ROBUST 2004

TŘEŠŤ
7. – 11. června 2004

ABSTRAKTY

Barbora Arendacká, ÚM SAV Bratislava

Konfidenčné intervaly pre variančný komponent v modeli s dvomi variančnými komponentami.

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, ktoré je možné riešiť pomocou zovšeobecnených p-hodnôt, resp. zovšeobecnených testovacích premenných (pojmy zaviedli Tsui a Weerahandi (1989) a ďalej rozpracoval Weerahandi (1995)). Problémom, ktorý pri tomto prístupe vyvstáva, je nejednoznačné určenie použiteľnej zovšeobecnenej testovacej premennej v prípade, keď matica W má viac ako dve rôzne vlastné hodnoty. Túto nejednoznačnosť ilustrovali Zhou a Mathew (1994) na príklade testovacej premennej závislej na kladných konštantách c_i a vo svojom článku konštatujú, že nie je jasné, ktorú testovaciu premennú v tomto prípade preferovať. Ako ukážeme, pri voľbe c_i je možné sa inšpirovať známymi štatistikami (napr. Michalski (2003)) používanými na testovanie nulovosti σ_1^2 . Ďalej uvedieme testovaciu premennú, ktorá nie je špeciálnym prípadom testovacej premennej Zhoua a Mathewa pre žiadnu kombináciu c_i a jednotlivé testovacie premenné navzájom porovnáme.

Literatura:

Michalski A. (2003). *On some aspects of the optimal statistical inference on variance components in mixed linear models*. Tatra Mt. Math. Publ. 26, 133 – 153.

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.

Weerahandi S. (1995). *Exact Statistical Methods for Data Analysis*. Springer-Verlag, New York.

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

Lucie Belzová, KPMS MFF UK Praha

Inference založená na sekvenčních pořadích

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

Viktor Beneš, KPMS MFF UK Praha

Časo-prostorové bodové procesy

Přednáška 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.

Martin Betinec, KS FF UK Praha

Poznámky ke shlukové analýze prvoků

Shluková analýza ve formě fylogenetických stromů pronikla i do nejrůznějších odvětví biologie a genetiky. V tomto příspěvku se budeme zabývat nejprve vlivem parametrů shlukové analýzy (tj. volbou kódování nukleotidů, volbou míry nepodobností a typem shlukování) na výslednou klasifikaci prvoků čeledi *Trichomonandinæ*. Dále zmíníme i některé koncepty odhadů spolehlivostí výsledných dendrogramů a jejich výpočetní vlastnosti. V závěru bude shluková analýza konfrontována s metodou hlavních komponent.

Marek Brabec, SZÚ Praha

Regrese s useknutými daty: analýza vývoje tělesné výšky ve Švédsku v 18. a 19. století

V našem příspěvku rozebereme zajímavý příklad analýzy časové řady výšek dospělé populace ve Švédsku v 18. a 19. století. Zaměříme se zejména na periodicitu této řady a ukážeme typické komplikace, s nimiž se analýza založená na historických datech potýká, zejména pak selektivitou výběru pro měření, z něhož jsou data k dispozici, a návazným modelem useknuté regrese, jež problém formalizuje. V tomto kontextu se budeme zabývat i citlivostí maximálně věrohodných odhadů k některým předpokladům, o něž se analýza opírá.

Václav Čapek, KPMS MFF UK Praha

Rozumí si statistika s medicínou?

Příspěvek si klade za úkol seznámit posluchače 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.

Součástí příspěvku bude prezentace některých zajímavých výsledků z oblasti medicíny. Na příkladech bude ukázáno, s jakými problémy se studenti a vědečtí pracovníci – nestatistici – často potýkají, a bude naznačeno, jaká by jim měla být nabídnuta pomoc ze strany statistické odborné veřejnosti.

Tomáš Cipra, KPMS MFF UK Praha

Zajištění v pojišťovnictví a jeho matematické aspekty

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.

Daný příspěvek nejprve představí 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 soustředí 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.

Literatura:
Cipra, T. *Zajištění a přenos rizik v pojišťovnictví.* Grada, Praha 2004.

Petr Dostál, KPMS MFF UK Praha

Asymptotická analýza strategií obchodování s akcií při existenci transakčních nákladů

Uvažujme situaci, kdy investor má možnost investovat své jmění do bezrizikového aktiva úročeného konstantní úrokovou mírou r a do akcie, jejíž tržní cena se chová jako geometrický Brownův pohyb. Omezujeme se na strategie, které udržují pozici investora v mezích intervalu $[\alpha, \beta]$, což znamená, že investor nakupuje akcii, pokud její poměrné zastoupení v portfoliu klesá pod hodnotu α a naopak tuto akcii prodává, pokud pozice investora roste nad hodnotu β. Nákupy a prodeje jsou zatíženy poplatky úměrné velikosti nákupů resp. prodeje. Jako kritérium optimality uvažujeme maximalizaci charakteristik vypovídající o asymptotickém chování střední hodnoty užitku z tržní hodnoty portfolia, přičemž užitková funkce je typu HARA (hyperbolic absolute risk aversion). Tento systém zahrnuje mocninné užitkové funkce a jako singulární případ také logaritmickou užitkovou funkci.

Příspěvek bude ilustrovat hlavní myšlenky postupu směřujícího k výpočtu příslušných charakteristik. V singulárním případě logaritmické užitkové funkce tento postup vede k explicitnímu výsledku v tom smyslu, že je možné najít přímo formuli, která vyjadřuje závislost hledaných charakteristik na parametrech uvažované strategie. V ostatních případech je možné alespoň najít formuli, která příslušnou funkci určuje implicitně.

Běžně používaný přístup k řešení otázky volby portfolia je založen na maximalizaci středního užitku plynoucího ze spotřeby realizované odprodejem části portfolia. Tento přístup vede na tzv. Hamilton-Jacobi-Belmanovu parciální diferenciální rovnici, která se redukuje na obyčejnou diferenciální rovnici 2. řádu, pokud předpokládáme, že užitková funkce pochází ze systému HARA. Tato rovnice se za předpokladu nenulových transakčních nákladů považuje za analyticky neřešitelnou. V případě nulových transakčních nákladů lze příslušnou úlohu tímto přístupem řešit, přičemž výsledek tohoto postupu je identický, co se týká strategie obchodování, s výsledky založenými na maximalizaci charakteristik popisující asymptotické chování středního užitku z hodnoty portfolia. Námi zvolený postup se také opírá o obyčejné diferenciální rovnice 2. řádu, které však analyticky řešitelné jsou, což je klíčový moment celého postupu.

V současné době se od geometrického Brownova pohybu jako modelu pro tržní cenu akcie přechází k modelu, který by se dal označit jako geometrický Lévyho proces. Podstatné je, že i v tomto případě bude možné použít (alespoň myšlenkově) uvedený postup. Bohužel už to pravděpodobně nepůjde bez použití výsledků funkcionální analýzy, což celou analýzu značně komplikuje.

Výhodou tohoto modelu je však to, že výsledné metody budou více robustní, neboť Brownův pohyb bude nahrazen Lévyho procesem.

Zdeněk Fabián, ÚI AV ČR Praha

Core vzdálenosti a testování hypotéz

Ve Sborníku ROBUST 2002 jsem prezentoval neotřelý pohled na strukturu náhodných veličin na neúplném nosiči, intervalu $S \neq R$: lze je chápat jako transformované veličiny X=jS(Y), kde Y je náhodná veličina s nosičem R, zvaná prototyp, a jS: RRS spojité zobrazení, které lze volit tak, aby bylo jednoznačné. Za těžiště rozdělení F veličiny X byl zvolen bod t=jS(m), kde m je těžiště (mód) rozdělení veličiny Y. V práci zavedená core funkce TF(x) veličiny X vyjadřuje relativní citlivost těžiště rozdělení F vůči hodnotě v bodě x. Parametrická rozdělení na S užívaná v praxi, která nemají toto t jako parametr, lze přeparametrizovat; jejich core funkce je pak úměrná skórové funkci pro t.

V letošním příspěvku zavedeme pro dané S pivotní proměnnou uS a ukážeme, že core funkci rozdělení lze obecně definovat jako skórovou funkci prototypu vyjádřeného pomocí uS. Core funkce použijeme ke konstrukci vzdáleností bodů ve výběrovém prostoru a v prostoru pravděpodobnostních měr. První z nich lze využít při testování hypotéz o parametru t, druhá je obecnější a lze ji použít jako určitou obdobu Kullbackovy vzdálenosti.

Lucie Fajfrová, ÚTIA AV ČR Praha

Rychlost konvergence k ekvilibriu systému hromadné obsluhy se stromovou strukturou

Uvažujme systém hromadné obsluhy, v němž stanice obsluhy umístíme do uzlů binárního stromu T o výšce N. Budeme se zajímat o počet zákazníků v jednotlivých frontách a sledovat vývoj celého systému v čase. V terminologii částicových systémů takovýto Markovský proces nazýváme zero range proces s přívlastky „na binárním stromě, v konečném objemu“, které specifikují propojení a počet front.

