UNIVERZITA KARLOVA V PRAZE

matematicko-fyzikální fakulta

Arnošt Komárek

Katedra pravděpodobnosti a matematické statistiky

Regrese s korelovanými intervalově cenzorovanými daty zatíženými nepřesnou klasifikací události

ROBUST'18

Jetřichovice, 19. – 24. ledna 2014

Příspěvek založen na spolupráci

s **Maríou José García-Zattera**

a **Alejandrem Jarou**

Pontificia Universidad Católica de Chile

Santiago de Chile

Část I

Intervalově cenzorovaná data zatížená nepřesnou klasifikací události

Signal Tandmobiel® (*mobilní zub*) studie

Longitudinální zubní studie ve Flandrech, 1996 – 2001.

Stratifikovaný náhodný výběr dětí navštěvujících v roce 1996 1. třídu ZŠ.

2 315 chlapců, 2 153 dívek (7.3 % vlámské populace daného věku).

Každoroční (v 1. – 6. třídě) detailní zubní vyšetření jedním z 16 proškolených zubařů.

- Dané dítě neprohlíží v jednotlivých letech nutně stejný zubař.

V prvním roce též dotazník ohledně zubní hygieny, stravovacích návyků atd. vyplněný rodiči.

Korelované časy do události

Pro pamětníky: ROBUST 2006 (Lhota nad Rohanovem), ROBUST 2004 (Třešť)

$T_{(i,j)} \in \mathbb{R}_+$: čas do vzniku zubního kazu u $j$tého zubu $i$tého dítěte
($i = 1, \ldots, N$, $j = 1, \ldots, J$)

- obecně: čas do události u $j$té jednotky $i$tého subjektu.

$\mathbf{x}_{(i,j)}$: vektor regresorů, které mohou vysvětlit chování $T_{(i,j)}$.

Hlavní cíl: Regresní model pro závislost $T_{(i,j)}$ na $\mathbf{x}_{(i,j)}$.

Potřeba vzít v potaz:

složky vektoru $\boldsymbol{T}_i = \left(T_{(i,1)}, \ldots, T_{(i,J)}\right)^\top$ nejsou nekorelované.

Kontrola, zda nastala událost, je prováděna pouze v (předem) daných momentech.

K_i: počet kontrol u $i$tého subjektu.

Časy kontrol u $i$tého subjektu:

$$0 = v_{(i,0)} < v_{(i,1)} < v_{(i,2)} < \cdots < v_{(i,K_i)} < v_{(i,K_i+1)} = \infty.$$

Čas $T_{(i,j)}$ není pozorován přesně, je pouze známo

$$T_{(i,j)} \in \left(v_{(i,l_{(i,j)}-1)},\ v_{(i,l_{(i,j)})}\right] \quad \text{pro } l_{(i,j)} \in \{1, \ldots, K_i + 1\}$$

➠ intervalově cenzorovaná data

- $l_{(i,j)} \leq K_i$: $v_{(i,l_{(i,j)})}$ je čas kontroly, při které byla událost poprvé detekována;
- $l_{(i,j)} = K_i + 1$: $v_{(i,l_{(i,j)})} = \infty$ ➠ událost nenastala do času poslední kontroly.

Intervalové cenzorování zatížené nepřesnou klasifikací události

Misclassified interval-censored data

Klasifikace události (zda nastala či nikoliv) může být zatíženo chybou

- examinátor (diagnostický test, ...) s nenulovou pravděpodobností falešně pozitivního, resp. falešně negativního výsledku.

Pro (i,j)tou jednotku pozorujeme 0/1 posloupnost

$$\boldsymbol{Y}_{(i,j)} = \left(Y_{(i,j,1)}, \ldots, Y_{(i,j,K_i)}\right)^\top.$$

Examinátor (test) indikoval, že událost do $k$té kontroly (v čase $v_{(i,k)}$)

- nastala ➠ $Y_{(i,j,k)} = 1$;
- nenastala ➠ $Y_{(i,j,k)} = 0$.

