříd při výpočtech χ^2 statistik z celých výběrů (v programu MS Excel) a z useknutých výběrů (v programu Statgraphics for Windows). ([6]).

Sociální skupina	Stupeň kontaminace	
	ε_d	ε_h
všechny	0%	3,055%
1	3,817%	0%
2	0%	3,261%
3	0%	3,210%
4	0%	3,053%
5	0%	5,128%
6	0%	3,374%
7	0%	1,676%
8	0%	2,692%
0	0%	3,390%

Tabulka 1: Odhad stupně kontaminace zdola a shora rozdělení příjmů na domácnost t-testem standardizovaných reziduí ($\alpha = 0{,}05$).

Sociální skupina	Stupeň useknutí		χ^2 test
	α_d	α_h	p-value
0	1%	9%	0,869695
2	2%	2%	0,432965
4	7,5%	1,5%	0,739458
5	2%	4%	0,797002
6	0%	1%	0,153245
8	6%	1%	0,848642

Tabulka 2: Odhad optimálního stupně useknutí zdola a shora v souborech příjmů na domácnost pomocí numerické minimalizace χ^2 statistiky.

Z důvodu nejednotnosti vlivu useknutí podle výše odhadnutého stupně kontaminace na shodu empirického rozdělení četností s logaritmicko - normálním modelem byla provedena úprava odhadu stupně useknutí zdola $\hat{\alpha}_d$ a shora $\hat{\alpha}_h$ pomocí numerické minimalizace χ^2 statistiky (viz tab. 2). Odhady byly realizovány v programu MS Excel. Z tabulky vyplývá, že bylo dosaženo výrazného zvýšení shody empirického rozdělení s teoretickým, a proto můžeme považovat logaritmicko - normální rozdělení za vhodný model pro většinu sociálních skupin.

Další odhad stupně kontaminace příjmových rozdělení v jednotlivých sociálních skupinách, tj. počtů hodnot odlehlých zdola r_d a shora r_h, byl realizován prostřednictvím dvou sekvenčních testů – klasické a modifikované veze ESD testu. Modifikace ESD testu spočívala v tom, že procesy identifikace aktuální „podezřelé“ hodnoty a jejího testování na odlehlost probíhají současně – v témž iteračním kroku. Metoda vychází ze skutečnosti, že nadhodnocení předpokládaného počtu odlehlých pozorování r má na efektivnost

Sociální skupina	Rozsah	Klas. ESD		Modif. ESD		Num. optimalizace	
	n	r_d	r_h	r_d	r_h	r'_d	r'_h
0	236	0	0	1	3	3	27
2	1748	0	3	0	1	36	36
4	131	0	2	1	4	15	3
5	195	0	0	5	0	4	8
6	1156	0	1	0	1	0	12
8	260	0	0	0	0	18	3

Tabulka 3: Odhady počtu odlehlých hodnot příjmů na domácnost určené sekvenčními testy a odhady jejich horních hranic určené numerickou optimalizací.

použitého testu minimální vliv ([7]) a zároveň musí být splněna nerovnost $r \leq [\frac{n}{4}]$. Při realizaci modifikované verze ESD testu je setříděný soubor příjmů na domácnost nejprve symetricky maximálně redukován, tzn. že z každé strany je useknuto $[\frac{n}{4}]$ hodnot, takže prvním členem posloupnosti podmnožin výběru $\{A_r, A_{r-1}, \ldots, A_0\}$ je množina $A_r = \{x_{([\frac{n}{4}]+1)}, \ldots x_{(n-[\frac{n}{4}])}\}$, každý následující člen pak odpovídá rekurzivnímu vztahu $A_{i-1} = A_i + \{x(A_i)\}$. Iterační krok spočívá ve vyhledání a otestování odlehlosti „nejbližší“ useknuté hodnoty $x(A_i)$, která má od průměru aktuálního redukovaného souboru $\bar{x}(A_i)$ minimální vzdálenost. Pro každou hodnotu $x(A_i)$ je určena statistika ESD_{i+1}, která je porovnána s hodnotou příslušného kvantilu L_{i+1}. Pokud je splněna nerovnost $ESD_{i+1} > L_{i+1}$, iterační cyklus končí a $r = r_d + r_h = i$. Rozdíl mezi klasickou a modifikovanou formou testu je především ve startovacím bodě testovacího procesu a v určení hodnoty $\bar{x}(A_i)$. Iterační procedura obou sekvenčních testů byla realizována v programu MS Excel.

K odhadu počtu hodnot vhodných k useknutí zdola r'_d a shora r'_h prostřednictvím numerické optimalizace shody empirického rozdělení s dvouparametrickým logaritmicko - normálním modelem byla použita statistika $MSE = \sum_{i=1}^{k}(p_i - \pi_i)^2$. Optimalizační procedura byla realizována v programu Matlab. Výsledky (viz tab. 3) ukazují, že počty identifikovaných odlehlých hodnot příjmů na domácnost, získané prostřednictvím obou sekvenčních testů, jsou ve všech sociálních skupinách srovnatelné, nezávislé na rozsahu souborů a relativně velmi malé $(0 \leq r_d \leq 5)$, $(0 \leq r_h \leq 4)$. Naproti tomu při numerické optimalizaci bylo ve většině sociálních skupin dosaženo maximální shody empirického rozdělení s teoretickým modelem až po useknutí většího počtu hodnot příjmů $(0 \leq r'_d \leq 36)$, $(3 \leq r'_h \leq 36)$. Ani zde nebyla prokázána závislost optimálního počtu useknutých hodnot na rozsahu souboru. Vyjádříme-li si procentuální velikosti kontaminace $\varepsilon_d = \frac{r_d}{n}$, $\varepsilon_h = \frac{r_h}{n}$ a procentuální velikosti optimálních useknutí $\alpha_d = \frac{r'_d}{n}$, $\alpha_h = \frac{r'_h}{n}$, zjistíme, že ve všech sociálních skupinách je zachována platnost vztahu $(\varepsilon_d \leq \alpha_d) \wedge (\varepsilon_h \leq \alpha_h)$, to znamená, že odhady useknutí lze ve všech případech považovat za horní hranice odhadů kontaminace.

3.2 Závěry

Probíhající transformace hospodářství České Republiky z plánované formy na tržní, která byla zahájena před více než deseti lety, se projevila v úrovni a struktuře čistých ročních peněžních příjmů domácností získaných z Mikrocensu 1996 pouze částečně. Došlo především k výrazné diferenciaci příjmů, tj. ke vzniku (malého počtu) domácností s výrazně vysokými a s výrazně nízkými příjmy, které způsobují narušení teoretického modelu a snižují jeho statistickou významnost. Naproti tomu uvedená analýza rozdělení ročních příjmů domácností získaných z Mikrocensu 1996 prokázala u většiny sociálních skupin platnost logaritmicko - normálního modelu, kontaminovaného malým podílem odlehlých hodnot. Pro nalezení optimálního statistického modelu rozdělení příjmů je proto vhodné nejprve provést v každé sociální skupině detekci odlehlých pozorování, popřípadě optimalizaci stupně useknutí souboru. K odhadu charakteristik modelu je z výše zmíněných důvodů vhodné použít některou z robustních metod odhadu.

Reference

[1] Anděl J. (2002). *Základy matematické statistiky.* Preprint MFF UK, Praha.

[2] Antoch J., Vorlíčková D. (2004). *Vybrané metody statistické analýzy dat.* ACADEMIA, Praha.

[3] Barnett V., Lewis T. (1978). *Outliers in statistical data.* 1st edn. John Wiley, Chichester

[4] Bartošová J. (2003). *Robustní metody odhadů.* Oeconomica, Praha, 234 – 246.

[5] Bartošová J. (2003). *Příjmové modely.* Výpočtová štatistika, SŠDS, Bratislava, 7 – 11.

[6] Bartošová J. (2004)

[7] Jain R. B., Pingel L. A. (1981). *A procedure for estimating the number of outliers.* Commun. Statist. Theor. Meth. **10**, 10029 – 10041.

[8] Jurečková J. (2001). *Robustní statistické metody.* Karolinum, Praha.

[9] Militký J., Militká D. (1985). *Moderní matematicko-statistické metody v hutnictví.* Základní statistické metody III. Dvůr Králové.

[10] Scott, D. W. (1992). *Multivariate density estimation. Theory, practice and visualization.* J. Willey, New York.

Adresa: J. Bartošová, Vysoká škola ekonomická, Fakulta managementu, katedra managementu informací, Jarošovská 1117/II, 377 01 Jindřichův Hradec, ČR

E-mail: barto-ji@fm.vse.cz

NOVÉ CHARAKTERISTIKY ROZDĚLENÍ A VÝBĚRŮ Z ROZDĚLENÍ

Zdeněk Fabián

Klíčová slova: Základní charakteristiky, core funkce, rozdělení s těžkými chvosty.

Abstrakt: V článku je definujeme core funkci, Johnsonovo těžiště a Johnsonovu disperzi spojitého pravděpodobnostního rozdělení a ukážeme, že výběrové těžiště a výběrová Johnsonova disperze rozumně popisují náhodné výběry i z rozdělení, která nemají momenty.

1 Úvod

Core funkce T_F absolutně spojitého rozdělení F s hustotou f regulární na nosiči, kterým je otevřený interval $\Sigma = (a, b) \subseteq R$, je ve sbornících Robustu [1]-[3] a v článcích [4]-[5] zavedena jako

$$T_F(x) = T_G(\eta(x)), \tag{1}$$

kde $T_G(y) = -g'(y)/g(y)$ je skórová funkce *prototypu* $G = F\eta^{-1}$ s hustotou g a kde $\eta : \Sigma \to R$ je diferencovatelné rostoucí zobrazení. V tomto článku je η pro $\Sigma = R$ identické zobrazení a pro $\Sigma \neq R$ je dáno předpisem

$$\eta(x) = \begin{cases} \log(x-a) & \text{pokud} \quad \Sigma = (a, \infty) \\ \log \dfrac{(x-a)}{(b-x)} & \text{pokud} \quad \Sigma = (a, b) \\ \log(b-x) & \text{pokud} \quad \Sigma = (-\infty, b). \end{cases} \tag{2}$$

Zobrazení (2) zavedl Johnson [6]. Přiřazuje jednoduše matematicky vyjádřeným F jednoduché prototypy a prototyp lognormálního rozdělení je normální. Pro některá rozdělení je však možno nalézt vhodnější zobrazení, viz [2] a [3]. V tomto článku se omezíme na zobrazení (2) a budeme pro stručnost mluvit o core funkci rozdělení, i když budeme mít na mysli „Johnsonovu core funkci". Snad nás k tomu opravňuje fakt, že Johnsonova core funkce je pro mnoho ve statistice užívaných rozdělení tou nejjednodušší core funkcí.

V předešlých článcích bylo nutné definici (1) zobecnit pro parametrická rozdělení. V tomto článku půjdeme jinou cestou: vymezíme vhodný parametrický prostor, ukážeme, že prakticky libovolné rozdělení umíme převést na *upravené rozdělení* s parametry v tomto prostoru a pro upravená parametrická rozdělení definujeme (Johnsonovu) core funkci.

V práci [3] jsme zavedli pojem těžiště τ rozdělení F s nosičem $\Sigma \neq R$ jakožto nuly core funkce a střední informaci I_τ jakožto Fisherovu informaci o těžišti. V tomto článku definujeme Johnsonovu disperzi jako $\sigma_J^2 = I_\tau^{-1}$ a ukážeme na příkladech, že dvojice (τ, σ_J) popisuje i rozdělení, která nemají momenty a že dvojice jejich odhadů $(\hat{\tau}, \hat{\sigma}_J)$ pěkně charakterizuje náhodné výběry z takových rozdělení.