Stacionární rozdělení (ekvilibrium) tohoto Markovského procesu je součinová míra μ na N^T s geometrickými marginály v případě otevřeného systému, respektive podmíněné rozdělení μ_K míry μ za podmínky, že celkový počet zákazníků v systému je K, v případě uzavřeného systému s K zákazníky. Víme tedy, že přechodové pravděpodobnosti $p_t(x, A)$ za čas t z konfigurace x do množiny A v uzavřeném systému konvergují při $t \to \infty$ k $\mu_K(A)$ pro každé x. Přirozenou otázkou, které se budeme v příspěvku věnovat, je rychlost této

konvergence, s čímž úzce souvisí spektrální vlastnosti matice intenzit Markovského procesu. Jednoduše řečeno, naším úkolem je nalézt tzv. „spectral gap“, rozdíl mezi největším a druhým největším vlastním číslem matice intenzit, anebo alespoň jeho dolní odhad. Samozřejmě budeme chtít vědět, jak se rychlost konvergence vyvíjí při N rostoucím do nekonečna.

Stejný problém pro jinou skupinu částicových systémů je řešen v článku Caputo P. *Spectral gap inequalities in product spaces with conservation laws*, Proceedings of the conference *Stochastic analysis on large scale interacting systems*, 2002.

Marie Forbelská, KAM PřF MU Brno

Klasifikační pravidla pro elipticky vrstevnicová rozdělení

Diskriminační analýza patří mezi klasifikační metody vícerozměrné statistické analýzy. Používá klasifikační postupy, pomocí kterých se objekt popsaný vícerozměrným znakem zařadí do jedné z konečného počtu existujících tříd. Postup klasifikace je založen na určitých předpokladech o vlastnostech klasifikovaných 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 pravidla lze najít i pro mnohem širší třídu vícerozměrných rozdělení, a to pro tzv. elipticky vrstevnicová rozdělení.

Michal Friesl, KM FAV ZČU Plzeň

Neparametrické bayesovské odhady v Koziolově-Greenově modelu náhodného cenzorování

Mezi bayesovskými metodami jsou jako neparametrické označovány ty, které počítají s apriorním rozdělením nikoli pro konečný počet parametrů určitého rozdělení, ale pro obvykle nekonečněrozměrné parametry, jako je např. distribuční funkce. Ferguson (1973) 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 Koziolově-Greenově modelu náhodného cenzorování se předpokládá, že distribuční funkce F dob života a distribuční funkce G dob cenzorování, nezávislých náhodných veličin, jsou ve vztahu $(1 - G) = (1 - F)^\gamma$ s určitou

konstantou $\gamma > 0$. Rozdělení cenzoru je tak prostřednictvím parametru γ přímo svázáno s rozdělením dob života.

V příspěvku uplatníme neparametrický bayesovský přístup (v uvedeném smyslu) na odhadování v Koziolově-Greenově modelu a odvodíme příslušné bayesovské odhady pro dobu života.

Literatura:

Doksum K., 1974. *Tailfree and neutral random probabilities and their posterior distributions*, Ann. Probability 2, 183 – 201.

Ferguson T.S., 1973. *A Bayesian analysis of some nonparametric problems*, Ann. Statist. 1, 209 – 230.

Koziol J.A. and Green S.B., 1976. *A Cramér-von Mises statistic for randomly censored data*, Biometrika 63, 465 – 474.

Salinas-Torres V.H., Pereira C.A.B. and Tiwari R.C., 2002. *Bayesian nonparametric estimation in a series system or a competing-risks model*, J. Nonparametr. Stat. 14, 449–458.

Sethuraman J., 1994. *A constructive definition of Dirichlet priors*, Statist. Sinica 4, 639 – 650.

Walker S. and Muliere P., 1997. *Beta-Stacy processes and a generalization of the Pólya-urn scheme*, Ann. Statist. 25, 1762 – 1780.

Eva Gelnarová, KPMS MFF UK Praha

Predikce regrese rakovinného bujení po radikální prostatektomii

Sledujeme pacienty, kterým byla diagnostikována rakovina prostaty a kteří se podrobili radikální prostatektomii. I přes radikální léčbu může dojít k znovupropuknutí, regresi, rakovinného bujení. Bude prezentována metoda umožňující, na základě dostupných údajů o pacientovi (v okamžiku operace či krátce po ní), predikovat regresi rakovinného bujení. Metoda je demonstrována na konkrétních datech.

Marek Hanyš, KS VŠE Praha

Určení optimálního rozsahu výběrového souboru pro vytváření klasifikačního modelu

Některé metody dolování dat, jejichž cílem je klasifikace či predikce, vyžadují, aby byl vstupní datový soubor rozdělen na dvě množiny dat, a to trénovací množinu a testovací množinu. Obě obsahují stejné proměnné, přičemž jedna množina dat je používána pro odhad modelu (trénovací) a druhá pro jeho testování (testovací). V tomto příspěvku soustředím pozornost na postupy,

které se zaměřují na nalezení takové velikosti množiny trénovacích dat, která je optimální ze dvou hledisek, a to přesnosti klasifikačního modelu a nákladů vzniklých v souvislosti s tvorbou modelu.

Zdeněk Hlávka, KPMS MFF UK Praha

Estimation of State Price Densities

The fair price of European option with payoff $(S_T - K)_+ = \max(S_T - K, 0)$, with S_T denoting the price of the stock at time T, K the strike price, and r the risk free interest rate, can be written as:

$$C_t(K,T) = \exp\{-r(T-t)\} \int_0^{+\infty} (S_T - K)_+ f(S_T) dS_T,$$

i.e., as the discounted expected value of the payoff with respect to the so-called state price density $f(S_T)$. The state price density (SPD) bears important information on the behaviour and expectations of the market.

Prices $C_t(K,T)$ of European options with strike price K observed at time t and expiring at time T allow to deduce the state price density in the following form Breeden and Litzenberger:

$$f(K) = \exp\{r(T-t)\} \frac{\partial^2 C_t(K,T)}{\partial K^2}.$$

Kernel smoothers were in this framework proposed and successfully applied by, e.g., Sahalia and Duarte. Another, more sophisticated approach based on nonparametric least squares which allows to include the required constraints is described and applied on simulated data in Yatchew and Härdle.

Using nonlinear least squares, we will construct a simple estimate of the state price density satisfying all of the shape constraints which follow from the theoretical properties (no-arbitrage assumptions). The method is then applied to the DAX option prices observed in 1995.

Literatura:

Sahalia Y., Duarte, J. *Nonparametric Option Pricing under Shape Restrictions*. Journal of Econometrics, 116, 9 – 47.

Breeden D., Litzenberger, R. *Prices of state-contingent claims implicit in option prices*. Journal of Business 51, 621 – 651.

Yatchew A., Härdle W. *Nonparametric state price density estimation using constrained least squares and the bootstrap*. Journal of Econometrics, to appear.

Daniel Hlubinka, KPMS MFF UK Praha

Stereologický problém extrémů; sféroidy

V příspěvku se zaměříme na problém odhadu extrémní hodnoty tvaru či velikosti sféroidu pozorovaného pouze pomocí řezu. Tak by se ve stručnosti dala charakterizovat celá práce. Předpokládáme, že v nějakém neprůhledném materiálu (typicky kovu) se vyskytují drobné částice ve tvaru zploštělých sféroidů. Tyto mohou způsobit v materiálu praskliny či jiné vady a pravděpodobnost, že se tak stane, vzrůstá s velikostí a zploštěností částice.

Budeme předpokládat, že částice je sféroid se dvěma stejně dlouhými hlavními poloosami a jednou menší poloosou. Pak lze takový sféroid charakterizovat *velikostí*, což je délka hlavní poloosy x, a *tvarem* $t = x^2/v^2 - 1$, kde v je délka vedlejší poloosy. Tvar $t = 0$ odpovídá kouli. Nechť velikost a tvar částice jsou náhodné veličiny, které nezávisejí na orientaci a umístění částice. Proces umístění a orientace částice předpokládáme izotropní.

Proveďme náhodný řez materiálem. To, co uvidíme z částic jsou elipsy s velikostí (délkou hlavní poloosy) y a tvarem (obdobně zavedeným) z. Tvar $t = 0$ si zřejmě vynucuje $z = 0$, a obecně platí $t \geq z$ a $x \geq y$.

Nejprve nás zajímá chování výběrových extrémů velikosti a tvaru elips (pozorovaných profilů), předpokládáme-li nějaké chování chvostů pro velikost a tvar sféroidů. Ukážeme si, že za podmínek určité stejnoměrnosti vynucené vztahy $t \geq z$ a $x \geq y$ se sféroidy i elipsy nacházejí ve stejné oblasti přitažlivosti. Od tohoto zjištění je již jen krok k nějakému uvažovanému chování chvostů a odhadu extrémních hodnot sféroidů i elips. Naším dalším úkolem tedy je hledat vhodné modely splňující zmíněnou podmínku stejnoměrnosti a z největších pozorovaných hodnot u elips přejít k odhadům největších hodnot sféroidů. K tomuto účelu použijeme normovací konstanty získané z modelu pro chvosty rozdělení velikosti a tvaru sféroidu.

Tato práce vznikla za podpory grantu GAČR 201/03/0946 *Modely stochastické geometrie a prostorová statistika* a výzkumného záměru MŠMT ČR MSM 113200008 *Matematické metody ve stochastice*.