➠ intervalové cenzorování zatížené nepřesnou klasifikací události (*misclassified interval-censored data*)

Intervalové cenzorování zatížené nepřesnou klasifikací události

Misclassified interval-censored data

S perfektním examinátorem/testem (nulová pravděpodobnost falešně pozitivních i negativních výsledků) by posloupnost $\boldsymbol{Y}_{(i,j)}$ byla (samozřejmě) monotónní.

Chování examinátora/testu v čase $v_{i,k}$ nezávislé na výsledcích získaných v minulosti ($Y_{(i,j,l)}, l < k$):

- posloupnost $\boldsymbol{Y}_{(i,j)}$ není nutně monotónní.

Regrese s intervalové cenzorovanými daty zatíženými nepřesnou klasifikací události

Hlavní cíl:
i nadále regresní model závislosti času do události $T_{(i,j)}$ na prediktorech $\mathbf{x}_{(i,j)}$.

Čas $T_{(i,j)}$ však pozorujeme pouze skrze $\boldsymbol{Y}_{(i,j)}$
– sadu 0/1 indikátorů události zatížených klasifikační chybou.

Zopakování značení

Pro $i$tý subjekt (dítě), $i = 1, \ldots, N$

$\boldsymbol{T}_i = \left(T_{(i,1)}, \ldots, T_{(i,J)}\right)^\top$: časy do události pro J jednotek (zuby)

➠ „skrytá“ data.

$\boldsymbol{Y}_i = \left(\boldsymbol{Y}_{(i,1)}^\top, \ldots, \boldsymbol{Y}_{(i,J)}^\top\right)^\top$: posloupnosti 0/1 indikátorů události zatíženými klasifikační chybou

➠ pozorovaná data.

$\mathbf{x}_i = \left(\mathbf{x}_{(i,1)}^\top, \ldots, \mathbf{x}_{(i,J)}^\top\right)^\top$: regresory pro vysvětlení chování časů $\boldsymbol{T}_i$.

$\mathbf{v}_i = \left(v_{(i,1)}, \ldots, v_{(i,K_i)}\right)^\top$: časy kontrol stavu události.

Hierarchický model

Model pro pozorovaná data (věrohodnost) specifikujeme hierarchicky.

Sdružená věrohodnost skrytých a pozorovaných dat $i$tého subjektu

$$p(\boldsymbol{Y}_i, \boldsymbol{T}_i) = p(\boldsymbol{Y}_i \mid \boldsymbol{T}_i)\, p(\boldsymbol{T}_i).$$

- $p(\boldsymbol{Y}_i \mid \boldsymbol{T}_i)$: model pro klasifikační proces
  - časy $\mathbf{v}_i$ kontrol stavu události zde budou vystupovat v roli regresorů.
- $p(\boldsymbol{T}_i)$: (regresní) model pro (korelované) časy do události
  - $\mathbf{x}_i$ zde bude vystupovat klasicky v roli regresorů.

Věrohodnost pozorovaných dat $i$tého subjektu

$$p(\boldsymbol{Y}_i) = \int_{\mathbb{R}_+^J} p(\boldsymbol{Y}_i, \boldsymbol{T}_i)\, \mathrm{d}\boldsymbol{T}_i.$$

Nezávislost mezi subjekty

Budeme předpokládat nezávislost mezi subjekty, tj. pro věrohodnosti odpovídající všem subjektům:

$$p(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N, \boldsymbol{T}_1, \ldots, \boldsymbol{T}_N) = \prod_{i=1}^{N} p(\boldsymbol{Y}_i, \boldsymbol{T}_i) = \prod_{i=1}^{N} \Big\{ p(\boldsymbol{Y}_i \mid \boldsymbol{T}_i)\, p(\boldsymbol{T}_i) \Big\},$$

$$p(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N) = \prod_{i=1}^{N} p(\boldsymbol{Y}_i) = \prod_{i=1}^{N} \int_{\mathbb{R}_+^J} p(\boldsymbol{Y}_i, \boldsymbol{T}_i)\, \mathrm{d}\boldsymbol{T}_i.$$

Část II

Model pro klasifikační proces

Model pro klasifikační proces

Nyní se budeme zabývat $p(\boldsymbol{Y}_i \mid \boldsymbol{T}_i)$ částí sdružené věrohodnosti skrytých a pozorovaných dat.