2 Struktura parametrického prostoru

Označme stejně jako v [3] Π_Σ množinu regulárních rozdělení na nosiči $\Sigma \subseteq R$. Je-li $G \in \Pi_R$, automorfizmus $[\mu, s]$: $R \to R$ definovaný pro $(\mu, s) \in R \times (0, \infty)$ vztahem

$$[\mu, s]\,(y) = \mu + sy, \qquad y \in R$$

určuje rodinu $\mathcal{G} = \{G_{\mu,s} = G\,[\mu, s]^{-1}\}$ s rodičem G. Platí $\mathcal{G} \subset \Pi_R$ a hustota rozdělení $G_{\mu,s} \in \mathcal{G}$ je $g_{\mu,s}(y) = s^{-1} g((y - \mu)/s), y \in R$, kde g je hustota G.

Definujme pro $(\tau, s) \in \Sigma \times (0, \infty)$ jedno-jednoznačné zobrazení $[\tau, s]$: $\Sigma \to R$ vztahem

$$[\tau, s] = \eta(\tau) + s\eta(x), \qquad x \in \Sigma. \tag{3}$$

Zobrazení (3) definuje pro každou $F \in \Pi_\Sigma$ parametrickou rodinu

$$\mathcal{F} = \left\{ F_{\tau,s} = F\,\{\tau, s\}^{-1} : (\tau, s) \in \Sigma \times (0, \infty) \right\}$$

s rodičem F. Parametr $\tau \in \Sigma$,

$$\tau = \eta^{-1}(\mu), \tag{4}$$

jsme nazvali těžištěm (zde: *Johnsonovým těžištěm*) rozdělení $F_{\tau,s}$. Pro všechna rozdělení $F_{\tau,s} \in \mathcal{F}$ platí $F_{\tau,s} = G_{\eta(\tau),s}\eta$, kde η je definováno v (2) a $G_{\eta(\tau),s}$ je prvkem rodiny $\mathcal{G}$ s rodičem $G = F\eta^{-1}$. Dále $\mathcal{F} \subset \Pi_\Sigma$ a pro hustoty $f_{\tau,s}$ rozdělení $F_{\tau,s} \in \mathcal{F}$ platí [5, Proposition 5])

$$f_{\tau,s}(x) = \frac{1}{s} g\left(\frac{\eta(x) - \eta(\tau)}{s}\right) \eta'(x), \qquad x \in \Sigma.$$

Buď nyní $\tilde{F} \in \Pi_\Sigma$. Najdeme jeho prototyp $\tilde{G} = \tilde{F}\eta^{-1}$. Pokud je $\tilde{G}$ unimodální, určíme jeho mód $y^*(\tilde{G})$ a položíme $\mu = y^*(\tilde{G})$. Není-li $\tilde{G}$ unimodální, je třeba μ zvolit nějak jinak. $\tilde{G}$ pak upravíme na tvar G_{θ_R} s vektorem parametrů $\theta_R = (\mu, s, \lambda) \in \Theta_R$, kde $\Theta_R = R \times (0, \infty) \times (0, \infty)^{m-2}$ a kde $\lambda \in (0, \infty)^{m-2}$ je vektor tvarových parametrů. To jde vždy. G_{θ_R} bude mít hustotu

$$g(y; \theta_R) = \frac{1}{s} g_\lambda\left(\frac{y - \mu}{s}\right), \qquad y \in R. \tag{5}$$

Buď dále $\Theta_\Sigma = \Sigma \times (0, \infty) \times (0, \infty)^{m-2}$ a položme $\theta_\Sigma = (\tau, s, \lambda) \in \Theta_\Sigma$. Na Σ zkonstruujeme obraz prototypu, $F_{\theta_\Sigma} = G_{\theta_R}\eta$, který bude upraveným tvarem původního rozdělením $\tilde{F}$ a bude mít hustotu

$$f(x; \theta_\Sigma) = \frac{1}{s} g_\lambda\left(\frac{\eta(x) - \eta(\tau)}{s}\right) \eta'(x), \qquad x \in \Sigma. \tag{6}$$

Množinu rozdělení na Σ s hustotami (6) označíme $\mathcal{F}_{\Theta_\Sigma}$. Budeme předpokládat, že splňují obvyklé podmínky regularity (např. [7, str. 462]).

Příklad 1. Exponenciální rozdělení s nosičem $\Sigma=(0,\infty)$ má hustotu $f(x;\tau)=\frac{1}{\tau}e^{-x/\tau}$. Z (2) máme $\eta(x) = \log x, \eta'(x) = 1/x$ a $\eta^{-1}(y) = e^y$. Protože $f(x;\tau) = \frac{1}{x}\frac{x}{\tau}e^{-x/\tau}$, je hustota prototypu $g(y;\tau) = \frac{1}{\tau}e^y e^{-\frac{1}{\tau}e^y}$. Ta má mód $y^*(G) = \log\tau$, takže parametr τ exponenciálního rozdělení je Johnsonovým těžištěm a exponenciální rozdělení je v požadovaném tvaru.

3 Core funkce

Veličině $u_R = (y-\mu)/s, y \in R$ budeme říkat pivotní proměnná (jako nepřesný překlad termínu pivotal quantity, [8, str. 101]).

Definice 1. Buď X náhodná veličina s nosičem $\Sigma \subseteq R$ s rozdělením v upraveném tvaru $F_{\tau,s,\lambda} \in \mathcal{F}_{\Theta_\Sigma}$ s parametrem těžiště (4) a $G_{\mu,s,\lambda} = F_{\tau,s,\lambda}\eta^{-1}$ jeho Johnsonův prototyp s hustotou ve tvaru (5). *Pivotní proměnnou na* $\Sigma \neq R$ nazveme veličinu

$$u_\Sigma = \frac{\eta(x) - \eta(\tau)}{s}. \tag{7}$$

Core funkci náhodné veličiny X definujeme jako

$$T_F(x;\tau,s,\lambda) = -\frac{g'_\lambda(u_\Sigma)}{g_\lambda(u_\Sigma)}.$$

Core funkce je tedy skórovou funkcí svého prototypu vyjádřeného pomocí pivotní proměnné na Σ. V této stručné formulaci jsme třikrát vynechali přívlastek Johnsonova/y. Definice je ekvivalentní té v předešlých článcích [1]-[3]. Připomeňme si nejdůležitější vlastnost core funkcí [3, Věta 1]:

$$\frac{\partial}{\partial\tau}\log f(x;\tau,s,\lambda) = \frac{\eta'(\tau)}{s}T_F(x;\tau,s,\lambda), \tag{8}$$

t.j. *core funkce je úměrná věrohodnostnímu skóru parametru těžiště.*

4 Těžiště rozdělení

V [1]-[3] jsme definovali $k-$tý core moment náhodné veličiny X s rozdělením $F \in \mathcal{F}_{\Theta_\Sigma}$ a core funkcí T_F vztahem $M_k(F) = ET_F^k(x) = \int_\Sigma T_F^k(x)\ dF(x)$ a ukázali, že $M_1(F) = 0$. *Těžiště* x_T rozdělení $F \in \mathcal{F}_{\Theta_\Sigma}$ jsme definovali jako nulu core funkce, t.j. $x_T : T_F(x;\theta_\Sigma) = 0$ a ukázali, že pro parametrická rozdělení z $\mathcal{F}_{\Theta_\Sigma}$ je $x_T = \tau$ kde τ je hodnota parametru těžiště pro dané rozdělení. V případě že $\Theta_\Sigma = \Sigma$ a $\mathbf{X_n} = (X_1,\ldots,X_n)$ je náhodným výběrem z rozdělení $F_\tau \in \mathcal{F}_{\Theta_\Sigma}$, určíme *výběrové těžiště* $\hat\tau_n$ z rovnice

$$\hat\tau_n : \qquad \frac{1}{n}\sum_{i=1}^{n} T_F(x_i;\tau) = 0.$$

Z (8) je patrné, že $\hat\tau_n$ je maximálně věrohodným odhadem těžiště rozdělení s nosičem $\Sigma \neq R$ a zároveň i „těžištěm“ datového souboru $\mathbf{X_n}$.

5 Johnsonova disperze

Připomeňme, že obvyklé podmínky regularity zajišťují existenci druhého core momentu, který navíc nezávisí na parametrech τ a $s = 1/\beta$, což snadno dokážeme: Buď T_G core funkce prototypu $G = F\eta^{-1}$ rozdělení F. Podle Definice 1 je $T_F(x;\theta_\Sigma) = T_{G_\lambda}(u_\Sigma)$. Použijeme ještě vztahy (6) a (7) máme

$$\begin{aligned} M_2(F) = \int_\Sigma T_F(x;\theta_\Sigma)^2 f(x;\theta_\Sigma)\,dx &= \int_\Sigma T_{G_\lambda}(u_\Sigma)^2 g_\lambda(u_\Sigma)\eta'(x)\,dx/s \\ &= \int_R T_{G_\lambda}(u_R)^2 g_\lambda(u_R)\,du_R. \end{aligned}$$

Poslední integrál je pouze funkcí λ, pišme tedy $M_2(\lambda)$ místo $M_2(F)$. Z (8) vyplývá vztah

$$I_\tau(\theta_\Sigma) = \left(\frac{\eta'(\tau)}{s}\right)^2 M_2(\lambda)$$

mezi Fisherovou informací o těžišti a druhým core momentem, kterého využijeme k definici veličiny popisující variabilitu rozdělení.

Definice 2. Buď X náhodná veličina s rozdělením $F \in \mathcal{F}_{\Theta_\Sigma}$. Její *Johnsonovu disperzi* σ_J^2 definujeme jako

$$\sigma_J^2 = I_\tau(\theta_\Sigma)^{-1} = \frac{s^2}{\eta'(\tau)^2 M_2(\lambda)}. \tag{9}$$

σ_J nazveme *Johnsonovou směrodatnou odchylkou.*

Výběrovou Johnsonovu disperzi pak přirozeně definujeme jako $\hat\sigma_J^2 = \hat s^2/[\eta'(\hat\tau)^2 M_2(\hat\lambda)]$, kde $(\hat\tau, \hat s, \hat\lambda)$ je vektor maximálně věrohodných odhadů parametrů. Platí $\eta'(\tau) > 0$ a $M_2(\lambda) > 0$. Předpokládáme-li dále spojitost M_2 vzhledem k λ, je $\hat\sigma_J^2$ konzistentním odhadem Johnsonovy disperze (9). V dalším ukážeme, že i docela rozumným odhadem variability rozdělení včetně rozdělení, která nemají momenty.

6 Příklady

Studujeme závislost Johnsonovy směrodatné odchylky $\sigma_J = s\tau/\sqrt{M_2(\lambda)}$ a výběrové Johnsonovy směrodatné odchylky $\hat\sigma_J = \hat s\hat\tau/\sqrt{M_2(\hat\lambda)}$ na parametrech některých rozdělení s nosičem $\Sigma = (0,\infty)$ (kde $\eta'(\tau) = 1/\tau$). V některých případech uvádíme výsledky simulací, standardně jsme generovali 1000 náhodných výběrů délky 50 bodů z daného rozdělení a parametry odhadovali metodou maximální věrohodnosti pomocí programů z knihovny Matlab.