Klára Hornišová, ÚM SAV Bratislava

Aproximácie nelineárnych regresných modelov metódou hlavných komponentov

Všeobecný nelineárny regresný model možno aproximovať modelom z užšej triedy, ktorej vlastnosti sú lepšie preskúmané. Pre prípad, že je známe apriórne rozdelenie neznámeho parametra strednej hodnoty, v článku Pázman

(2001) sa navrhovalo merať optimalitu takej aproximácie jej apriórnou strednou kvadratickou chybou. Tamže sa aproximácia odvodila pre triedu lineárnych modelov, resp. triedu lineárnych kombinácií daných funkcií parametra. V triede vnútorne lineárnych modelov sa dá riešenie vyjadriť pomocou hlavných komponentov náhodného (vzhľadom na apriórne rozdelenie) vektora η - bodu z plochy stredných hodnôt pôvodného modelu, viz Hornišová (2004). V špeciálnych prípadoch (analogicky ako v článku El-Shaarawi a Shah, 1980) možno metódu použiť aj na aproximáciu podmodelov, ktoré neobsahujú rušivé parametre. Uvedieme príklady inferencie založenej na tejto metóde a spomenieme možné rozšírenia na ďalšie triedy aproximujúcich modelov.

Literatúra:
El-Shaarawi A. a Shah K.R., 1980. *Interval estimation in non linear models*. Sankhya B 42, 229 – 232.
Pázman A., 2001. *Linearization of nonlinear regression models by smoothing*. Tatra Mt. Math. Publ. 22, 13 – 25.
Hornišová K. 2004. *Intrinsic linearization of nonlinear regression by principal components method*. AMUC, to appear.

Ivana Horová a Jiří Zelinka, KAM PřF MU Brno

Odhady rizikové funkce

Cílem příspěvku je prezentovat neparametrickou metodu pro cenzorovaná data. Budeme se zabývat modelem, ve kterém jsou data cenzorovaná zprava. S tímto typem dat se setkáváme v mnoha aplikacích, zejména v klinickém výzkumu.

V roce 1958 navrhli Kaplan a Meier odhad funkce přežití. V tomto příspěvku se zaměříme na jádrové odhady rizikové funkce a její druhé derivace. Jádrové odhady patří mezi efektivní neparametrické odhady. Tyto odhady závisí na vyhlazovacím parametru jádra a řádu jádra. Pojednáme zejména o volbě vyhlazovacího parametru. Uvedené odhady jsou aplikovány na onkologická data, která byla poskytnuta Masarykovým onkologickým ústavem v Brně.

Dušan Húsek a Hana Řezanková, ÚI AV ČR a KS VŠE Praha

Shlukování a textové dokumenty

Každý textový dokument může být reprezentován vektorem, jehož prvky charakterizují výskyt slov (resp. termů) obsažených v dokumentech. 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 $n \times m$, 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).

Základní úlohou při analýze takových dat je shlukování dokumentů. Pomocí hierarchické shlukové analýzy lze nalézt různé úrovně skupin dokumentů. Na základě zjištěných skupin mohou být 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 dimenze, jednak speciální postupy pro shlukování. Například shlukování náhodně vybraných dokumentů, opakované pro různé výběry, může vést ke stanovení množiny slov, která je vhodná pro charakterizování sledované kolekce dokumentů (viz [5]). 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 [3] je navrženo použít metodu k-průměrů a z každého vytvořeného shluku vybrat dokument, který je nejbližší centroidu.

Literatura:

[1] Ding C., He X.. *Cluster Structure of K-means Clustering via Principal Component Analysis*. PAKDD 2004, LNAI 3056, Springer-Verlag, Berlin, 2004, 414 – 418.

[2] Dobrynin V., Patterson D., Rooney N. *Contextual Document Clustering*. ECIR 2004, LNCS 2997, Springer-Verlag, Berlin, 2004, 167 – 180.

[3] Kang J., Ryu K.R., Kwon H. *Using Cluster-Based Sampling to Select Initial Training Set for Active Learning in Text Classification*. PAKDD 2004, LNAI 3056, Springer-Verlag, Berlin, 2004, 384 – 388.

[4] Mylonas P., Wallace M., Kollias S. *Using k-Nearest Neighbor and Feature Selection as an Improvement to Hierarchical Clustering*. SETN 2004, LNAI 3025, Springer-Verlag, Berlin, 2004, 191 – 200.

[5] Volk D., Stepanov M.G. *Resampling Methods for Document Clustering*. V tisku.

[6] Zhang Y., Zincir-Heywood N., Milios E. *Term-Based Clustering and Summarization of Web Page Collections*. Canadian AI 2004, LNAI 3060, Springer-Verlag, Berlin, 2004, 60 – 74.

Jana Husová, KPMS MFF UK Praha

Slabá konvergence suprema náhodných procesů

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 $\mathcal{A} \subset \mathcal{P}([\prime, \infty])$ procesy $(Y_n(A), A \in \mathcal{A})$ konvergují v distribuci. Prvky procesů $(X_n(t), t \in T)$ uvažujeme jako funkce v $C(T)$, $l^{+\infty}(T)$, $D(T)$.

Marie Hušková, KPMS MFF UK Praha

Alternativní hodnocení postupů ve statistické kontrole jakosti

Při statistické kontrole jakosti je obvykle kvalita rozhodovacího postupu na základě tzv. ARL (average run length), tj. průměrného zpoždění mezi dobou změny a dobou jejího odhalení. V příspěvku bude diskutováno kritérium založené na kvantilech zpoždění. Přístup bude ilustrován na situaci, kdy proces je pod kontrolou, jestliže sledovaná charakteristika je pod tzv. prahovou hodnotou a proces není pod kontrolou, jestliže sledovaná charakteristika tuto hodnotu překročí.

Martin Janžura a Jan Nielsen, ÚTIA AV ČR Praha

Některé příklady využití metody simulovaného žíhání ve statistických úlohách

Mnoho úloh statistického zpracování dat, počínaje odhadem parametru, přes výběr modelu, klasifikaci, až po např. filtraci signálu a rekonstrukci obrazu vede přirozeným způsobem na řešení optimalizační úlohy ve tvaru

$$F(Model) = Dist(Data, Model) + Pen(Model),$$

kde $Dist$ je vhodná míra vzdálenosti a Pen je penalta odrážející nějakou apriorní informaci o neznámém modelu, který zde považujeme za hledanou proměnnou patřící do nějaké vhodné třídy přípustných modelů. Pokud sledujeme bayesovský přístup a metodu maximální aposteriorní pravděpodobnosti, obdržíme první člen (až na konstantu) jako logaritmus podmíněného rozdělení a druhý jako logaritmus apriorního rozdělení. K obdobné formulaci dospějeme současně i racionální ad hoc úvahou, neboť naším cílem vždy nutně je,

aby identifikovaný model dobře vysvětloval pozorovaná data a přitom nebyl např. neúměrně složitý.

Numerické řešení takovéto úlohy pak samozřejmě závisí na konkrétním tvaru optimalizované funkce F a na struktuře množiny přípustných řešení. Pokud je tato množina sice diskrétní, ale poměrně veliká (typicky mnohorozměrná), není možná obvyklá iterační optimalizační technika a prosté ani sofistikovanější metody prohledávání nebývají z časových důvodů reálné. Je však možné použít metodu *simulovaného žíhání*, což je simulační metoda typu Markov Chain Monte Carlo a probíhá v podstatě jako cílené znáhodněné prohledávání.

Široké možnosti fungování celého postupu bude dokumentováno na několika příkladech. Nejprve bude připomenuta úloha proložení časové řady funkcí po částech konstantní (viz Janžura a Nielsen, 2002). Následně bude tato metoda modifikována na úlohu s funkcemi po částech lineárními. Nakonec bude předvedena aplikace metody na problém učení bayesovských sítí ze statistických dat.

Literatura:

Janžura M. a Nielsen J. (2002). *Metoda segmentace v „change-point" problému*. In: ROBUST'02 (Antoch J., Dohnal G. a Klaschka J. ed.). JČMF, Praha, pp. 163 – 177.

Daniela Jarušková, KM FSV ČVUT Praha

Extrémy gaussovských posloupností a procesů

Přednáška se zabývá asymptotickou teorií extrému gaussovských posloupností při zvětšujícím se počtu pozorování. První část je věnována limitnímu chování maxima stacionárních gaussovských posloupností při $n \to \infty$.

V druhé části se studují vlastnosti gaussovských procesů se spojitými trajektoriemi. Zvláštní pozornost je věnována derivovatelným procesům, kde je možno odvodit explicitní formuli pro střední počet překročení úrovně (Rieszova věta). Pro procesy s derivovatelnými i nederivovatelnými trajektoriemi bude uvedeno limitní chování maxima na pevném intervalu při rostoucí mezi překročení.

V třetí části bude naznačeno jak se teorie extrému použije při hledání asymptotického rozdělení testových statistik v teorii detekce bodu změny. Problematika bude ilustrována na řade příkladů se simulovanými i reálnými daty.

Jan Kalina, KPMS MFF UK Praha a Universität Duisburg–Essen

Analysis of a human face

The goal of this very applied work is to detect objects in pictures of human faces. To locate eyes, we use one or more eye templates and look for the position, scale and rotation, which lead to the best fit. Based on the position of the eyes, we find other facial features.

We try to construct the computer program to be rotation-invariant, which is the main advantage over existing face detection algorithms, usually based on neural networks or support vector machines.

A special attention is paid to preliminary transformations of the data, namely to removing noise and possible outliers. We compare the performance of several robust estimation methods.

This work has been supported by the grant GAČR 402/03/0084 of the Grant Agency of the Czech Republic.