Budeme předpokládat, že klasifikace události pro danou jednotku v daném čase ($Y_{(i,j,k)}$) je (podmíněně) nezávislá na

(a) klasifikaci události u jiných jednotek (jiné j) v libovolném čase (libovolné k);

(b) klasifikaci události u stejné jednotky (stejné j) v jiném čase (jiné k);

(c) času události u jiných jednotek (jiné j).

To jest, předpokládáme, že lze psát

$$p(\boldsymbol{Y}_i \mid \boldsymbol{T}_i) = \prod_{j=1}^{J} \prod_{k=1}^{K_i} p(Y_{(i,j,k)} \mid T_{(i,j)}).$$

V dalším už se budeme zabývat jenom tím, jak může vypadat $p(Y_{(i,j,k)} \mid T_{(i,j)})$.

Jednoduchý model pro klasifikační proces

Jeden examinátor/test

α: sensitivita examinátora/testu.

η: specificita examinátora/testu.

To jest,

$$\alpha = \mathrm{P}\left(Y_{(i,j,k)} = 1 \,\middle|\, T_{(i,j)} \leq v_{(i,k)}\right),$$

$$\eta = \mathrm{P}\left(Y_{(i,j,k)} = 0 \,\middle|\, T_{(i,j)} > v_{(i,k)}\right).$$

α, η: neznámé parametry modelu.

$\boldsymbol{Y}_{(i,j,k)} \mid T_{(i,j)}$ část modelu:

$$p(Y_{(i,j,k)} \mid T_{(i,j)}) = p(Y_{(i,j,k)} \mid T_{(i,j)};\, \alpha,\, \eta,\, v_{i,k})$$

$$= \begin{cases} \alpha^{Y_{(i,j,k)}} (1-\alpha)^{1-Y_{(i,j,k)}}, & \text{pokud } T_{(i,j)} \leq v_{(i,k)} \\ & \text{(správné } Y_{(i,j,k)} \text{ je rovno 1),} \\ (1-\eta)^{Y_{(i,j,k)}} \eta^{1-Y_{(i,j,k)}}, & \text{pokud } T_{(i,j)} > v_{(i,k)} \\ & \text{(správné } Y_{(i,j,k)} \text{ je rovno 0).} \end{cases}$$

O něco složitější model pro klasifikační proces

Q examinátorů/testů, různé sensitivity/specificity pro různé jednotky (různá j)

Bylo potřeba v kontextu studie mobilního zubu.

Klasifikaci kazu provádělo $Q = 16$ různých zubařů:

- každý má jinou schopnost detekovat zubní kaz;
- potřeba uvážit závislost sensitivit/specificit na examinátorovi.

Různá $j \equiv$ různé zuby (šestka vlevo dole, čtyřka vpravo nahoře, . . .):

- detekce kazu je na různých zubech různě obtížná;
- potřeba uvážit závislost sensitivit/specificit na j (jednotce).

O něco složitější model pro klasifikační proces

Q examinátorů/testů, různé sensitivity/specificity pro různé jednotky (různá j)

Další „regresor“ v modelu: $\xi_{(i,k)} \in \{1, \ldots, Q\}$

- index (číslo) examinátora, který prováděl klasifikaci události u (všech jednotek) $i$tého subjektu během jeho $k$té kontroly v čase $v_{(i,k)}$.