Příklad 2. Hustoty Weibullova and Fréchetova rozdělení na $\Sigma = (0,\infty)$ jsou

$$f_W(x;\tau,\beta) = \frac{\beta}{x}\left(\frac{x}{\tau}\right)^\beta e^{-(\frac{x}{\tau})^\beta}, \qquad f_F(x;\tau,\beta) = \frac{\beta}{x}\left(\frac{x}{\tau}\right)^{-\beta} e^{-(\frac{x}{\tau})^{-\beta}}$$

a jejich momenty

$$\mu_k^W = \tau^k \Gamma(1 + k/\beta), \qquad \mu_k^F = \tau^k \Gamma(1 - k/\beta).$$

Momenty μ_k^F existují pouze v případě, že $k < \beta$. Pro core funkce platí

$$T_W(x; \tau, \beta) = (x/\tau)^\beta - 1, \qquad T_F(x; \tau, \beta) = 1 - (\tau/x)^\beta,$$

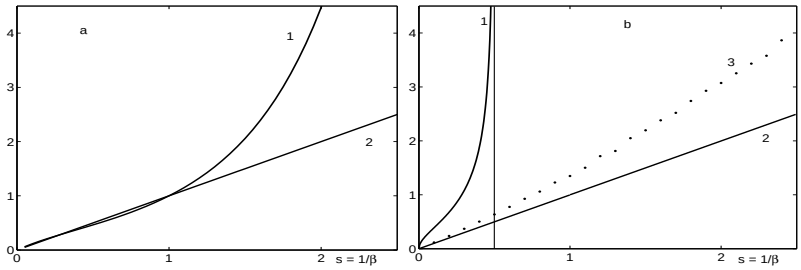
druhé core momenty jsou $M_2 = 1$ a $\sigma_C^2 = \tau^2/\beta^2$ pro obě rozdělení.

Pro $\tau = 1$ a sadu hodnot $s = 1/\beta$ a jsem generoval výběry z Fréchetova rozdělení a určil odhady $\hat{\tau}_i, i = 1, \ldots, 1000$. Jejich průměr $\bar{\tau}$ je uveden v Tabulce 1 spolu s průměrnými hodnotami aritmetického ($\bar{m}$) a harmonického ($\bar{h}$) průměru. Pro hodnoty $s \geq 1$ neexistuje střední hodnota Fréchetova rozdělení. $\bar{\tau}$ zůstává blízký jedné. Odhad těžiště je robustní, core funkce rozdělení je pro velká x ohraničená.

s	0.5	0.7	0.9	1	1.1	1.3	1.5
$\bar{m}$	1.77	2.83	6.61	-	-	-	-
$\bar{h}$	1.136	1.110	1.058	1.021	0.975	0.894	0.790
$\bar{\tau}$	1.008	1.009	1.017	1.021	1.020	1.035	1.047

Tabulka 1: Průměrné hodnoty výběrového těžíště, aritmetického průměru a harmonického průměru pro různé hodnoty parametru měřítka Fréchetova rozdělení.

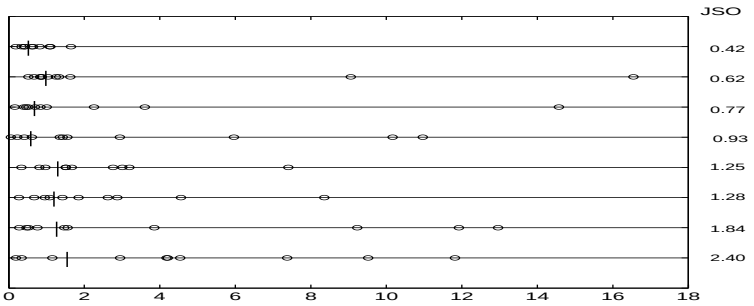
Na obr. 1 jsou grafy závislosti σ a σ_J na parametru $s = 1/\beta$ Weibullova (a) a Fréchetova (b) rozdělení.



Obrázek 1: Míry variability Weibullova (a) a Fréchetova (b) rozdělení $1 - \sigma, 2 - \sigma_J, 3 - \hat{\sigma}_{MAD}$.

Na obr. 1b je i průběh MAD odhadu $\hat{\sigma}_{\mathrm{MAD}} = 1.483 * \mathrm{med}(|X_i - \mathrm{med}(X)|)$, určený simulací. Johnsonova standardní odchylka má pro obě rozdělení lineární průběh, což je v obou případech triviální, ale robustní (porovnejte s průběhem $\hat{\sigma}_{\mathrm{MAD}}$ Fréchetova rozdělení).

Na obr. 2 jsou zakresleny nezávislé jedenáctibodové výběry z Fréchetova rozdělení $f_F(x; 1, 0.8)$, které nemá střední hodnotu ani rozptyl. Výběrová těžiště a výběrové Johnsonovy směrodatné odchylky se zdají docela dobře



Obrázek 2: $\hat{\tau}_n$ (čárky) a $\hat{\sigma}_J$ (JSO) náhodných výběrů z Fréchetova rozdělení.

charakterizovat datové soubory. Odhady jsou robustní a odhadnuté hodnoty nejsou ovlivněny odlehlými body, které se v některých výběrech vyskytly (a nevešly se do obrázku).

Příklad 3. Uvažujme rodinu $\mathcal{F}_{TB}$ na $\Sigma = (0, \infty)$ s hustotami ve tvaru

$$f_{TB}(x; \tau, \beta, p, q) = \left(\frac{q}{p}\right)^q \frac{\beta}{B(p,q)x} \frac{(x/\tau)^{\beta p}}{[(x/\tau)^\beta + q/p]^{p+q}}. \tag{10}$$

Je to upravený tvar transformované rodiny beta s Johnsonovým parametrem τ, parametrem měřítka $s = 1/\beta$ a tvarovými parametry $\lambda = (p, q)$. Členy rodiny jsou např. log-logistické, Fisher-Snedecorovo, beta-prime a Burrovo XII rozdělení. Položme $\nu = p/q$. k-tý moment rodiny (10) jsem spočetl jako

$$\mu_k = \left(\frac{\tau}{\nu^{1/\beta}}\right)^k \frac{\Gamma(\nu q + k/\beta)\Gamma(q - k/\beta)}{\Gamma(\nu q)\Gamma(q)}$$

pro $k < \beta q$. Core funkce rodiny mají tvar

$$T_{TB}(x; \tau, \beta, q, \nu) = q\frac{(x/\tau)^\beta - 1}{(x/\tau)^\beta + 1/\nu},$$

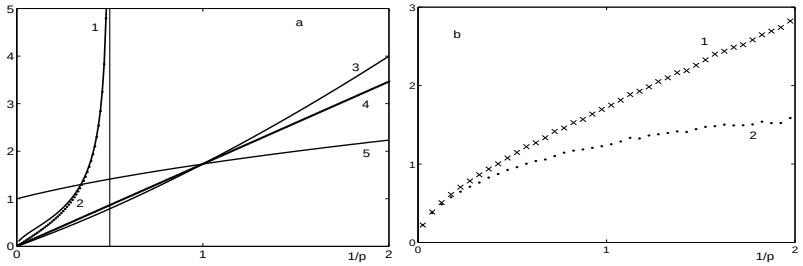
druhý core moment je $M_2(\nu, q) = \nu q^2/[(\nu + 1)q + 1]$.

V Tabulce 2 nalezneme hustotu, core funkci, obor existence rozptylu a vzorec pro Johnsonovu disperzi „elementárních“ náhodných veličin rodiny TB.

f_{TB}	T_{TB}	exist. of σ^2	σ_J^2
$f_{TB}(x; 1, \beta, 1, 1) = \frac{\beta x^{\beta-1}}{(x^\beta+1)^2}$	$\frac{x^\beta-1}{x^\beta+1}$	$\frac{1}{\beta} < \frac{1}{2}$	$3/\beta^2$
$f_{TB}(x; 1, 1, q, 1) = \frac{\Gamma(2q)}{\Gamma^2(q)} \frac{x^{q-1}}{(x+1)^{2q}}$	$q\frac{x-1}{x+1}$	$\frac{1}{q} < \frac{1}{2}$	$(2 + 1/q)/q$
$f_{TB}(x; 1, 1, 1, \nu) = \frac{x^{\nu-1}}{(x+1/\nu)^{1+\nu}}$	$\frac{x-1}{x+1/\nu}$	neexistuje	$1 + 2/\nu$

Tabulka 2: Trojice ementárních rozdělení transformované rodiny beta.

„Elementárními rozděleními“ myslíme rozdělení, která mají hustoty právě s jedním parametrem různým od jedné. Grafy závislosti σ a σ_J na převrácených hodnotách jednotlivých parametrů jsou na obr. 3a. Všechna σ_J rostou s rostoucí převrácenou hodnotou parametru přibližně lineárně. Průběh průměrné σ_J výběrů z rozdělení beta-prime s hustotou $f(x;p,q) = f_{TB}(x;p/q,1,q,p/q)$ v závislosti na $1/p$ pro $q = p$ je zachycen na obr. 3b spolu s průměrnou hodnotou $\hat{\sigma}_{\mathrm{MAD}}$.



Obrázek 3: Míry variability transformovaného beta rozdělení (a) elementární: $1-\sigma(1/\alpha), 2-\sigma(1/\beta), 3-\sigma_J(1/\alpha), 4-\sigma_J(1/\beta), 5-\sigma_J(1/\nu)$, (b) beta-prime $(q=p)$: 1-$\hat{\sigma}_C$, 2-$\hat{\sigma}_{MAD}$

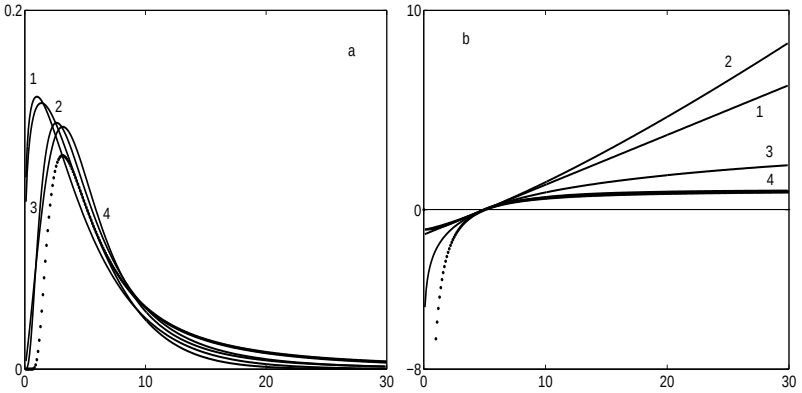
V následující tabulce uvádíme přehled vzorců.

rozdělení	$f(x)$	$T_F(x)$	τ	σ_J^2
Cauchyovo	$\frac{1}{\pi s(1+(\frac{x-\mu}{s})^2)}$	$\frac{2(\frac{x-\mu}{s})}{1+(\frac{x-\mu}{s})^2}$	μ	$2s^2$
lognormální	$\frac{\beta}{\sqrt{2\pi}x}e^{-\frac{1}{2}\log^2(x/\tau)^\beta}$	$\log\left(\frac{x}{\tau}\right)^\beta$	τ	$1/\beta^2$
exponenciální	$\frac{1}{\tau}e^{-x/\tau}$	$\frac{x}{\tau}-1$	τ	τ^2
Weibullovo	$\frac{\beta}{x}(\frac{x}{\tau})^\beta e^{-(\frac{x}{\tau})^\beta}$	$(\frac{x}{\tau})^\beta - 1$	τ	τ^2/β^2
Fréchetovo	$\frac{\beta}{x}(\frac{\tau}{x})^\beta e^{-(\frac{\tau}{x})^\beta}$	$1-(\frac{\tau}{x})^\beta$	τ	τ^2/β^2
gamma	$\frac{\gamma^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\gamma x}$	$\gamma x-\alpha$	α/γ	α/γ^2
chi-kvadrát	$\frac{1}{2^{\nu/2}\Gamma(\nu/2)}x^{\nu/2-1}e^{-x/2}$	$\frac{\nu}{2}\left(\frac{x}{\nu}-1\right)$	ν	2ν
log-logistické	$\frac{\beta}{x}\frac{(x/\tau)^\beta}{(1+(x/\tau)^\beta)^2}$	$\frac{(x/\tau)^\beta-1}{(x/\tau)^\beta+1}$	τ	$3\tau^2/\beta^2$
beta	$\frac{1}{B(p,q)}x^{p-1}(1-x)^{q-1}$	$(p+q)x-p$	$\frac{p}{p+q}$	$\frac{pq(p+q+1)}{(p+q)^4}$
beta-prime	$\frac{1}{B(p,q)}\frac{x^{p-1}}{(x+1)^{p+q}}$	$\frac{qx-p}{x+1}$	$\frac{p}{q}$	$\frac{p(p+q+1)}{q^3}$

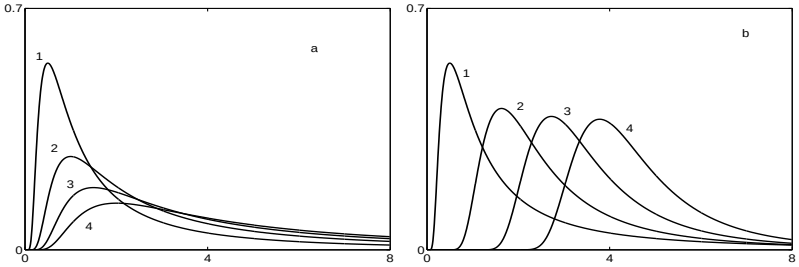
Tabulka 3: Core funkce, těžiště a Johnsonova disperze některých rozdělení.