Arnošt Komárek a Emmanuel Lesaffre

Log-lineární regresní model pro cenzorovaná data s vyhlazeným rozdělením chyb odhadovaný metodou penalizované maximální věrohodnosti

Nejčastěji používaným modelem pro regresní analýzu cenzorovaných dat (typicky časů do nějaké události) je bezesporu Coxův regresní model, který specifikuje logaritmus rizikové funkce jako lineární funkci vysvětlujících proměnných. Méně používanou, avšak v některých situacích vhodnější alternativou je tzv. AFT (*accelerated failure time*) model. Užitím tohoto přístupu získáme regresní model pro logaritmus časů do dané události, t.j.

$$\log(T_i) = \alpha + \beta^T x_i + \sigma\, \varepsilon_i, \qquad i\ = 1, \ldots, n,$$

kde T_i je čas do sledované události pro i-tou jednotku datového souboru, x_i je vektor vysvětlujících proměnných pro i-tou jednotku, α a β jsou regresní parametry, σ parametr měřítka a ε_i náhodná veličina s hustotou $f(e)$.

Výhodou prvně zmíněné Coxovy regrese je fakt, že není třeba specifikovat rozdělení časů do události (model je tzv. semiparametrický). AFT model se stejnou vlastností byl sice studován několika autory již od konce 70. let minulého století, avšak žádná z navržených metod dlouho (do roku 2003, viz Jin, Lin, Wei, Ying, *Biometrika*) nenabízela snadno aplikovatelnou či za rozumných předpokladů vždy konvergující metodu pro získání odhadů.

Dalším problémem komplikujícím život je případný výskyt intervalově cenzorovaných dat, s kterými se nutně setkáme v okamžiku kdy je čas do sledované události zjištěn bud' laboratorním testem (např. HIV test) či kontrolou odborníka (např. zubaře, jedná-li se o dobu do tvorby zubního kazu). V této situaci jsou obdržená data tvaru $(T_i^S, T_i^H]$, přičemž o skutečném času do sledované události T_i je pouze známo $T_i^S < T_i \leq T_i^H$. Vzhledem k tomu, ze výše uvedené semiparametrické přístupy jsou založeny na pořadích časů T_i, představuje intervalové cenzorování dosti složitý problém.

Příspěvek pojedná o metodě, která specifikuje rozdělení náhodné chyby ε_i z modelu jako směs vyššího, předem daného počtu normálních rozdělení s pevnými středními hodnotami a konstantním rozptylem, t.j.

$$\varepsilon_i \sim \sum_{j=1}^{k} w_j N(\mu_j, \sigma_0^2),$$

motivovanou vyhlazováním křivek pomocí B-splinů (viz Eilers a Marx, 1996, *Statistical Science*). Neznámé váhy $w_1, \ldots, w_k$, regresní parametry α a β a parametr měřítka σ jsou odhadovány maximalizováním penalizované věrohodnosti.

Užitím navržené metody nepřináší intervalové cenzorování žádné další komplikace v porovnání s obvyklejším cenzorováním zprava či zleva. Navíc, i když je navržený přístup ve své podstatě parametrický, skutečnost, že libovolné spojité rozdělení lze za mírných předpokladů libovolně přesně aproximovat normální směsí znamená, že prakticky nečiníme žádné předpoklady ohledně rozdělení náhodné chyby ε_i.

Lenka Komárková, FM VŠE Jindřichův Hradec

MOSUM-type tests for a change-point problem with censored data

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 changes 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. Particularly, under equal censorship, the permutation principle can be used. The consistency of the test procedure is also studied. Theoretical results are accompanied by simulations.

Literatura:
Hušková M., Neuhaus G. (2004). *Change Point Analysis for Censored Data.* Journal of Statistical Planning and Inference, accepted for publication.
Koblížková L. (2002). *Rank Tests for a Change in Censored Data.* In: J. Antoch, Ed., Proceedings ROBUST'02, JČMF, Prague, 178 – 185.

Michala Kotlíková, Hana Mašková, Arnoštka Netrvalová, Pavel Nový, Dagmar Spíralová, František Vávra, David Zmrhal

Informace a dezinformace - statistický pohled

Při práci se statistickými odhady sdílené informace nastávají některé zdánlivě neočekávatelné situace. Protože pravděpodobnosti (hustoty) pro nás nejsou dostupné, pracujeme s jejich odhady nebo s jejich některou parametrickou reprezentací. Situace, kdy nelze předpokládat znalost ani marginálů je nepoměrně komplikovanější. Zde se jako vhodný koncept jeví pojem dezinformace, jímž se ve svém příspěvku budeme zabývat.

Alena Koubková a Jaroslav Král, KSI MFF UK Praha

Pravděpodobnost a matematická statistika v informatických oborech

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, kde je i přes nesporný pokrok stále ještě co zlepšovat. Především je nutné studenty informatiky přesvědčit, že statistika je něco, co ve svém oboru potřebují, což je velmi obtížný úkol.

Vycházíme z vlastních zkušeností získaných působení m na katedře softwarového inženýrství MFF UK. Na několika konkrétních příkladech (analýza algoritmů, experimentální algoritmika, datové inženýrství apod.) se snažíme ukázat, že v informatice je prostor nejen pro elementární počtářskou rutinu, ale i pro náročné výpočty a teoretické úvahy, které samotný informatik bez hlubšího matematického vzdělání jen těžko zvládne. Jsme toho názoru, že je důležité nejen učit informatiky alespoň základům pravděpodobnosti a matematické statistiky, ale rovněž přesvědčit statistiky, že informatika může být pro ně zajímavou a perspektivní aplikací.

Alena Koubková, KPMS MFF UK Praha

A statistical proposal for sequential clinical trials in different cancer locations.

Každý nově vyvinutý lék musí projít mnoha testy, než je možné ho začít používat v běžné lékařské praxi. Nejprve je třeba zjistit řadu jeho vlastností, jako jsou účinnost, jedovatost, nežádoucí účinky a další. Testovací postup k tomuto určený sestává z chemických testů, testů na zvířatech a v poslední fázi i testů na lidech.

Můj příspěvek je zaměřen na plánování rakovinových klinických pokusů druhé fáze. Jde o testy na lidech, kde hlavním úkolem je odhadnout účinnost nových léků proti rakovině. Jelikož různé typy nádorů rakovinového původu (dále jen typy nádorů) mají řadu podobných vlastností, není nijak překvapující, že jeden lék může být účinný proti více těmto nemocem. Před zahájením testů se však většinou neví, proti kterému nádoru bude lék zabírat nejvíce.

V této práci byla navržena metoda určující pořadí typů nádorů, v jakém je nejvýhodnější, aby vstupovaly do klinických pokusů druhé fáze. Hlavní myšlenka je využít dostupné informace o jednotlivých nádorech a také o testovaném léku k tomu, aby byla vybrána rakovina, pro kterou je zisk z provedení klinického pokusu druhé fáze největší. Navržený postup má dvě části, v první se odhadne rozdělení tzv. poměru reakce a ve druhé se pak vyhodnotí rozhodovací funkce. Poměr reakce je pravděpodobnost, že pro náhodně vybraného pacienta s daným typem nádoru bude lék účinný.

Naše metoda využívá následující vlastnosti rakovin a nového léku: podobnost mezi jednotlivými nádory popsaná korelační maticí, podobnost nového léku s již známými léky, nebezpečnost a výskyt jednotlivých typů nádorů, a také případně různé ohodnocení důležitosti úspěšného či neúspěšného léčení pacienta.

RNDr. Milena Kovářová, Botanický ústav AV ČR Třeboň

Projevy globálních změn v Biosférické rezervaci Třeboňsko

V roce 1976 byla v rámci projektu UNESCO Člověk a Biosféra vyhlášena Biosférická rezervace Třeboňsko. Pracovníci Botanického ústavu AV ČR v Třeboni vybudovali v rámci řešení tohoto projektu nedaleko svého pracoviště v přirozeném mokřadním ekosystému ve výtopě rybníka Rožmberk v území zvaném Mokré Louky profesionální meteorologickou stanici, kde byly od roku 1977 sledovány údaje o stavu mikroklimatu na tomto území. Tato dostatečně dlouhá řada klimatologických pozorování je v současné době vzhledem ke specifičnosti oblasti Biosférické rezervace Třeboňsko velice cenným zdrojem

informací o stavu a změnách klimatu na tomto území a může být použita též k objasnění globálních změn jako celku. Po dobu téměř 30 let byly v pravidelných denních intervalech zaznamenávány úhrny denních srážek, teploty půdy v různých hloubkách, hladiny podzemní vody na několika stanovištích, v hodinových intervalech byly zaznamenávány hodnoty teploty vzduchu, relativní vzdušné vlhkosti, globální a difúzní sluneční záření. Data byla uložena do databáze (Klimadata Bot. Inst. Ac. Sci. (2003)) a jsou pro vědecké a výzkumné účely k dispozici v knihovně Botanického ústavu AV ČR v Třeboni.

Data byla částečně vyhodnocena a ukazují na velmi zajímavé výsledky. Je všeobecně známo, že v posledních letech vzrůstá vzdušná teplota. Celosvětový nárůst teploty vzduchu se udává asi o jeden stupeň Celsia za posledních 100 let, přičemž tento nárůst se především v posledních zhruba 10 letech výrazně zrychluje. Přestože na některých místech se může i ochlazovat, všeobecně převažuje trend růstu teplot. Růst teplot v České republice přibližně odpovídá celosvětovému, avšak data z Mokrých Luk jsou zajímavá tím, že růst teplot na tomto specifickém území je mnohem vyšší. Další zajímavý poznatek se týká změn ve srážkové činnosti. Přes vysokou roční variabilitu srážkových úhrnů lze pozorovat, že přestože jsou dlouhodobé srážkové normály prozatím stabilní, mění se vnitřní rozložení srážek. Srážky se více akumulují do určitých období, které jsou střídány obdobími s podnormálními srážkovými úhrny, výrazně vzrůstají hodnoty velkých srážek, což pak může vést, jak jsme viděli v předchozích dvou letech, ke vzniku povodní nebo extrémního sucha.