Trochu více neznámých parametrů modelu ($q = 1, \ldots, Q$):

$$\boldsymbol{\alpha}_q = \left(\alpha_{(q,1)}, \ldots, \alpha_{(q,J)}\right)^\top,$$

$$\boldsymbol{\eta}_q = \left(\eta_{(q,1)}, \ldots, \eta_{(q,J)}\right)^\top.$$

$\alpha_{(q,j)}, \eta_{(q,j)}$: sensitivita a specificita klasifikace, jestliže ji provádí examinátor q na jednotce j, to jest,

$$\alpha_{(q,j)} = \mathrm{P}\left(Y_{(i,j,k)} = 1 \,\middle|\, T_{(i,j)} \leq v_{(i,k)};\ \xi_{(i,k)} = q\right),$$

$$\eta_{(q,j)} = \mathrm{P}\left(Y_{(i,j,k)} = 0 \,\middle|\, T_{(i,j)} > v_{(i,k)};\ \xi_{(i,k)} = q\right).$$

O něco složitější model pro klasifikační proces

Q examinátorů/testů, různé sensitivity/specificity pro různé jednotky (různá j)

Vektory všech (neznámých) sensitivit a specificit:

$$\boldsymbol{\alpha} = \left(\boldsymbol{\alpha}_1^\top, \ldots, \boldsymbol{\alpha}_Q^\top\right)^\top,$$

$$\boldsymbol{\eta} = \left(\boldsymbol{\eta}_1^\top, \ldots, \boldsymbol{\eta}_Q^\top\right)^\top.$$

$\boldsymbol{Y}_{(i,j,k)} \mid T_{(i,j)}$ část modelu, jenom nějaké indexy navíc ☺

$$p(Y_{(i,j,k)} \mid T_{(i,j)}) = p(Y_{(i,j,k)} \mid T_{(i,j)};\, \boldsymbol{\alpha},\, \boldsymbol{\eta},\, v_{i,k},\, \xi_{(i,k)})$$

$$= \begin{cases} \alpha_{(\xi_{(i,k)},j)}^{Y_{(i,j,k)}} \left(1 - \alpha_{(\xi_{(i,k)},j)}\right)^{1-Y_{(i,j,k)}}, & \text{pokud } T_{(i,j)} \leq v_{(i,k)} \\ & \text{(správné } Y_{(i,j,k)} \text{ je rovno 1),} \\ \left(1 - \eta_{(\xi_{(i,k)},j)}\right)^{Y_{(i,j,k)}} \eta_{(\xi_{(i,k)},j)}^{1-Y_{(i,j,k)}}, & \text{pokud } T_{(i,j)} > v_{(i,k)} \\ & \text{(správné } Y_{(i,j,k)} \text{ je rovno 0).} \end{cases}$$

Ještě složitější model pro klasifikační proces

Sensitivity/specificity lze dále modelovat pomocí charakteristik jednotlivých examinátorů/testů, resp. charakteristik jednotek, jsou-li nějaké charakteristiky k dispozici.

Lze použít např. logistickou regresi

- další hierarchická úroveň celkového modelu;
- detaily až po 22. hodině.

Tím máme z krku specifikaci $p(\boldsymbol{Y}_i \mid \boldsymbol{T}_i)$ části modelu.

Zbývá ještě $p(\boldsymbol{T}_i)$ část (model pro časy do události).

Část III

Model pro čas do události

Model pro časy do události

Připomínka

$\boldsymbol{T}_i = \left(T_{(i,1)}, \ldots, T_{(i,J)}\right)^\top$
$\equiv$ ne nutně nekorelované časy událostí u J jednotek $i$tého subjektu.

$\mathbf{x}_i = \left(\mathbf{x}_{(i,1)}^\top, \ldots, \mathbf{x}_{(i,J)}^\top\right)^\top$
$\equiv$ možné regresory vysvětlující chování časů do události.

Tvar $p(\boldsymbol{T}_i)$ lze principielně odvodit z libovolného regresního modelu pro korelovaná „přežívací“ data (o kterém jsme přesvědčeni, že se podle něho časy do události skutečně řídí):

- *frailty* Coxův model;
- *accelerated failure time (AFT)* model s náhodným absolutním členem;
- $\vdots$

AFT model s náhodným absolutním členem

Pro pamětníky: ROBUST 2006 (Lhota nad Rohanovem), ROBUST 2004 (Třešť)

$$\log\left(T_{(i,j)}\right) = \mathbf{x}_{(i,j)}^{\top}\boldsymbol{\beta} + b_i + \varepsilon_{(i,j)} \qquad i = 1, \ldots, N,\ j = 1, \ldots, J,$$