Hustoty a core funkce některých rozdělení z Tabulky 3 pro tutéž zvolenou hodnotu σ_J jsou zakresleny na obr.4. Z obrázku je patrné, že core funkce zvýrazňují chování chvostů rozdělení.

Konečně na obr. 5 porovnáváme hustoty Fréchetova rozdělení při různých hodnotách parametru těžiště pro konstantní parametr měřítka (a) a konstantní Johnsonovu disperzi (b). I z tohoto obrázku je vidět, že pojem Johnsonovy disperze má docela dobrý smysl.



Obrázek 4: Hustoty (a) a core funkce (b) rozdělení se stejnou hodnotou σ_J 1-gamma, 2-Weibullovo, 3-lognormální, 4-log-logistické, ... Fréchetovo.



Obrázek 5: Hustoty Fréchetova rozdělení pro $\tau = 1, 2, 3, 4$. (a) s =konst. (b) σ_J =konst.

Reference

[1] Fabián Z. (1997). *Geometrické momenty.* Sb. Robust'96, 49 – 62.
[2] Fabián Z. (2001). *MM-odhady.* Robust'2000, 33 – 41.
[3] Fabián Z. (2003). *Informace ve výběru z rozdělení.* Robust'2002, 95 – 106.
[4] Fabián Z. (2001). *Induced cores and their use in robust parametric estimation.* Communications in Statistics-Theory and Methods **30**, 537 – 556.
[5] Fabián Z., Vajda I. (2003). *Core functions and core divergences of regular distributions.* Kybernetika 39, **1**, 29 – 42.
[6] Johnson N. L. (1949). *Systems of frequency curves generated by methods of translations.* Biometrika **36**, 149 – 176.
[7] Lehmann E. L., Casella G. (2001). *Theory of point estimation.* Springer.
[8] Lindley J. K. (1996). *Parametric statistical inference.* Clarendon Press, Oxford.

Poděkování: Autor děkuje Igoru Vajdovi za významnou pomoc při přípravě rukopisu. Práci podpořil grant AV ČR IAA1075403.

Adresa: Z. Fabián, ÚI AV ČR, Pod Vodárenskou věží 2, 182 07 Praha 8
E-mail: zdenek@cs.cas.cz

ODHADY REGRESNÍCH PARAMETRŮ S NEÚPLNÝMI DATY

Michal Kulich

Klíčová slova: Regrese, chybějící data, odhadovací rovnice, eficientní odhad.

Abstrakt: Zabýváme se problémem odhadu regresních koeficientů za přítomnosti chybějících hodnot v regresorech. Nejjednodušší postup, vynechání pozorování s chybějícími daty, je neefektivní a může vést k vychýleným odhadům. Zavedeme dvě třídy vážených odhadů v obecném regresním modelu, ukážeme, že jsou navzájem ekvivalentní a najdeme asymptoticky optimální odhad v rámci těchto tříd. Tento optimální odhad využívá neúplná pozorování i data z doplňkových veličin, které nesou informaci o chybějících hodnotách. Poznatky shrneme v krátké diskusi.

1 Úvod

Chybějící nebo neúplná data jsou běžnou a neoddělitelnou součástí statistické praxe. Datový soubor, který obsahuje kompletní hodnoty všech veličin a pozorování, se vyskytuje snad jen v pohádkách a učebnicích statistiky. Naráží-li ale statistik tak často na data s chybějícími pozorováními, jak si s nimi má poradit v praxi? Nejběžnější přístup je všechna neúplná pozorování vynechat a analysovat pouze podmnožinu úplných pozorování. Tento přístup je jednoduchý, ale má dvě očividné nevýhody. Za prvé, funguje pouze tehdy, je-li neúplnost daného pozorování nezávislá na tom, co toto pozorování obsahuje, anebo co by obsahovalo, kdyby bylo úplné (Rubinova podmínka MCAR, *missing completely at random*, viz [1]). Jinými slovy, vynecháním neúplných pozorování předpokládáme, že úplná pozorování tvoří náhodný výběr z původního datového souboru. Druhá nevýhoda spočívá v tom, že ignorováním neúplných pozorování se zbavujeme informace v těchto pozorováních obsažené a spokojujeme se se suboptimální analýzou.

V tomto článku stručně načrtneme, jak analysovat data s chybějícími pozorováními tak, abychom mohli oslabit podmínku MCAR a abychom efektivněji využili informaci obsaženou v neúplných pozorováních. Předpokládejme, že Y je odezva čili závisle proměnná, jejíž rozdělení závisí na vektoru prediktorů čili nezávisle proměnných X, který má d složek. Budeme se zajímat o podmíněnou střední hodnotu Y, je-li dáno $X = x$, jež je dána modelem

$$\mathrm{E}\,(Y \mid X = x\,) = g(x^{\mathrm{T}}\beta_0), \tag{1}$$

kde g je nějaká monotonní funkce z $\mathbb{R}$ do $D \subseteq \mathbb{R}$. Ekvivalentně předpokládáme, že

$$Y = g(X^{\mathrm{T}}\beta_0) + \varepsilon, \qquad \mathrm{E}\,(\varepsilon \mid X\,) = 0.$$

U všech veličin předpokládáme existenci alespoň druhých momentů. Rozdělení X není nijak určeno, a dokonce ani rozdělení $\mathcal{L}(Y \mid X = x)$ (a tudíž $\mathcal{L}(\varepsilon \mid X = x)$) obecně nemusí být specifikováno. Zajímá nás samozřejmě odhad vektoru regresních parametrů β_0.

Tento dosti obecný model zahrnuje jak lineární regresní modely včetně analýzy rozptylu, tak zobecněné lineární modely a semiparametrické regresní modely založené na kvazivěrohodnosti.

Vezměme si n nezávislých (a zatím úplných) pozorování (Y_i, X_i) rozdělených stejně jako (Y, X). Uvažujeme odhady $\widehat{\beta}_F$ parametru β_0, které řeší odhadovací rovnici $\overline{U}_F(\widehat{\beta}_F) = 0$, kde

$$\overline{U}_F(\beta) = \sum_{i=1}^{n} U_i^F(\beta \mid Y_i, X_i).$$

Statistika $\overline{U}_F(\beta)$ se obecně nazývá pseudoskóre, náhodný vektor U_i^F se nazývá odhadovací (pseudoskórová) funkce.

V našem modelu můžeme zvolit

$$U_i^F(\beta \mid Y_i, X_i) = h(X_i)[Y_i - g(X_i^{\mathrm{T}}\beta)],$$

kde h je nějaká dostatečně hladká funkce $\mathbb{R}^d \to \mathbb{R}^d$ taková, že $h(X)$ má druhé momenty. Očividně $\mathrm{E}_i^F(\beta_0 \mid Y_i, X_i) = 0$, což je klíčová vlastnost pro zajištění konsistence $\widehat{\beta}_F$. Předpokládáme, že pro každé $\beta \neq \beta_0$ naopak platí $\mathrm{E}_i^F(\beta \mid Y_i, X_i) \neq 0$, že existují matice

$$\Sigma_F = \operatorname{var} U_i^F(\beta_0) < \infty \qquad \text{a} \qquad D_F = -\mathrm{E}\,\frac{\partial U_i^F(\beta_0 \mid Y_i, X_i)}{\partial \beta} < \infty$$

a že Σ_F je positivně definitní a D_F je regulární.

Za podmínek regularity existuje jednoznačně určené řešení $\widehat{\beta}_F$ rovnice $\overline{U}_F(\beta) = 0$, které je asymptoticky lineární, to jest platí

$$\sqrt{n}(\widehat{\beta}_F - \beta_0) = D_F^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_i^F(\beta_0 \mid Y_i, X_i) + o_P(1)$$

(viz monografie [3]). Náhodný vektor $D_F^{-1}U_i^F(\beta \mid Y_i, X_i)$ se nazývá vlivová funkce i-tého pozorování. Z centrální limitní věty pro nezávislé stejně rozdělené náhodné vektory plyne, že $\sqrt{n}(\widehat{\beta}_F - \beta_0) \xrightarrow{D} \mathrm{N}(0, D_F^{-1}\Sigma_F D_F^{-1})$. Volba funkce $h(X)$ ovlivňuje asymptotický rozptyl $\widehat{\beta}_F$. Lze ukázat, že nejmenší rozptyl je dosažen, je-li

$$h(X) = h^*(X) \equiv \frac{\partial g(X^{\mathrm{T}}\beta)}{\partial \beta}\{\operatorname{var}(\varepsilon \mid X)\}^{-1}.$$

2 Neúplná data v regresorech

Nyní rozšíříme náš model tak, aby zahrnoval určitý mechanismus vzniku neúplných pozorování. Předpokládejme, že odezvu Y pozorujeme vždy a že vektor prediktorů X je rozdělen na dvě části X^M a X^A tak, že X^A je vždy pozorováno celé, zatímco X^M je buď pozorováno celé anebo celé chybí. Dále zavedeme vektor doplňkových veličin V, které jsou vždy pozorovány. Doplňkové veličiny sice nehrají přímou roli v regresním modelu (1), ale mohou nést informaci o nepozorovaných hodnotách X^M. V X^M máme veličiny, které se měří obtížně anebo ne vždy úspěšně, v X^A máme veličiny, které je snadné změřit nebo dohledat, a ve V máme nepřesná měření složek X^M, prediktory složek X^M a veličiny, na nichž závisí pravděpodobnost, že X_M bude pozorováno.