V prezentaci budou ukázány průběhy klimatologických parametrů, především teploty a srážek, získané analýzou dat z Mokrých Luk a to jak měsíční průměry, tak i hodnoty a změny těchto charakteristik v letech. Pro objasnění měsíčních průběhů teplot a srážkových úhrnů v závislosti na místě budou též ukázány hlavní rozdíly mezi kontinentálním a přímořským charakterem podnebí.

David Kraus, KPMS MFF UK Praha

Testování dobré shody v Cox–Aalenově modelu

Cox–Aalenův regresní model pro intensity čítacích procesů, který navrhli Scheike a Zhang, rozšiřuje Coxův proporcionální model a Aalenův aditivní model. Zabýváme se testováním dobré shody tohoto modelu. Uvažujeme test založený na stratifikovaném martingalovém residuálním procesu. Asymptotické rozdělení tohoto procesu je (až na zvláštní případy) komplikované (gaussovský proces se složitou kovarianční strukturou). Není proto možné přímo použít například test typu Kolmogorova–Smirnova. Ukazujeme dvě možnosti

řešení tohoto problému. První možností je simulovat realisace z limitního rozdělení residuálního procesu, jak to v případě Coxova modelu navrhli Wei a Ying. Na základě tohoto výběru lze posoudit (graficky či numericky) neobvyklost pozorované trajektorie. Druhá varianta spočívá v transformování (kompensování) limitního procesu na gaussovský martingal podobně jako Khmaladze. Statistika typu Kolmogorova–Smirnova založená na transformovaném procesu již má rozdělení suprema Wienerova procesu.

Literatura:
Khmaladze E.V. (1981). *Martingale approach in the theory of goodness-of-fit tests*. Theory Probab. Appl. 26, 240 – 257.
Lin D.Y., Wei L.J. a Ying Z. (1993). *Checking the Cox model with cumulative sums of martingale-based residuals*. Biometrika 80, 557 – 572.
Scheike, T. H. & Zhang, M.-J. (2002). *An additive-multiplicative Cox–Aalen regression model*. Scand. J. Statist. 29, 75 – 88.

Michal Kulich, KPMS MFF UK Praha

Odhadování regresních parametrů s neúplnými daty

Vyjdeme z obecné teorie odhadu regresních koeficientů za přítomnosti chybějících nebo nepřesně změřených hodnot v závisle nebo nezávisle proměnných, jež byla navržena v článku Robins, Rotnitzky a Zhao (1994). Obecné výsledky těchto autorů dávají návod, jak odvozovat odhady regresních koeficientů za neúplných dat, ať už byla jejich neúplnost plánována nebo vznikla nahodile. Tyto odhady nekladou žádné předpoklady na populační rozdělení regresorů a jsou za určitých podmínek semiparametricky eficientní. Ukážeme, jak tuto teorii použít pro celou škálu praktických problémů, jak odhadovat jejich rozptyl a testovat hypotézy o regresních parametrech.

Pavla Kunderová, PřF UP Olomouc

Lineární regresní modely s rušivými parametry

Přednáška bude věnována regresním modelům, ve kterých je vektor parametrů prvního řádu rozdělen na dva podvektory: na vektor užitečných parametrů a na vektor parametrů rušivých. Existují zde dva základní přístupy.

První (tzv. strukturální) přístup respektuje strukturu modelu a hledá třídy 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 v obou modelech (v modelu s rušivými parametry i v modelu, kde rušivé parametry neuvažujeme).

Ve druhém (tzv. eliminačním) přístupu hledáme takové transformace původního modelu, které eliminují rušivé parametry. Taková transformace ale nesmí způsobit ztrátu informace o užitečných parametrech, tj. nový model musí umožňovat stejně kvalitní odhady užitečných parametrů jako model původní.

Bude uveden přehled dosažených výsledků a ukázány směry, ve kterých jsou ještě neřešené problémy.

Petr Lachout, KPMS MFF UK Praha

Data, rozdělení pravděpodobnosti, asymptotika

Při statistickém odhadování parametrů a testování hypotéz používáme ukazatele, statistiky založené na empirických pozorováních. Tyto statistiky jsou povětšinou založeny na empirickém rozdělení určeném danými pozorováními. Příspěvek se bude zabývat precizací pojmu statistiky a bude pojednávat o jejích asymptotických vlastnostech.

Cílem příspěvku je zdůraznit, které vlastnosti jsou určující pro asymptotické vlastnosti dané statistiky. Pokud jsou pozorování realizací náhodných veličin, které jsou i.i.d., silně stacionární, mixující, či svázané jiným vhodným modelem, pak víme, že naše pozorování mají potřebné vlastnosti s pravděpodobností jedna. To nám umožňuje studovat konzistenci, řád konzistence a asymptotickou normalitu daného odhadu, či konstruovat statistický test a diskutovat jeho sílu.

Jaroslav Marek and Eva Fišerová, KMAAM UP Olomouc

Statistical analysis of geodetical measurements

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 mantioned 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, i.e., it is the same as for the first kind of the model. The inaccuaracy in the second stage we will be called uncertainty of the type A.

The aim of the contribution is to compare standard estimators, H-optimum estimators in the twostage model and standard estimators by using the replication of the model in the second stage. It will be shown where it is better to reduce the influence of the uncertainty of the type B on the uncertainty of the type A by using H–optimum estimators and where it is meaningful to reduce the uncertainty of the type A by using the replication of the model.

Jan Nielsen, ÚTIA AV ČR Praha

Stochastické metody zpracování obrazu

Hlavním cílem příspěvku je představit algoritmus klasifikace textur v obrazu. Klasifikace je proces, jehož úkolem je nalézt v obrazu homogenní části textur, přičemž předem známe parametry modelů v obrázku hledaných textur. Texturou budeme chápat náhodnou realizaci z gaussovského-markovského náhodného pole.

Metodu samotnou provádíme užitím Bayesovy věty metodou maximální aposteriorní pravděpodobnosti. Metoda oproti běžným postupům, kdy se znáhodňuje oštítkování (přiřazení textury) jednotlivých bodů obrazu, znáhodňuje interakce mezi obrazovými body. Definujeme apriorní rozdělení modifikací Pottsova rozdělení, které respektuje typické rozložení textur v obrázku a redukuje některé nedostatky klasického Pottsova rozdělení. Jelikož se jedná o složitý výpočetní problém, odhad výsledku provedeme užitím oblíbené metody simulovaného žíhání za použití Metropolisova sampleru (viz [1]).

V případě, že předem neznáme parametry modelů hledaných textur, provedeme odhad těchto modelů. Tento proces se nazývá analýza obrazu. Ukážeme si tedy algoritmus analýzy obrazu, ve kterém nejprve obrázek rozdělíme na disjunktní části dostatečně veliké pro provedení odhadu a současně dostatečně malé, abychom mohli předpokládat, že se v dané části vyskytuje právě jedna textura. V každé takto vytvořené oblasti metodou maximální pseudověrohodnosti (viz [2]) získáme odhad modelu textury v oblasti. Takto získané modely následně použitím hiearchického shlukování (viz [3]) redukujeme na několik výrazných reprezentantů modelů textur vyskytujících se v daném obrázku. Nad takto získanou množinou provedeme výše popsaný algoritmus klasifikace.

Nakonec si ukážeme několik příkladů analýzy i klasifikace a jejich citlivosti na vstupní parametry.

Literatura:

[1] Winkler G.: *Image Analysis, Random Fields and Dynamic Monte Carlo Methods*. Springer, 1995.

[2] Janžura M.: *Asymptotic properties of the maximum pseudolikelihood estimate of Gauss-Markov random fields*. ÚTIA AV ČR, Praha, 1999, 18 s. (Research Report 1949).
[3] Everitt B.S., Landau S., Leese M.: *Cluster Analysis*. Edward Arnold; 4th edition (May 2001).

Petr Novotný, KPMS UK MFF Praha

Optimální přistup k segmentaci dat

Úkolem je proložit vektor s mnoha pozorováními po částech spojitou regresní funkcí tak, aby regresní model v daném úseku spojitosti závisel pouze na datech v tomto úseku. Dále požadujeme, aby počet bodů nespojitosti byl nejvýše $K-1$. Přitom chceme najít dělení, které minimlizuje zvolenou ztrátovou funkci, například reziduální součet čtverců.

Klasický algoritmus založený na dynamickém programování má paměťovou náročnost $O(N^2)$, kde N je délka celého vektoru. Mnou navržený algoritmus při yachování stejného počtu operací potřebuje pouze $(2 \times K + 1) \times N$ čísel v paměti. Je-li regresní funkce po částech konstantní, lze snížit časovou složitost z $O(N^3)$ na $O(N^2)$.