Pro zjednodušení výkladu budeme předpokládat, že známe pravděpodobnostní mechanismus, jenž vede k neúplnosti pozorování X. Tento mechanismus specifikujeme takto: nechť $\xi_1, \ldots, \xi_n$ jsou nezávislé nula-jedničkové náhodné veličiny. Pro pozorování z takzvaného validačního souboru $\mathcal{V} \equiv \{i \in \{1, \ldots, n\} : \xi_i = 1\}$ budiž X^M k disposici, takže jejich (úplná) data jsou: $(Y_i, X_i^M, X_i^A, V_i, \xi_i)$. Pro ostatní pozorování X^M chybí, takže máme k disposici pouze (Y_i, X_i^A, V_i, ξ_i). Pravděpodobnost výběru přitom může záviset na vždy pozorovaných datech $W = (Y, X^A, V)$ skrze vztah

$$\pi_i = \mathrm{P}\,(\xi_i = 1 \mid W_i\,) = \pi(W_i),$$

kde $\pi(\cdot)$ je známá funkce s hodnotami v $(\delta, 1]$, kde $\delta > 0$. Požadujeme, aby ξ_i bylo podmíněně nezávislé na X_i^M, je-li dáno W_i. Díky tomu tento mechanismus vzniku chybějících pozorování splňuje Rubinovu podmínku MAR (*missing at random*, viz [1]).

V praxi se takovéto dvoufázové výběrové schema často provádí v případech, že z finančních, časových anebo praktických důvodů veličiny X^M nelze změřit pro celý studovaný soubor.

3 Upravené vážené odhady

Zabývejme se nyní odhady parametrů modely (1) při neúplných datech. Budeme uvažovat třídu odhadů $\widehat{\beta}_M \equiv \widehat{\beta}_M(\varphi)$ (viz [2]), které řeší $\overline{U}_M(\widehat{\beta}_M) = 0$, kde $\overline{U}_M(\beta) = \sum_{i=1}^n U_i^M(\beta)$ a

$$U_i^M(\beta) \equiv U_i^M(\beta, \varphi) = \frac{\xi_i}{\pi_i} U_i^F(\beta \mid Y_i, X_i) - \frac{\xi_i - \pi_i}{\pi_i} \varphi(\beta \mid W_i). \qquad (2)$$

První člen váží příspěvky do $\overline{U}_F$ inversními pravděpodobnostmi výběru a je brán v úvahu pouze u pozorování patřících do validační skupiny. Druhý člen vnáší do odhadové rovnice informaci z veškerých pozorovaných dat pomocí nějaké vektorové funkce φ. Tato funkce musí být dostatečně hladká a musí splňovat $\operatorname{var} \varphi(\beta \mid W_i) < \infty$.

Zvolíme-li $\varphi = 0$, dostaneme klasický vážený odhad $\widehat{\beta}_M(0)$ (viz [4]) definovaný skrze

$$U_i^M(\beta, 0) = \frac{\xi_i}{\pi_i} U_i^F(\beta \mid Y_i, X_i). \tag{3}$$

Tento odhad je konsistentní, avšak nečerpá informaci ze všech dat.

Bez ohledu na tvar funkce φ platí, že

$$\mathrm{E}\,\frac{\xi_i}{\pi_i} U_i^F(\beta_0) = \mathrm{E}\,[\pi_i^{-1}\mathrm{E}\,(\,\xi_i \mid W_i\,)\,U_i^F] = \mathrm{E}\,U_i^F = 0,$$

a že

$$\mathrm{E}\,\frac{\xi_i - \pi_i}{\pi_i}\varphi(\beta, W_i) = \mathrm{E}\,[\pi_i^{-1}\{\mathrm{E}\,(\,\xi_i \mid W_i\,) - \pi_i\}\varphi(\beta, W_i)] = 0.$$

Z toho plyne, že pro každé φ platí $\mathrm{E}\,U_i^M(\beta_0, \varphi) = 0$, což je klíčová podmínka pro konsistenci $\widehat{\beta}_M(\varphi)$.

Za podmínek regularity (mj. $\pi_i > \delta > 0$) má rovnice $\overline{U}_M(\beta, \varphi) = 0$ s pravděpodobností konvergující k 1 jediné řešení $\widehat{\beta}_M$, které jest asymptoticky lineárním odhadem β_0 s vlivovou funkcí $D_F^{-1}U_i^M(\beta, \varphi)$ [2].

Tudíž platí

$$\sqrt{n}(\widehat{\beta}_M - \beta_0) \xrightarrow{D} \mathrm{N}(0, D_F^{-1}\Sigma_M D_F^{-1}),$$

kde $\Sigma_M = \mathrm{var}\,U_i^M(\beta_0) = \Sigma_F + \Sigma_E$. Pro Σ_E dostaneme po krátkém výpočtu vyjádření

$$\Sigma_E = \mathrm{E}\left\{\frac{1-\pi_i}{\pi_i}[U_i^F(\beta_0) - \varphi(\beta_0, W_i)]^{\otimes 2}\right\},$$

kde $a^{\otimes 2}$ je značení pro aa^{T}.

Funkce φ ovlivňuje rozptyl odhadu $\widehat{\beta}_M(\varphi)$ skrze Σ_E. Mezi všemi odhady typu (2) je rozptyl $\widehat{\beta}_M(\varphi)$ minimální, právě když

$$\varphi = \widetilde{\varphi} \equiv \mathrm{E}\left(U_i^F(\beta_0) \,\middle|\, W_i\right) \tag{4}$$

(viz [5], Věta 18, str. 57).

Optimální $\widetilde{\varphi}$ je neznámé, ale může být odhadnuto z pozorovaných dat. Nejprve spočteme jakýkoli konsistentní odhad β_0, např. $\widehat{\beta}_M(0)$. Jednotlivé složky $\mathrm{E}\left(U_i^F(\beta_0) \,\middle|\, W_i\right)$ odhadneme pomocí d regresních modelů s odpovídajícími složkami $U_i^F(\widehat{\beta}_M(0))$ jako odezvami a s vhodně zvolenými funkcemi W_i jako regresory. Spočteme odhady regresních parametrů pro tyto modely na datech z validační skupiny a získáme $\widehat{\varphi} \equiv \widehat{\mathrm{E}}\left(U_i^F(\widehat{\beta}_M(0)) \,\middle|\, W_i\right)$, které dosadíme do (2) za φ. Lze dokázat, že pokud regresní odhady v modelu pro $\widehat{\varphi}$ konvergují ke skutečným parametrům v řádu $O(n^{-1/2})$, výsledný odhad $\widehat{\beta}_M(\widehat{\varphi})$ má stejné asymptotické rozdělení jako $\widehat{\beta}_M(\widetilde{\varphi})$.

4 Vážené odhady s odhadnutými pravděpodobnostmi

Vraťme se k odhadům typu (3) a uvažujme odhady tohoto tvaru, které však neváží převrácenými hodnotami skutečných pravděpodobností π_i, ale odhadnutých pravděpodobností $\widehat{\pi}_i$. Ukážeme, že nahrazením skutečných a známých pravděpodobností π_i jejich odhady vygenerujeme celou třídu odhadů typu (2) a získáme jiný náhled na optimální odhad $\widehat{\beta}_M(\widetilde{\varphi})$.

Nechť skutečné pravděpodobnosti $\pi_i \equiv \pi(W_i)$ splňují logistický model

$$\log \frac{\pi_i}{1-\pi_i} = \alpha_0^{\mathrm{T}} u(W_i), \tag{5}$$

kde $u(W_i)$ je nějaká r-rozměrná transformace W_i a $\alpha_0 \in \mathbb{R}^r$ je známý parametr. Takový model zajisté existuje pro r dostatečně velké (např. $r \geq r_0$). Pravděpodobnosti výběru $\pi_i \equiv \pi(\alpha_0, W_i)$ odhadneme jako $\pi(\widehat{\alpha}, W_i)$, kde $\widehat{\alpha}$ je odhad α_0 v modelu (5). Tím jsme získali třídu odhadů $\widehat{\beta}_H$ definovaných jako řešení rovnice $\overline{U}_H(\beta) = 0$, kde $\overline{U}_H = \sum_{i=1}^n U_i^H(\beta)$ a

$$U_i^H(\beta) = \frac{\xi_i}{\pi(\widehat{\alpha}, W_i)} U_i^F(\beta \mid Y_i, X_i). \tag{6}$$

Jednotlivé odhady se navzájem liší specifikací $u(W_i)$ v logistickém modelu pro π_i. Data z neúplných pozorování jsou použita jen nepřímo, skrze odhadnuté pravděpodobnosti π_i.

Uveďme si zde výsledek poprvé publikovaný v práci Robins, Rotnitzky a Zhao [2].

Věta 1: *$\widehat{\beta}_H$ je asymptoticky lineární odhad s odhadovou funkcí*

$$\frac{\xi_i}{\pi_i} U_i^F(\beta) - \frac{\xi_i - \pi_i}{\pi_i} \varphi(\beta \mid W_i),$$

kde

$$\varphi(\beta \mid W_i) = \mathrm{E}\,[\widetilde{\varphi}(W_i)(1-\pi_i)u(W_i)]^{\mathrm{T}} I_\alpha^{-1} \pi_i u(W_i), \tag{7}$$

$\widetilde{\varphi}$ je definováno vztahem (4) a $I_\alpha = \mathrm{E}\,\pi_i(1-\pi_i)u(W_i)^{\otimes 2}$.

Robins a kolegové ukázali, že každou „rozumnou“ funkci $\varphi(\beta \mid W_i)$, která definuje upravený vážený odhad vztahem (2), lze vyjádřit ve tvaru (7) pro nějaké $u(W_i)$ v logistickém modelu pro π_i. Každý odhad z třídy $\widehat{\beta}_H$ je tudíž asymptoticky ekvivalentní nějakému odhadu $\widehat{\beta}_M$ s určitým φ a naopak každý odhad $\widehat{\beta}_M$ je ekvivalentní nějakému odhadu $\widehat{\beta}_H$ s určitým $u(W_i)$.

Věta 1 má i další zajímavé důsledky. Lze ukázat, že asymptotický rozptyl $\widehat{\beta}_H$ závisí na dimensi prostoru generovaném sloupci $u(W_i)$: s rostoucí dimensí $u(W_i)$ asymptotický rozptyl $\widehat{\beta}_H$ nikdy neroste. Přitom ale víme, že určité $u(W_i)$ (které ostatně známe) generuje zcela správný model (5) pro pravděpodobnosti výběru. Přidáváním dalších nadbytečných složek do $u(W_i)$, které musí mít skutečné parametry (odpovídající složky α_0) rovné nule, přesto nikdy nezvětšíme rozptyl odhadu $\widehat{\beta}_H$. Toto ovšem platí jen asymptoticky;

v praxi nemůžeme očekávat, že zvětšováním už tak bohatého regresního modelu pro π_i budeme roztyl neustále zmenšovat.

Z věty 1 je dokonce možné zjistit, který logistický model vede k odhadu $\widehat{\beta}_H$ s nejmenším rozptylem ve třídě (6). Vezmeme-li nějaký platný regresní model (5) a přidáme-li jako další regresory $\mathrm{E}\left(U_i^F(\beta_0) \,\middle|\, W_i\right)/\pi_i$, dostaneme odhad, který je asymptoticky ekvivalentní optimálnímu odhadu $\widehat{\beta}_M(\widetilde{\varphi})$. Toto lze ukázat rozpisem pravé strany rovnosti (7) pro tento speciální případ. Jelikož neznáme β_0 ani potřebnou podmíněnou střední hodnotu, v praxi použijeme odhad $\widehat{\mathrm{E}}\left(U_i^F(\widehat{\beta}_M(0)) \,\middle|\, W_i\right)$ podobně jako v případě optimálního $\widetilde{\varphi}$.

NÁZNAK DŮKAZU VĚTY 1:

Na rozdíl od publikace [2] zde přímo odvodíme asymptoticky lineární vyjádření pro pseudoskóre $\overline{U}_H(\beta_0)$. Máme

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_i^H(\beta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\xi_i}{\pi_i(\alpha_0)}U_i^F(\beta) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(\frac{1}{\pi_i(\widehat{\alpha})} - \frac{1}{\pi_i(\alpha_0)}\Big)\xi_i U_i^F(\beta).$$

Potřebujeme aproximovat druhý člen součtem nezávislých veličin. Rozložme nejprve $\dfrac{1}{\pi_i(\widehat{\alpha})} - \dfrac{1}{\pi_i(\alpha_0)}$. Z vlastností odhadu parametru logistické regrese dostaneme

$$\sqrt{n}(\widehat{\alpha} - \alpha_0) = \frac{1}{\sqrt{n}} I_\alpha^{-1} \sum_{j=1}^{n} [\xi_i - \pi(\alpha_0, W_i)]u(W_i) + o_P(1),$$

kde $I_\alpha = \mathrm{E}\,\pi_i(1-\pi_i)u(W_i)^{\otimes 2}$ je informační matice.

Nechť $f_i(\alpha) = \dfrac{1}{\pi(\alpha, W_i)} = \dfrac{1+\exp\{\alpha^{\mathrm{T}}u(W_i)\}}{\exp\{\alpha^{\mathrm{T}}u(W_i)\}}$. Z Taylorova rozvoje dostaneme

$$\sqrt{n}(f_i(\widehat{\alpha}) - f_i(\alpha_0)) \approx f_i'(\alpha_0)^{\mathrm{T}}\sqrt{n}(\widehat{\alpha} - \alpha_0),$$

kde

$$f_i'(\alpha) = \frac{\partial f_i(\alpha)}{\partial \alpha} = -\frac{1-\pi(\alpha, W_i)}{\pi(\alpha, W_i)}u(W_i).$$

Nyní dáme všechny kusy dohromady.

$$\begin{aligned}\frac{1}{n} \quad & \sum \sqrt{n}\Big(\frac{1}{\pi(\widehat{\alpha}, W_i)} - \frac{1}{\pi(\alpha_0, W_i)}\Big)\xi_i U_i^F(\beta_0) \\ &= -\frac{1}{n}\sum_{i=1}^{n}\frac{1-\pi(\alpha_0, W_i)}{\pi(\alpha_0, W_i)}u(W_i)^{\mathrm{T}}\frac{1}{\sqrt{n}}\sum_{j=1}^{n} I_\alpha^{-1}[\xi_j - \pi(\alpha_0, W_j)]u(W_j)\xi_i U_i^F(\beta_0) \\ &\quad + o_P(1)\end{aligned}$$

Ztransponujeme a přehodíme pořadí sčítání:

$$-\frac{1}{\sqrt{n}}\Big\{\sum_{j=1}^{n}[\xi_j-\pi(\alpha_0,W_j)]u(W_j)^{\mathrm{T}}I_\alpha^{-1} \times \frac{1}{n}\sum_{i=1}^{n}\frac{\xi_i}{\pi(\alpha_0,W_i)}[1-\pi(\alpha_0,W_i)]u(W_i)U_i^F(\beta_0)^{\mathrm{T}}\Big\}^{\mathrm{T}}+o_P \quad (1)$$

Součet přes i nám teď konverguje v pravděpodobnosti; nahradíme jej jeho limitou $\mathrm{E}\,[(1-\pi_i)u(W_i)U_i^F(\beta_0)^{\mathrm{T}}] = \mathrm{E}\,[(1-\pi_i)u(W_i)\mathrm{E}\left(U_i^F(\beta_0)^{\mathrm{T}} \,\middle|\, W_i\right)]$ a celý výraz zpětně ztransponujeme. Důkaz je hotov.

5 Závěr

Ukázali jsme dva různé způsoby, jak získat optimální odhad regresního parametru β_0 v dosti obecném modelu (1) za přítomnosti neúplných dat v regresorech. Optimální upravený vážený odhad $\widehat{\beta}_M(\widehat{\varphi})$ a optimální odhad $\widehat{\beta}_H$ odvozený v předchozí kapitole jsou asymptoticky ekvivalentní. Z praktického hlediska je šikovnější odhad $\widehat{\beta}_H$, neboť jej lze spočítat s použitím standartních výpočetních procedur. Jeho výpočet vyžaduje: (i) výpočet $\widehat{\beta}_M(0)$; (ii) výpočet regresních modelů pro $\widehat{\mathrm{E}}\left(U_i^F(\widehat{\beta}_M(0)) \,\middle|\, W_i\right)$; (iii) výpočet logistické regrese pro $\pi(\widehat{\alpha},W_i)$; a (iv) výpočet vážené regrese pro model (1) s vahami $\pi^{-1}(\widehat{\alpha},W_i)$. Oproti tomu výpočet $\widehat{\beta}_M(\widehat{\varphi})$ vyžaduje numerické řešení rovnice $\overline{U}_M(\widehat{\beta}_M)=0$ místo kroků (iii) a (iv).

Uvažované třídy odhadů $\widehat{\beta}_H$ a $\widehat{\beta}_M$ vycházejí z odhadové funkce $U_i^F = h(X_i)[Y_i - g(X_i^{\mathrm{T}}\beta)]$, jejíž asymptotické vlastnosti závisí na volbě funkce h. Zmínili jsme, že při úplných datech je rozptyl minimalisován pro $h^*(X) = [\partial g(X^{\mathrm{T}}\beta)/\partial\beta]\{\mathrm{var}\,(\varepsilon \mid X)\}^{-1}$. Při neúplných datech ovšem h^* již není optimální. Plně eficientní odhad dostaneme, pokud h^* nahradíme jinou funkcí, která řeší jistou integrální rovnici (viz [2]). Tuto rovnici lze vyřešit, pokud jsou naše data plně diskrétní; pro spojité regresory anebo spojitou odezvu je její řešení dost těžký oříšek. Proto se v praxi raději spokojujeme s funkcí h^*, i když víme, že není úplně optimální.

Teoretické principy, které jsme zde ukázali, lze zobecnit i na složitější regresní modely a obecnější struktury chybějících dat. Kromě chybějících regresorů můžeme uvažovat i chybějící hodnoty závisle proměnné. Všechny tyto principy se týkají i situace, kdy chybějící hodnoty vznikají nahodile. Pak ale nemáme mechanismus vzniku chybějících dat zcela pod kontrolou a neznáme správný logistický model (5). Můžeme se pokusit takový model co nejlépe sestavit, nicméně riskujeme, že dostaneme vychýlené odhady π_i a tudíž vychýlený odhad β_0. V některých případech správný logistický model pro π_i sestavit ani nelze. To se stane, pokud neplatí předpoklad MAR, to jest v případech, kdy pravděpodobnost, že určitá hodnota chybí, může záviset na oné chybějící hodnotě. Platnost předpokladu MAR není bohužel možné na daném datovém souboru otestovat.

Na závěr ještě poznamenejme, že předkládané výsledky jsou pouze asymptotické aproximace. O tom, jak se tyto metody chovají při malém počtu pozorování, není ještě zdaleka dost známo.

Reference

[1] Rubin D.B. (1976). *Inference and missing data.* Biometrika **63**, 581 – 592.

[2] Robins J., Rotnitzky A., Zhao L.P. (1994). *Estimation of regression coefficients when some regressors are not always observed.* JASA **89**, 846 – 866.

[3] Manski C.F. (1988). *Analog estimation methods in econometrics.* Chapman and Hall, New York.

[4] Horvitz D.G., Thompsom D.J. (1952). *A generalization of sampling without replacement from a finite universe.* JASA **47**, 663 – 685.

[5] Anděl J. (1985). *Matematická statistika.* SNTL, Praha.

Poděkování: Děkuji anonymnímu recensentovi nejmenovaného zahraničního časopisu za podnět ke studiu této problematiky.

Adresa: M. Kulich, Katedra pravděpodobnosti a matematické statistiky, Matematicko-fyzikální fakulta University Karlovy, Sokolovská 83, 186 75 Praha 8

E-mail: `kulich@karlin.mff.cuni.cz`

STOCHASTICKÉ ALGORITMY V ODHADU PARAMETRŮ REGRESNÍCH MODELŮ

Josef Tvrdík

Klíčová slova: Globální optimalizace, řízené náhodné prohledávání, nelineární regrese.

Abstrakt: V článku je popsán algoritmus řízeného náhodného prohledávání (Controlled Random Search, CRS) a jeho zobecnění využívající soutěž heuristik pro generování nového bodu populace. Dále jsou uvedeny výsledky aplikace dvou variant tohoto algoritmu na řadě obtížných úloh odhadu parametrů nelineárních regresních modelů.

1 Úvod

Úlohou nalezení globálního minima účelové funkce $f : D \to \mathcal{R}, \quad D \subseteq \mathcal{R}^d$ je nalezení bodu $\mathbf{x}^* \in D$ s nejnižší funkční hodnotou, $\mathbf{x}^* = \arg\min_{\mathbf{x}\in D} f(\mathbf{x})$. V řadě statistických metod je potřeba nalézt globální minimum (nebo maximum) v souvislé oblasti $D = \prod_{i=1}^{d} \langle a_i, b_i \rangle,\ a_i < b_i,\ i = 1, 2, \ldots, d$, a účelovou funkci $f(\mathbf{x})$ umíme vyhodnotit s požadovanou přesností v každém bodu $\mathbf{x} \in D$. Příklady takových úloh jsou odhady parametrů nelineárních regresních modelů metodou nejmenších čtverců, robustní odhady parametrů atd. Jelikož účelová funkce může být multimodální a odhady parametrů navíc bývají korelované, je nalezení správných hodnot odhadů algoritmicky obtížné. Pro hledání globálního minima je však možné užít stochastických algoritmů pro globální optimalizaci, zejména evolučních algoritmů.

2 Řízené náhodné prohledávání

Řízené náhodné prohledávání (Controlled Random Search, CRS) je velmi jednoduchý stochastický algoritmus pro hledání globálního minima. Původní verzi tohoto algoritmu publikoval před čtvrt stoletím Price [5], některé modifikace tohoto algoritmu jsou uvedeny v [1], [2], [9]. V algoritmu CRS se na počátku vygeneruje náhodně populace $\mathcal{P}$, tvořená N body v prohledávaném prostoru D. Počet vygenerovaných bodů populace N je větší než dimenze d prohledávaného prostoru D. Pro generování nového bodu $\mathbf{y}$ v každém iteračním kroku Price užíval reflexi (1) v simplexu [4], kdy simplex je vytvořen $d + 1$ body náhodně vybranými z populace:

$$\mathbf{y} = 2\mathbf{g} - \mathbf{x}_H, \tag{1}$$

kde $\mathbf{x}_H$ je bod simplexu s největší hodnotou účelové funkce a $\mathbf{g}$ je těžiště zbývajících d bodů simplexu. Nahrazením nejhoršího bodu populace novým bodem $\mathbf{y}$ dosahujeme toho, že populace se koncentruje v okolí dosud nalezeného bodu s nejmenší funkční hodnotou. Nový bod $\mathbf{y}$ lze však generovat

i jinou heuristikou než reflexí v simplexu podle (1). Tak dostáváme zobecněný algoritmus řízeného náhodného prohledávání, který lze zapsat takto:

Algoritmus 1. Zobecněný CRS
generuj populaci $\mathcal{P}$, tj. N bodů náhodně v D;
repeat
  najdi $\mathbf{x}_{\max} \in \mathcal{P}$, $f(\mathbf{x}_{\max}) \geq f(\mathbf{x}), \quad \mathbf{x} \in \mathcal{P}$;
  repeat
    užij nějakou heuristiku k vygenerování nového bodu $\mathbf{y} \in D$;
  until $f(\mathbf{y}) < f(\mathbf{x}_{\max})$;
  $\mathbf{x}_{\max} := \mathbf{y}$;
until podmínka ukončení;

Jako příklady heuristik uvedeme ty, které byly užity v implementaci algoritmu ověřované v této práci. Znáhodněná reflexe v simplexu je popsána vztahem

$$\mathbf{y} = \mathbf{g} + U\ (\mathbf{g} - \mathbf{x}_H), \tag{2}$$

U je náhodná veličina vhodného rozdělení. Zde je užito rovnoměrné rozdělení na $[s, \alpha - s)$, $\alpha > 0$ a $s \in (0, \alpha/2)$ jsou vstupní parametry heuristiky. Střední hodnota je $EU = \alpha/2$. Tato heuristika je v dalším textu označena REFL.