Marek Omelka, KPMS UK MFF Praha

Test hypotézy úplné specifikace normálního rozdělení

Pro výběr $X_1, \ldots, X_n$ z normálního rozdělení testujeme hypotézu $(\mu, \sigma^2) = (0, 1)$ proti alternativě $(\mu, \sigma^2) \neq (0, 1)$. Vedle třech klasických testů, tj. testu poměrem věrohodnosti, testu Waldova typu a Raova skórového testu, budeme uvažovat také další speciální testy, které byly pro tento problém navrženy v literatuře. Z těchto testů nás bude zajímat zejména lokálně nestranný test, který lokálně maximalizuje průměrnou sílu, dále Isaacsonova aproximace testu typu D a nakonec test založený na Fischerově kombinaci nezávislých statistik. U všech výše zmíněných testů budeme zkoumat jak lokální, tak globální vlastnosti. Numerická studie nám mimo jiné ukazuje, že požadovat od testů lokální nestrannost je opravdu rozumné a že koncept Bahadurovy eficience se jeví pro globální hodnocení testů jako vhodnější než koncept lokální optimality.

Jan Picek, KAM TUL Liberec

Odhady a testy Paretova indexu

Nechť $X_1,\ X_2, \ldots$ jsou nezávislé stejně rozdělené náhodné veličiny s distribuční funkcí F. Zajímá nás chování maxima $M_n = \max(X_1, \ldots, X_n)$. Podle

Fisherovy-Tippettovy věty dostaneme, že jestli vhodně standardizovaná maxima konvergují v distribuci k nedegenerované limitě, potom limitní rozdělení musí být rozdělení extrémních hodnot $G(x)$.

Podle hodnot parametru γ (někdy zvaný Paretův index) rozděluje distribuční funkce do tří tříd: Fréchetova ($\gamma > 0$), Gumbelova ($\gamma = 0$) a Weibullova ($\gamma < 0$).

V příspěvku se zaměříme na odhady parametru γ (především pro Fréchetovu třídu) a testy o parametru γ, zvláště se bude jednat testy hypotézy $\gamma = 0$ proti alternativě $\gamma > 0$.

Literatura:

Embrechts P., Klüppelberg C. and Mikosch T. (1997). *Modelling Extremal Events for Insurance and Finance*. Springer.

Hasofer A.M. and Wang Z. (1992). *A test for extreme value domain of attraction*. JASA, 87, 171 – 177.

Hill B.M. (1975). *A simple general approach to inference about the tail of a distribution*. Ann. Statist. 3, 1163 – 1174.

Jurečková J., Picek J. (2001).*A class of tests on the tail index*. Extremes 4, 165 – 183.

Neves C., Picek J., Fraga Alves M. I. (2003). *A Test for Gumbel Domain of attraction – ratio of maximum and mean of excesses*. Submitted.

Pickands J. III (1975). *Statistical inference using extreme order statistics*. Ann. Statist. 3, 119 – 131.

Segers J. and Teugels J. (2001). *Testing the Gumbel hypothesis by Galton's ratio*. Extremes 3, 291 – 303.

Pavel Plát, FJFI ČVUT

Nejmenší vážené čtverce a heteroskedasticita disturbancí

Znalost asymptotické normality a asymptotické reprezentace odhadu regresních koeficientů v lineární regresním modelu metodou nejmenších vážených čtverců (LWS) nám umožňuje využít myšlenku H. Whita a získat tak pro nejmenší vážené čtverce modifikaci Whiteova testu homoskedasticity disturbancí. Numerický příklad pak prezentuje výsledky získané s využitím LWS při zpracování dat s heteroskedastickými disturbancemi.

Research was supported by grant of GA ČR no. 402/03/0084.

Literatura:

Mašíček L. (2003). *Diagnostika a senzitivita robustních modelů*. PhD disertace, MFF UK, Praha.

Plát P. (2003). *Odhad metodou nejmeníchch vážených čtverců*. Diplomová práce, FJFI ČVUT, Praha.

Rousseeuw P.J. a Leroy, A.M. (1987). *Robust Regression and Outlier Detection*. J. Wiley & Sons, New York.

White H. (1980). *A heteroskesticity-consistent covariance matrix estimator and a direct test for heteroskedasticity*. Econometrica 48, 817 – 838.

Zuzana Prášková, KPMS UK MFF Praha

Metoda bootstrap - 25 let

V letošním roce je to právě 25 let, kdy byl publikován dnes již proslavený Efronův článek o metodě bootstrap (*B. Efron, Bootstrap methods: Another look at the jackknife, Ann. Statist. 7 (1979), 1 – 26).* Těchto 25 let prokázalo, že bootstrap, využívající intenzívním způsobem počítačové algoritmy, je velmi účinným nástrojem, který umožňuje dělat statistické závěry i v těch situacích, kdy řešení statistických problémů ryze matematickými prostředky je obtížné či dokonce nemožné.

V naší přednášce vysvětlíme základní principy této metody, její přednosti i nedostatky, budeme se zabývat problémem konzistence a vyšetříme asymptotické vlastnosti bootstrapových statistik a odhadů.

V další části přednášky se zaměříme na aplikace a modifikace metody bootstrap pro případy závislých pozorování, zejména pro stacionární i nestacionární časové řady. Zabývat se budeme jak metodami, které využívají znalosti modelu (model based), tak metodami, které informaci o modelu nevyužívají (model free). Kromě konzistence a eficience se budeme zabývat také otázkami jak volit rozsahy bootstrapových výběrů, počty opakování simulací a délky bloků pozorování, které jsou vybírány jako jeden celek.

Luboš Prchal, KPMS UK MFF Praha

Jak jsem se učil modelovat realitu

Příspěvek se zabývá statistickou analýzou závislosti intenzity radioaktivního záření na výšce nad zemským povrchem. V úvodní kapitole popisuje způsob měření a pohled meteorologů na analyzovaná data. Druhá kapitola je věnována hledání tvaru a parametrizace nelineárních regresních modelů vhodných k popisu vertikálních profilů radiace. Použití numerických metod pro odhad parametrů navrhovaných modelů a porovnání dostupného software se diskutuje v kapitole III. Výsledkům a srovnání odhadnutých modelů je pak věnována čtvrtá kapitola.

Tato práce ukazuje cestu, jak statistickými metodami modelovat a odhadnout fyzikálně zatím nepopsaný funkcionální vztah mezi radiací a nadmořskou výškou. Odhadnuté modely navíc můžeme chápat jako „vstupní“ funkcionální data pro další statistické analýzy této problematiky; analýzy, které by měly odpovědět na aktuální meteorologické otázky, např. zda existuje sezónnost v chování radiokativity či jak identifikovat „odlehlá“ pozorování.

Soňa Reisnerová, KPMS MFF UK Praha

Analýza přežití pro diskrétní čas s aplikací na Coxův regresní model

Tento příspěvek patří do oblasti analýzy přežití. Nejprve uvádí pár motivačních příkladů z praxe, vysvětluje pojem cenzorování dat a poté se zaměřuje na jeden z velmi často používaných modelů analýzy přežití. Jedná se o Coxův regresní model proporcionálních rizik, který je zde použit pro případ diskrétní povahy dat. Konkrétně jde o situaci, kdy předpokládáme, že pro každý časový interval máme konstantní základní rizikovou funkci. Z matematického hlediska dojde ke zjednodušení, neboť se kumulativní riziková funkce vyjádří jako suma přes jednotlivé hodnoty základní rizikové funkce. Budeme uvažovat konečný horizont pozorování, cenzorování zprava a v regresním modelu na čase nezávislé kovariáty. Pro tento případ ukazujeme, jakými metodami se získávají odhady parametrů modelu a jaké jsou asymptotické vlastnosti těchto parametrů. Bez povšimnutí nezůstanou ani 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. Data mají povahu kvartálních pozorování a nezaměstnaní jsou rozdělení podle pohlaví, věku, dosaženého vzdělání a oblasti.

Monica Rencová, KM FSV ČVUT Praha

Extremal Theory in Temperature Series

The broadly accepted hypothesis of global warming stimulated an interest for temperature series. Many scientists assume that the change does not necessarily occur in the mean of the series but rather in some other characteristics, e.g. appearance of some extreme events or decrease of difference between summer and winter temperatures etc.

From the statistical point of view it is important to study behavior of maximal and minimal annual temperatures. According to the extreme value theory it seems natural to expect that the distribution of annual maximal, resp. minimal temperatures follows one of the extreme value distributions.

However, the available data contradict this assumption. There are several reasons why the Gumbel distribution does not fit the data well:

- The maximal/minimal annual temperature is a maximum/minimum of independent identically distributed random variables. Nevertheless the study shows that there is a strong correlation between the daily temperature values. The correlation between two following days is for all series very close to 0.8.
- The annual extremes are affected by the variation of the temperature during the year.
- A role plays also the type of the distribution of daily temperatures.

The contribution studies the statistical properties of very long European temperature series measured in Padova, Milan, Uppsala, Stockholm,St.Petersburg, Cadiz and Belgium. For modelling data on climatological factors such as maximum/minimum daily temperatures is widely applied the Weibull distribution. For each temperature series we found parameter estimates of the three parametric Weibull distribution. From the analysis it follows that the three parameter Weibull model fits the data well.

Patrícia Rexová, EuroMISE centrum, ÚI AV ČR Praha

Obtížnost položek didaktického testu

Vezměme v úvahu situaci, kdy hodnotíme znalost n studentů pomocí didaktického testu složeného z k položek. Odpovědi i-tého studenta na j-tou položku označme x_{ij}. Zabývejme se pro jednoduchost pouze případem, kdy je odpověď hodnocena jako správná ($x_{ij} = 1$) či jako nesprávná ($x_{ij} = 0$).