Znáhodněnou modifikací reflexe popsané v [1] je REFL-B, kdy nový bod $\mathbf{y}$ se generuje podle (2), ale do simplexu je vždy zařazen nejlepší bod populace $\mathbf{x}_{\min}$ s funkční hodnotou $f_{\min}$, $f_{\min} \leq f(\mathbf{x})$, $\mathbf{x} \in \mathcal{P}$, a d bodů simplexu se pak vybere náhodně z ostatních bodů populace.

U heuristik vycházejících z diferenciální evoluce [7] se nejdříve generuje bod $\mathbf{u}$

$$\mathbf{u} = \mathbf{r}_1 + F(\mathbf{r}_2 - \mathbf{r}_3), \tag{3}$$

$\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3$ jsou navzájem různé body náhodně vybrané z populace $\mathcal{P}$, $F > 0$ je vstupní parametr. Nový vektor $\mathbf{y}$ vznikne „křížením“ vektoru $\mathbf{u}$ a náhodně vybraného vektoru $\mathbf{x}$ tak, že kterýkoli jeho prvek x_j je nahrazen hodnotou u_j s pravděpodobností C. Pokud žádné x_j nebylo přepsáno hodnotou u_j nebo při volbě $C = 0$, nahrazuje se jeden náhodně vybraný prvek vektoru $\mathbf{x}$:

$$y_j = \begin{cases} u_j & \text{když } R_j \leq C \quad \text{nebo } j = I \\ x_j & \text{když } R_j > C \quad \text{a } j \neq I\ , \end{cases} \tag{4}$$

kde I je náhodně vybrané celé celé číslo z $\{1, 2, \ldots, d\}$, $R_j \in (0, 1)$ jsou voleny náhodně a nezávisle pro každé j a $C \in [0, 1]$ je vstupní parametr. Ali a Törn [1] navrhli určovat hodnotu parametru F v každém iteračním kroku podle adaptivního pravidla

$$F = \begin{cases} \max(F_{\min}, 1 - |\frac{f_{\max}}{f_{\min}}|) & \text{if } |\frac{f_{\max}}{f_{\min}}| < 1 \\ \max(F_{\min}, 1 - |\frac{f_{\min}}{f_{\max}}|) & \text{jinak,} \end{cases} \tag{5}$$

kde $f_{\min}$, $f_{\max}$ jsou minimální a maximální funkční hodnoty v populaci $\mathcal{P}$ a $F_{\min}$ je vstupní parametr, který zabezpečuje, aby bylo $F \in [F_{\min}, 1)$. Předpokládá se, že tento způsob výpočtu F udržuje prohledávání diverzifikované v počátečním stadiu a intenzivnější v pozdější fázi prohledávání, což má zvyšovat spolehlivost hledání i rychlost konvergence. Tato heuristika je dále označena DE-ADP.

V práci [8] byl popsán evoluční algoritmus se soutěžícími heuristikami. Stejný přístup můžeme užít i v řízeném náhodném prohledávání, neboť CRS je jen speciální případ zmíněného evolučního algoritmu. Mějme k dispozici h heuristik a v každém kroku ke generování nového bodu vybíráme náhodně i-tou heuristiku s pravděpodobností $q_i,\ i = 1, 2, \ldots, h$. Pravděpodobnosti q_i měníme v závislosti na úspěšnosti i-té heuristiky v průběhu vyhledávacího procesu. Heuristiku považujeme za úspěšnou, když generuje nový bod $\mathbf{y}$ takový, že $f(\mathbf{y}) < f(\mathbf{x}_{\max}))$. Pokud n_i je dosavadní počet úspěchů i-té heuristiky, pravděpodobnost q_i je úměrná tomuto počtu úspěchů

$$q_i = \frac{n_i + n_0}{\sum_{j=1}^{h}(n_j + n_0)}, \tag{6}$$

kde $n_0 > 0$ je vstupní parametr algoritmu. Nastavením $n_0 \geq 1$ zabezpečíme, že jeden náhodný úspěch heuristiky nevyvolá příliš velkou změnu v hodnotě q_i. Algoritmus užívající k hodnocení úspěšnosti heuristik (6) je v dalším textu označen COMP1. Jinou možností, jak ohodnotit úspěšnost heuristiky, je vážit úspěšnost relativní změnou v hodnotě funkce. Váha w_i se určí jako

$$w_i = \frac{f_{\max} - \max(f(\mathbf{y}), f_{\min})}{f_{\max} - f_{\min}}. \tag{7}$$

Hodnoty w_i jsou v intervalu $(0,\ 1)$ a pravděpodobnost q_i se pak vyhodnotí jako

$$q_i = \frac{W_i + w_0}{\sum_{j=1}^{h}(W_j + w_0)}\ , \tag{8}$$

kde W_i je součet vah w_i v předcházejícím hledání a $w_0 > 0$ je vstupní parametr algoritmu. Algoritmus užívající takové hodnocení úspěšnosti je označen COMP4.

Aby se zabránilo potlačení možnosti výběru některé z heuristik, lze zadat vstupní parametr δ a klesne-li kterákoli hodnota q_i pod hodnotu δ, jsou pravděpodobnosti výběru heuristik nastaveny na jejich počáteční hodnoty $q_i = 1/h$.

Zobecněný CRS algoritmus s osmi soutěžícími heuristikami byl ověřován na řadě testovacích funkcích užívaných při porovnávání algoritmů pro globální optimalizaci [9]. Výsledky ukázaly, že algoritmus byl jak spolehlivější, tak rychlejší než diferenciální evoluce, která je považována za velmi efektivní stochastický algoritmus pro tento typ úloh [7]. Navíc pro různé funkce bylo pozorováno různé rozdělení četnosti užitých heuristik, což ukazuje, že tento algoritmus se soutěžícími heuristikami je schopen se adaptovat podle aktuálně řešené úlohy.

3 Odhad parametrů nelineární regrese

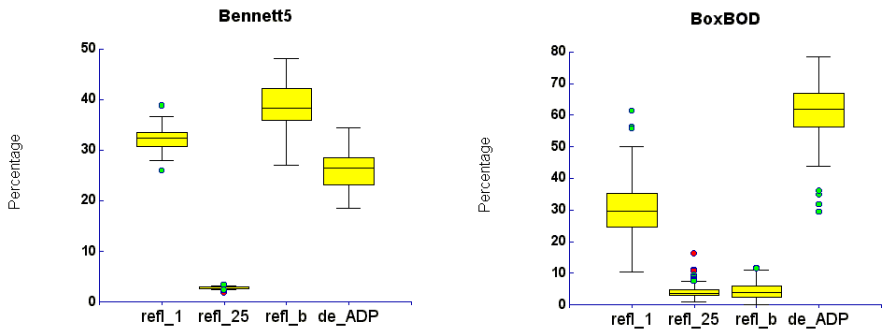
V referenční databázi NIST [6] je dvacet sedm testovacích úloh odhadu parametrů nelineárních regresních modelů. Pro první fázi experimentálního ověřování bylo odtud vybráno všech osm úloh s nejvyšším vyznačeným stupněm obtížnosti. Tyto úlohy jsou obtížně řešitelné pomocí standardního statistického softwaru, ve kterém se pro nalezení minima užívají deterministické algoritmy (různé modifikace Levenberg-Marquardtova nebo Gauss-Newtonova algoritmu, simplexová metoda). NCSS 2001, PLUS 4.5, SPSS 10.0 a SYSTAT 8.0, zhruba v polovině z těchto úloh nenalezly žádné řešení nebo skončily v lokálním minimu. Podrobněji jsou výsledky uvedeny v [10].

Na těchto úlohách byly ověřovány i různé varianty CRS se soutěžícími heuristikami. Pro každou úlohu bylo provedeno sto opakování. Sledován byl způsob ukončení (zda bylo nalezeno řešení dostatečně blízké certifikovanému v [6]) a počet vyhodnocení účelové funkce potřebný k dosažení podmínky ukončení, tj. rozdíl v indexu determinace $R2$ mezi nejhorším a nejlepším bodem populace je menší než 1×10^{-12} a několik dalších veličin charakterizujících průběh prohledávání. Vstupní parametry algoritmu byly nastaveny takto: velikost populace $N = 10d$, $n_0 = 1$, $w_0 = 0.5$, $\delta = 1/8h$. Vymezení prohledávaného prostoru D pro jednotlivé úlohy je uvedeno v [10] nebo na webové stránce [11]. Pro algoritmus s jedenácti soutěžícími heuristikami [8] bylo hledání pro šest z osmi testovaných úloh stoprocentně spolehlivé, u zbývajících dvou byla spolehlivost zhruba tříčtvrtinová, ale průměrné počty vyhodnocení účelové funkce byly u některých úloh dosti vysoké, u úlohy Bennet5 dokonce přes 78 tisíc, což na běžném PC vyžaduje několik minut [10]. Algoritmus s osmi soutěžícími heuristikami [9] byl sice rychlejší i spolehlivější, ale u některých úloh ne příliš významně. Zřejmě samotná soutěž heuristik nezaručuje takovou adaptibilitu algoritmu, aby z relativně velkého množství heuristik byly přednostně vybírány ty nejvhodnější. Volba heuristik významně ovlivňuje jak spolehlivost hledání, tak rychlost konvergence. Je známo, že u většiny úloh odhadu parametrů nelineárních regresních modelů algoritmická obtížnost spočívá spíše ve zvládnutí údolí účelové funkce, kde gradient je velmi malý, než v nějaké „divoké“ multimodalitě časté u funkcích užívaných pro testování optimalizačních algoritmů [9]. Proto byly pro další variantu algoritmu zvoleny ty heuristiky, které byly nejčetněji vybírány pro takové funkce (např. Rosenbrockovu funkci) nebo heuristiky, u kterých lze očekávat podobné vlastnosti. Úspěšně fungoval algoritmus CRS se čtyřmi heuristikami. Dvě heuristiky byly typu REFL s parametry $\alpha = 2$, $s = 0.5$ a $\alpha = 5$, $s = 1.5$. Další heuristikou byla REFL-B s parametry $\alpha = 2$, $s = 0.5$, Čtvrtou heuristikou byla DE-ADP s parametry $F_{\min} = 0.4$, $C = 0.9$. Výsledky jsou uvedeny v tabulce 1, označení úloh je shodné s NIST. Ve sloupcích R je uvedena spolehlivost v procentech, s jakou bylo nalezeno řešení blízké certifikovanému globálnímu minimu (shoda v součtu reziduálních čtverců alespoň na 7 platných míst), ve sloupcích označených $\overline{NE}$ je průměrný počet vyhodnocení účelové funkce potřebný k dosažení podmínky ukončení, ve sloupcích *vc* je

koeficient variace vyjádřený v procentech, *suc* je relativní četnost (v procentech) úspěšných bodů $\mathbf{y}$, kdy $f(\mathbf{y}) < f_{\max}$, $R2$ je index determinace, *rst* je průměrný počet resetů na 1000 vyhodnocení účelové funkce (reset se provede, když klesne kterákoli hodnota q_i pod hodnotu δ) a *cpu* je průměrný čas v milisekundách potřebný na jedno vyhodnocení účelové funkce (na PC 667 MHz). Z tabulky 1 vidíme, že zejména algoritmus COMP4 byl vysoce spolehlivý s vcelku přijatelnými časovými nároky. U většiny úloh byl čas na nalezení minima několik sekund, minutu přesáhl jen u úlohy Bennett5. Jelikož jsou to časy pro testovací verzi algoritmu se zaznamenáváním několika veličin pro sledování průběhu vyhledávání, lze efektivnějším naprogramováním dobu výpočtu snížit. Jak ukazuje obrázek 1, kde jsou porovnány dvě úlohy lišící se nejvíce v časové náročnosti, je algoritmus adaptivní v tom ohledu, že relativní četnost heuristik užitých při vyhledávání se přizpůsobuje řešené úloze.