Klasická teorie předpokládá pro náhodnou veličinu X_{ij} (jejíž je x_{ij} realizací) model

$$X_{ij} = a_i - b_j + e_{ij},$$

kde a_i popisuje úroveň znalosti i-tého studenta (ať už jako neznámý paramentr či jako náhodný efekt), b_j je neznámý parametr popisující obtížnost j-té položky a $e_{ij} \sim N(0, \sigma^2)$ je náhodná chyba. Jak je patrné, model opomíjí skutečnost, že odezva může nabývat pouze nul a jedniček.

Modernějším přístupem je popisovat situaci pomocí logistické regrese:

$$P(X_{ij} = x_{ij}) = \frac{exp[x_{ij}(a_i - b_j)]}{1 + exp(a_i - b_j)},$$

kde a_i a b_j mají obdobný význam jako v modelu klasickém.

V příspěvku budou diskutovány dva výše zmíněné modely. Mimo jiné budou předvedeny analogie mezi odhadováním parametru obtížnosti v daných

dvou modelech. Jako aplikace bude představen evaluační systém ExaME vyvíjený v rámci EuroMISE centra.
Poděkování: Tato práce vznikla v rámci grantu MŠMT uděleným pod číslem LN00B107.

Alexander Savin, ÚM SAV Bratislava

Testy a konfidenčné intervaly pre strednú hodnotu v modeloch jednoduchého triedenia.

Kenward a Roger (1997) 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, či už s pevnými alebo náhodnymi efektmi v prípade heteroskedasticity, a porovnaná už so známymi metódami pre inferenciu o strednej hodnote spomenutých v Hartung, J., Argaç a Makambi K.H. (2002) pre pevné efekty a v Witkovský, Savin a Wimmer G. (2003) pre náhodné efekty.

Výskum bol podporený grantom z Vedececkej Grantovej Agentury Slovenskej Republiky VEGA 1/0264/03.

Literatura:

Hartung J., Argaç D. a 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.

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.

Ivan Saxl a Lucie Ilucová, MÚ AV ČR Praha

Historie grafického zobrazování statistických dat

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. Na požadavek „graphical statistics“ poskytne vyhledávač Google 988 000 odkazů, další tisíce produkují hesla „statistical graphics“, „statistical graphs“ atd.

Na samém počátku grafického zobrazování dat jsou pochopitelně mapy, z nichž t.č. nejstarší známá byla nakreslena před 8000 až 9000 tisíci let. Tématická kartografie, jejímž předmětem jsou mapy doplněné daty popisujícími osídlení, floru i faunu, vzdělanost, obchod, dopravu, počasí či šíření chorob atd., je však stará pouze několik málo staletí (E. Halley, 1701). Na počátku statistické grafiky je diagram zachycující nadhodnocené odhady rozdílů v zeměpisných délkách Toleda a Říma podle řady historických map (M. F. van Langren, 1644). Systematický rozvoj statistické grafiky se však objevuje až koncem XVIII. století (W. Playfair, J.H. Lambert, A.F.W. Crome). Grafy mají vesměs politicko-ekonomickou tématiku, pravoúhlé souřadnice jsou sice nejběžnější, objevují se však také diagramy kruhové a polární. K rozvoji statistické grafiky výrazně přispěla i různá zařízení k automaticky zapisující data, jako byly klimatické hodiny zapisující teplotu a směr větru (Ch. Wren, 1663) či indikátor tlaku páry v parním stroji (J. Watt, 1796). Přesto téměř do konce XVIII. století byly grafy nakreslené těmito přístroji považovány za bezcenné pro vědeckou analýzu a byly přepisovány do tabulek.

Velký rozmach grafického zobrazování dat probíhá v XIX. století. Je do značné míry dílem francouzských inženýrů, většinou žáků Gasparda Monge, soustředěných kolem Ecole des Ponts et Chaussées. Vrcholí pak pracemi Ch. J. Minarda. Souběžně se ovšem grafika uplatňuje ve společenských studiích (L.A.J. Quetelet, v lékařství (J. Snow, Florence Nightingalová), v biologii (F. Galton) a objevují se i ve školních učebnicích. Specifický příspěvek podal W.S. Jevons svými diagramy zachycujícími „komerční bouře“ typu objev australského zlata.

V současné době se rozvíjejí systematicky budované teorie obrazového zpracování dat, detailně zachycující působení jednotlivých obrazových prvků jako je tvar, barva, pseudo-dimenze aj. (E. Tufte, W.S. Cleveland), na internetu jsou vystaveny učebnice statistické grafiky, pořádají se semináře k této problematice, softwarové firmy vydávají speciální grafické manuály. Pozornost je věnována také skutečnosti, že grafické zobrazování dat má ve srovnání s tabulkami výrazně emocionální účinek, který snadno (úmyslně či neúmyslně) může vést k vyvolání dojmu, jenž je s prezentovanými daty v rozporu.

Literatura (uvedeny jsou jen nejobsažnější internetové stránky):
http://www.math.yorku.ca/SCS/Gallery/
http://www.math.yorku.ca/SCS/StatResource.html
http://www.phil.uni-sb.de/FR/Medienzentrum/Grafikexperiment/verweise.html
http://cm.bell-labs.com/cm/ms/departments/sia/doc/index.html
http://www.eia.doe.gov-neic/graphs/

Miroslav Šiman, KPMS MFF UK Praha

Portmanteau testy a jejich aplikace na testování podmíněné heteroskedasticity

V tomto článku si nejprve shrneme poznatky o dosud navržených portmanteau testech, poté se zamyslíme nad vhodností jejich použití při testování podmíněné heteroskedasticity a nakonec pro její detekci zkusíme navrhnout lepší portmanteau testy, než jsou ty dosud běžně používané. Při jejich konstrukci se pokusíme skloubit dohromady několik myšlenek, které již byly jednotlivě při tvorbě nových portmanteau testů úspěšně vyzkoušeny: vhodně přeuspořádat zkoumanou řadu před testováním, použít pořadí místo skutečných hodnot, aplikovat transformaci stabilizující rozptyl a přiřadit odlišnou důležitost různým hodnotám autokorelační funkce. Měřítkem optimality navrhovaných testů nám bude zejména velikost a síla jimi dosažená při testování nulové hypotézy náhodnosti (tj. nezávislých, stejně rozdělených náhodných veličin) proti alternativám GARCH(1,1) modelu.

Protože praxe silně preferuje snadno srozumitelné a výpočetně jednoduché metody, budeme se zde zajímat jen o testy počítané v časové doméně, založené na myšlence korelačního koeficientu a nevyžadující použití simulačních metod.

Petr Šimeček a Milan Studený, ÚTIA AV ČR Praha

Využití Hilbertovy báze k ověření shodnosti strukturálních a kombinatorických imsetů

Jedním ze způsobů, jak reprezentovat struktury (podmíněné) nezávislosti mezi n náhodnými veličinami X_1, X_2, ..., X_n, je pomocí tzv. „imsetů“. Imsetem rozumíme zobrazení potenční množiny $\mathcal{P}(\{\infty, \in, \ldots, \setminus\})$ do množiny celých čísel Z. Tato metoda napravuje nešvar v praxi i literatuře častějšího popisu pomocí grafů: těch je totiž řádově méně ($\approx 2^{poly(n)}$, kde polynom $poly(n)$ je určen typem hran v grafu) nežli všech možných struktur podmíněné nezávislosti ($\approx 2^{2^n}$). Podrobné informace o metodě popisu struktur podmíněných nezávislostí pomocí imsetů lze nalézt například ve Studený (2004). Jedním z otevřených problémů z hlediska implementace uvedené metody zůstává otázka shodnosti tzv. strukturálních a kombinatorických imsetů (viz níže).

Semielementárním imsetem odpovídajícím nezávislosti mezi $\{X_i; i \in A\}$ a $\{X_i; i \in B\}$ dáno $\{X_i; i \in C\}$, kde A, B a C jsou po dvou disjunktní podmnožiny $\{1, 2, \ldots, n\}$, budeme rozumět zobrazení, jež $A \cup B \cup C$ a C přiřadí

1, $A \cup C$ a $B \cup C$ přiřadí -1 a zbylým prvkům potenční množiny 0. Množinu semielementárních imsetů budeme značit $\mathcal{D}_\setminus$.

Z teoretického, ale především praktického hlediska je velmi zajímavá otázka, zda-li $\mathcal{C}_\setminus = \mathcal{S}_\setminus$. Odpověď můžeme získat nalezením tzv. celočíselné Hilbertovy báze (viz Schrijver (1998)) kónického obalu $\mathcal{S}_\setminus$. V praxi se však budeme potýkat s velkou výpočetní složitostí tohoto problému. Náš článek ukazuje, jak v reálném čase provést výpočet této báze pro $n = 3$ a $n = 4$, a zároveň jsou prezentovány částečné výsledky pro $n = 5$.

Literatura:
Studený M. (2004): *On Probabilistic Conditional Independence Structures*. Springer.
Schrivjer A., (1998): *Theory of Linear and Integer Programming*. John Wiley.

Marie Šimečková, KPMS MFF UK Praha

Metoda nejmenších useknutých čtverců (LTS) jako diagnostický nástroj

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.

Příspěvek vznikl s podporou grantu GA ČR 402/03/0084.

Literatura:
Rousseeuw P.J. a Leroy, A.M. (1987). *Robust regression and outlier detection*. Wiley, New York.
Víšek J.Á, (2000): *On the diversity of estimates*. *Computational Statistics & Data Analysis*, Vol. 34, 67 – 89.