	COMP1			COMP4						
Úloha	R	$\overline{NE}$	*vc*	R	$\overline{NE}$	*vc*	*suc*	1-$R2$	*rst*	*cpu*
Bennett5	100	42367	33.0	100	37620	38.2	44	1.06e-5	12.1	3.23
BoxBOD	100	1023	9.4	100	1018	8.9	51	1.20e-1	21.8	2.65
Eckerle4	100	2089	5.9	100	2118	5.9	50	2.94e-3	18.7	2.70
MGH09	100	9552	10.2	100	9316	9.8	44	5.94e-3	14.8	2.69
MGH10	100	20709	7.6	100	21031	10.3	50	4.70e-8	8.5	2.50
Rat42	100	2357	5.7	100	2368	6.2	44	1.73e-3	16.2	2.51
Rat43	100	3777	4.8	100	3779	5.0	42	8.16e-3	16.5	2.76
Thurber	91	12343	3.7	97	12309	3.7	38	4.92e-4	13.9	3.21

Tabulka 1: CRS se čtyřmi soutěžícími heuristikami.



Obrázek 1: Četnosti využití heuristik.

Algoritmus COMP4 byl dále ověřován experimentálně na 14 úlohách [1] z článku [3] a na ostatních 19 úlohách NIST [6]. Ve srovnání s algoritmem MCRS [3]

[1] Při porovnání těchto testovacích úloh bylo zjištěno, že úloha Model-5 je shodná s úlohou MGH10, odlišné je jen vymezení prohledávaného prostoru D.

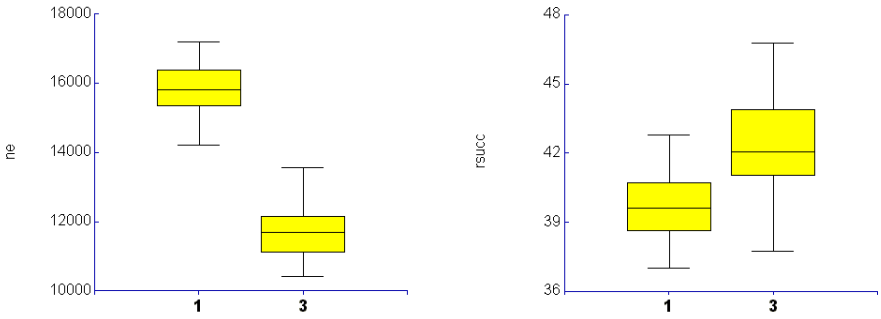
Model	R	$\overline{NE}$	vc	suc	1-$R2$	rst	cpu
1	100	3177	5.2	48	3.45e-3	13.3	2.45
2	80	3038	12.0	47	4.70e-8	12.1	2.46
3	100	1404	6.3	56	8.08e-4	19.8	2.48
4	100	3071	9.2	42	1.62e-3	14.5	2.34
5	100	21590	6.6	47	6.20e-8	11.3	2.37
6	100	1036	8.8	48	3.77e-1	25.0	2.40
7	100	37061	1.9	48	2.38e-2	15.5	2.61
8	100	32349	4.1	47	2.80e-6	15.6	2.73
9	98	2051	13.1	41	3.84e-3	18.4	2.36
10	100	30428	40.6	36	3.21e-5	12.8	2.80
11	85	24485	33.6	35	3.70e-8	11.5	2.56
12	98	2942	5.1	45	1.89e-3	16.3	2.54
13	100	14297	3.9	42	1.11e-5	16.0	3.11
14	100	19483	3.7	49	4.57e-4	13.5	2.47

Tabulka 2: Algoritmus COMP4, úlohy z článku [3].

Úloha	R	$\overline{NE}$	vc	suc	1-$R2$	rst	cpu
chwirut1	100	2423	5.1	52	2.00e-2	16.4	3.34
chwirut2	100	2382	6.0	52	1.40e-2	16.5	2.97
danwood	100	1278	9.0	44	5.67e-4	21.7	2.91
enso	91	19554	12.3	35	4.02e-1	12.6	4.44
hahn1	35	15844	4.9	40	1.96e-4	14.6	4.64
kirby2	100	6834	3.1	40	2.85e-5	14.1	3.44
misra1a	100	1802	8.0	44	1.84e-5	18.3	2.81
misra1b	100	1553	9.3	42	1.12e-5	17.4	2.95
misra1c	100	1743	9.8	44	6.10e-6	18.7	2.94
misra1d	100	1763	9.9	44	8.30e-6	16.6	2.79
nelson	100	5263	5.9	47	6.98e-2	12.7	3.18
roszman1	95	5145	4.5	41	1.59e-3	16.1	2.99

Tabulka 3: Lehké a středně obtížné úlohy NIST, experimentální data.

má COMP4 ve většině úloh menší časové nároky, počet vyhodnocení účelové funkce je zhruba o třetinu menší. Pouze u dvou úloh ze čtrnácti byla časová náročnost u COMP4 vyšší, a to u úlohy 11 (dosažená spolehlivost je ale o 9% vyšší) a u úlohy 7. Relativně nižší spolehlivost nalezení globálního minima u úloh 2 a 11 lze snad zdůvodnit velmi malým reziduálním rozptylem (viz sloupec 1-$R2$). Také pro lehké a středně obtížné úlohy NIST v tabulce 3 jsou výsledky testování algoritmu COMP4 přijatelné s výjimkou úlohy Hahn1, kdy je spolehlivost přes značnou časovou náročnost nízká. Jak ukazuje porovnání úspěšných a neúspěšných vyhledávání minima na ob-



Obrázek 2: Porovnání úspěšných (1) a neúspěšných (3) běhů- Hahn1.

Úloha	*eps*	*R*	$\overline{NE}$	*vc*	*suc*	1-*R*2	*rst*	*cpu*
gauss1	1e-15	55	22436	22.9	42	3.04e-3	13.4	4.09
gauss2	1e-15	18	20540	13.8	42	3.51e-3	14.7	4.23
lanczos1	1e-28	100	37359	14.0	40	1.34e-26	15.9	3.07
lanczos2	1e-16	100	28013	19.6	42	2.10e-12	16.4	3.05
lanczos3	1e-15	100	30323	17.6	43	1.51e-9	16.8	3.08
mgh17	1e-12	100	9235	4.6	42	4.74e-5	15.4	2.90
gauss3	1e-16	0	(21468)	(18.7)	(41)	(2.53e-2)	(14.3)	(4.07)

Tabulka 4: Lehké a středně obtížné úlohy NIST, generovaná data.

rázku 2 (*ne* je počet vyhodnocení účelové funkce, *rsucc* je relativní četnost úspěchu), algoritmus zřejmě často končí prohledávání v lokálním minimu, neboť jsou přednostně vybírány heuristiky s vyšší relativní úspěšností, které neumožňují únik z oblasti předčasné konvergence. Zde tedy adaptivita této varianty algoritmu není dostatečná.

Podobně vypadalo porovnání úspěšných a neúspěšných vyhledávání minima pro úlohu Model-2 s nejnižší dosaženou spolehlivostí z tabulky 2. I zde zřejmě algoritmus nezajišťoval diverzivitu vyhledávání dostatečnou k úniku z oblasti lokálního minima, i když rozdíl v relativní úspěšnosti úspěšných a neúspěšných běhů nebyl tak výrazný jako u úlohy Hahn1. Problematické jsou výsledky testování úloh NIST, kdy data nebyla z experimentů, ale generovaná, viz tabulka 4. Při nastavení hodnot vstupních parametrů algoritmu z předchozích testů nebylo u většiny z těchto úloh nacházeno globální minimum. U třech těchto úloh je to pochopitelné, neboť je u nich extrémně malý reziduální rozptyl a pomohlo zpřísnění podmínky ukončení, viz sloupec *eps* v tabulce. U úloh Gauss1, Gauss2 a Gauss3 ani toto zpřísnění výrazně nepomohlo. Je nutno konstatovat, že algoritmus COMP4 v těchto úlohách selhal, v případě úlohy Gauss3 dokonce nenalezl globální minimum ani v jednom ze sta běhů.

4 Závěr

Přes velmi povzbudivé výsledky na osmi obtížných úlohách NIST nelze po dalším testování algoritmus COMP4 považovat za dostatečně spolehlivý pro odhad parametrů nelineárních regresních modelů metodou nejmenších čtverců. Je však rozhodně spolehlivější než algoritmy ve standardním statistickém softwaru a může být užíván přinejmenším jako alternativní postup. Implementace tohoto algoritmu v Matlabu (prozatím nepříliš uživatelsky přátelská) je přístupná na webové stránce [11].

Reference

[1] Ali M.M., Törn A. (2004). *Population set based global optimization algorithms: Some modifications and numerical studies.* Computers and Operations Research **31**, 1703 – 1725.

[2] Křivý I., Tvrdík J. (1995). *The controlled random search algorithm in optimizing regression models.* Comput. Statist. and Data Anal. **20**, 229 – 234.

[3] Křivý I., Tvrdík J. Krpec, R. (2000). *Stochastic algorithms in nonlinear regression.* Comput. Statist. and Data Anal. **33**, 278 – 290.

[4] Nelder J.A., Mead R. (1964). *A simplex method for function minimization.* Computer J. **7**, 308 – 313.

[5] Price W. L. (1977). *A controlled random search procedure for global optimization.* Computer J. **20**, 367 – 370.

[6] Statistical Reference Datasets. *Nonlinear regression.* NIST Information Technology Laboratory. `http://www.itl.nist.gov/div898/strd/`. December 1, 2001.

[7] Storn R., Price K. (1997). *Differential evolution – a simple and efficient heuristic for global optimization.* J. Global Optimization **11**, 341 – 359.

[8] Tvrdík J., Mišík L., Křivý I. (2002). *Competing heuristics in evolutionary algorithms.* Intelligent Technologies - Theory and Applications, IOS Press, Amsterdam, 159 – 165.

[9] Tvrdík J. (2004). *Generalized controlled random search and competing heuristics.* MENDEL 2004, 10th International Conference on Soft Computing (Matoušek R. and Ošmera P. eds). University of Technology, Brno, 228 – 233, 2004.

[10] Tvrdík J., Křivý I. (2004). *Comparison of algorithms for nonlinear regression estimates.* COMPSTAT 2004 (J. Antoch ed.), Physica-Verlag, 1917 – 1924.

[11] `http://albert.osu.cz/tvrdik/`

Poděkování: Tento příspěvek byl podporován z institucionálního výzkumného záměru J09/98:179000002.

Adresa: J. Tvrdík, Přírodovědecká fakulta OU, 30. dubna 22, 701 03 Ostrava

E-mail: `tvrdik@osu.cz`