Milan Stehlík, KPMŠt UKo Bratislava

Covariance related properties of D-optimal correlated designs

The aim of this paper is discussion on particular properties of the D-optimal designs within the intrinsically stationary random fields with corelated errors. Here we focus mainly on the relation of the D-optimal design and covariance

structure. We show that design points can collapse under the presence of some covariance structures. This enables to include so called nugget effect (see Cressie) by natural way and we also introduce the linear and exponential variograms. Some numerical examples are also included.

Literatura:
Banerjee S., Carlin B., Gelfand A., *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall/CRC, 2004.
Cressie N.A.C., *Statistics for Spatial Data*. Wiley, New York, 1993.
Müller W.G. and Stehlík M., *An example of D-optimal designs in the case of correlated errors*. Accepted to Compstat 2004 Proceedings, 2004.
Stehlík M., *Further aspects on an example of D-optimal designs in the case of correlated errors*. Report 1 of the Research Report Series of the Department of Statistics and Mathematics, Wirtschaftsuniversität Wien, Austria, 2004.

Josef Tvrdík, KI OU Ostrava

Stochastické algoritmy v obtížných úlohách odhadu parametrů regresních modelů

V příspěvku bude popsán algoritmus řízeného náhodného prohledávání (Controlled Random Search, CRS), který vyhledává globální minimum $\mathbf{x}^*$ pro $f : D \to \mathcal{R},\ D \subset \mathcal{R}^d$, $\mathbf{x}^* = \arg\min_{\mathbf{x}\in D} f(\mathbf{x})$, kde $D = \prod_{i=1}^{d}[a_i, b_i]$, $a_i < b_i,\ i = 1, 2, \ldots, d$.

Dále bude popsáno zobecnění algoritmu CRS, kdy heuristika pro generování nového bodu v populaci je v každém kroku vybírána ze seznamu dovolených heuristik náhodně s pravděpodobností úměrnou její dosavadní úspěšnosti.

Budou uvedeny výsledky dosažené několika variantami tohoto algoritmu v obtížných úlohách odhadu parametrů nelineárních regresních modelů a možná i vyslovena doporučení pro aplikaci algoritmů tohoto typu v „rutinních“ úlohách.

Literatura:
Ali M. M., Törn A. (2004). *Population set based global optimization algorithms: Some modifications and numerical studies*. Computers and Operations Research 31 (10), 1703 – 1725, 2004.
Price W.L. (1977). *A Controlled Random Search Procedure for Global Optimization*. Computer J. 20, 367 – 370.
Statistical Reference Datasets. *Nonlinear regression*. NIST Information Technology Laboratory. http://www.itl.nist.gov/div898/strd/

Tvrdík J., Mišík L., Křivý I. (2002). *Competing Heuristics in Evolutionary Algorithms*. In: Sinčák P. et al. (Eds.), 2nd Euro-ISCI, *Intelligent Technologies – Theory and Applications*, IOS Press, Amsterdam, 159 – 165.

Jan Ámos Víšek, FSV UK Praha

GMM weighted estimation

The point estimation was from the very beginning of the statistics (and econometrics) one of the key topics. In the early days, the unbiasedness was assumed to play the crucial role but later the *(weak) consistency* overtook the governance.

The (classical and/or robust) statistics developed a bunch of "principles", heuristics of which promised to yield the estimators being not only consistent but also premiant in a widely considered competition (achieving e. g. efficiency). Some of them worked, some not (Martin et al. (1989)). Under special circumstances, they induced occasionally the same "estimating formula", e.g. LS and ML. Moreover, the blind pursuit for *effiiency* made sometimes from the good servant the bad boss, Mizon (1995).

Typically, the estimator was given as a solution of a (vector) equation (*normal equations*) - interpretable as p-tuple of orthogonality conditions (of residuals to the columns of design matrix, e. g.).

The econometrics went directly (brutally?) to the goal - *consistency of the estimation* - leaving aside all marketing (of users) by heuristics (Taking this liberty due to absence of a need to attract the clients (contrary to the statistics), since in the economics a stochastic processing data became a must. The recent titles of monographs as *Macroeconometrics* or *Microeconometrics* as well as the topics of interests of Nobel prize winners (2003), Clive W. J. Granger and Robert F. Engle or (2001) James J. Heckman and Daniel L. McFadden, don't leave too much space for hesitations.) The consistency requires orthogonality of residuals to the (estimated) model and hence the estimators are defined as solution of a q-tuple of orthogonality conditions ($p \leq q$), Hasman (1982). It allows for direct employment of additional information about the parameter in question, Wooldridge (2001) (we speak about *Generalized Method of Moments estimation*).

Despite the forty years of robust studies, the econometrics haven't taken seriously (possible) fatal consequences of a slight deviation of the assumed model from the underlying one (Fisher (1922)) or of a few contaminating observations (Hampel et al.(1986), Huber (1981)).

Since the weighting down the order statistics of squared residuals appeared to be powerful tool for influential-points-recognition (Plát (2004)),

the present paper offers an idea of the *generalized method of moments weighted estimators* and shows that *the least weighted squares* (Víšek (2000)) are special case of them.

Research was supported by grant of GA ČR number 402/03/0084.

Gejza Wimmer a Viktor Witkovsky, ÚM SAV Bratislava

Konfidenčné intervaly pre efekt ošetrenia v klinických pokusoch

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 Kenward & Roger (1997) a pre prípad medzilaboratórnych porovnávacích štúdii podrobne študovanú v práci Witkovský, Savin & Wimmer (2003) pre prípad spoločnej strednej hodnoty v medzilaboratórnych porovnávacích štúdiach. Simulačná štúdia ukazuje vlastnosti (resp. ohraničenia) navrhnutého konfidenčného intervalu.

Literatura:

Kenward M.G. a Roger J.H. *Small sample inference for fixed effects from restricted maximum likelihood*. Biometrics 53, 1997, 983 – 997.

Witkovský V., Savin A. a Wimmer G. *On small sample inference for common mean in heteroscedastic one-way model*. Discussiones Mathematicae: Probability and Statistics 23, 2003, 123 – 145.

Jitka Zichová, KPMS MFF UK Praha

Grafické modely v analýze finančních dat

Grafické (nebo jak je někdy v české terminologii uváděno grafové) modely jsou jedním z nástrojů mnohorozměrné statistické analýzy. V poslední době byly s úspěchem aplikovány v oblasti financí, o čemž svědčí např. práce [1], [2], [3].

Grafickým modelem rozumíme množinu pravděpodobnostních rozdělení k-rozměrného náhodného vektoru X, která splňují podmínky dané tzv. grafem podmíněných nezávislostí. Graf podmíněných nezávislostí je tvořen množinou vrcholů $V = \{1, \ldots, k\}$ a množinou hran, přičemž chybějící hrana mezi vrcholy i, j, indikuje podmíněnou nezávislost i-té a j-té složky vektoru X při pevných hodnotách ostatních složek.

Má-li X mnohorozměrné normální rozdělení, hovoříme o gaussovských grafických modelech, ale metodologie je vypracována i pro vektory, jejichž složky jsou kategoriální proměnné. Závislosti regresního typu popisují modely s řetězovými grafy, v nichž orientované hrany značí vztahy mezi závisle a nezávisle proměnnými.

Je zřejmé, že v daném k-rozměrném vektoru lze strukturu podmíněných nezávislostí jeho složek modelovat grafy o k vrcholech a různém počtu hran. Graf obsahující všechny hrany představuje model bez omezení podmíněnými nezávislostmi. V aplikacích je hlavním úkolem navrhnout grafický model, který dobře popisuje data, což vede na problém testování shody modelu s daty. Testovou statistikou je deviance odvozená od věrohodnostního poměru. Pro její výpočet potřebujeme odhady parametrů rozdělení vektoru X. Pro případ mnohorozměrného normálního rozdělení a rozdělení daného k-rozměrnou kontingenční tabulkou byly na MFF UK v Praze vypracovány algoritmy v programu Mathematica řešící nejen problém odhadu, ale i selekci modelu pro daná data a testování shody modelu s daty.

Tyto algoritmy byly uplatněny při analýze dat z finanční praxe, ať už se jednalo o studium závislostí mezi burzovními indexy nebo aplikaci v credit scoringu, což je posuzování žadatelů o bankovní úvěry.

Literatura:

[1] Giudici P. *Data Mining*. McGraw-Hill, Milano, 2001.

[2] Hand D.J., Mc Conway K.J. a Stanghellini E. *Graphical models of applicants for credit*. IMA Journal of Mathematics Applied in Business and Industry 8, 1997, 143 – 155.

[3] Stanghellini E., Mc Conway,K.J. a Hand, D.J. *A discrete variable chain graph for applicants for credit*. Applied Statistics 48, 1999, 239 – 251.

[4] Whittaker, J. *Graphical Models in Applied Multivariate Statistics*. Willey, New York, 1990.

[5] Diplomové práce posluchačů MFF UK 2001-2003 (Lněnička, Randáková, Chýna, Zelinková, Svobodová), školitel J. Zichová.

Ivan Žežula a D. Klein, PF UPJŠ Košice

Robustnost v modelu růstových křivek

V práci uvažujeme dvě varianty modelu růstových křivek - replikovaný a model se speciálními variančními strukturami. Byla provedena simulační studie zkoumající robustnost modelu vůči porušení některých předpokladů. Ukazuje se, že model je velmi citlivý na přítomnost systematické chyby měření, na druhou stranu je dostatečně robustní vůči porušení předpokladu normality. V modelu se speciálními variančními strukturami se ukazuje nespolehlivost odhadu MSE rozptylu.