- $\boldsymbol{\beta}$: neznámé regresní koeficienty;
- $\varepsilon_{(1,1)}, \ldots, \varepsilon_{(N,J)}$: i.i.d. náhodné veličiny s hustotou $g_{\varepsilon}(\cdot)$ s nulovou střední hodnotou;
- $b_1, \ldots, b_N$: i.i.d. náhodné veličiny s hustotou $g_b(\cdot)$
  - b_i (společné pro všechna j) indukuje jistou formu závislosti mezi $T_{(i,1)}, \ldots, T_{(i,J)}$;
- $\varepsilon_{(1,1)}, \ldots, \varepsilon_{(N,J)}$, $b_1, \ldots, b_N$ nezávislé.

Příslušná transformace hustot g_ε a g_b určuje rozdělení jednotlivých časů do události a naše $p(\boldsymbol{T}_i)$.

Zde:

$$g_\varepsilon(\cdot) \sim \mathcal{N}(0, \sigma_\varepsilon^2),$$

$$g_b(\cdot) \sim \mu + \tau \underbrace{\sum_{l=-M}^{M} w_l \mathcal{N}(\kappa_l, \zeta^2)}_{\text{penalizovaná normální směs}}$$

- Neznámé parametry: σ_ε^2, $\mathbf{w} = \left(w_{-M}, \ldots, w_M\right)^\top$, μ, τ.
- Penalizovaná normální směs:

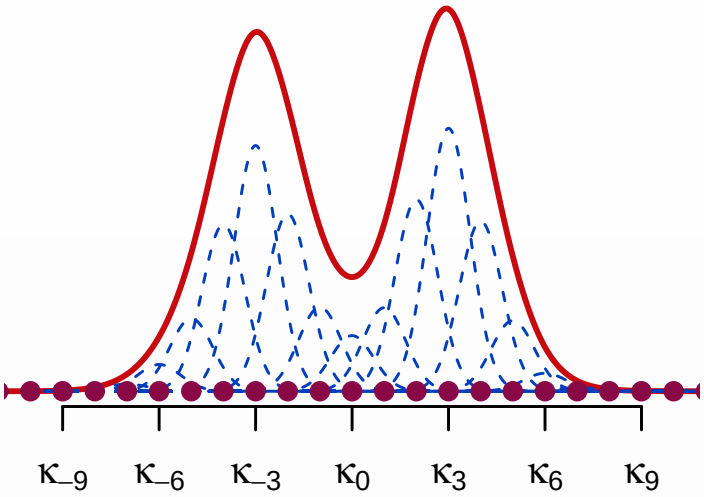
  $M \approx 15$, $\zeta \approx 0.2$,

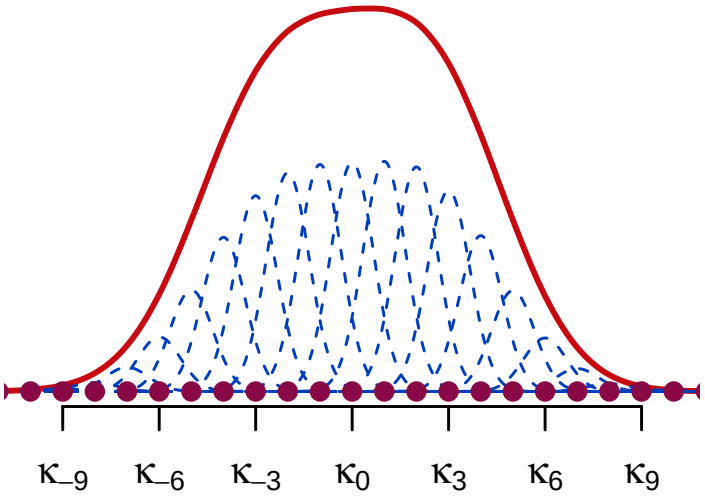
  $\kappa_{-M}, \ldots, \kappa_M$: ekvidistantní uzly na intervalu přibližně $[-4.5, 4.5]$;

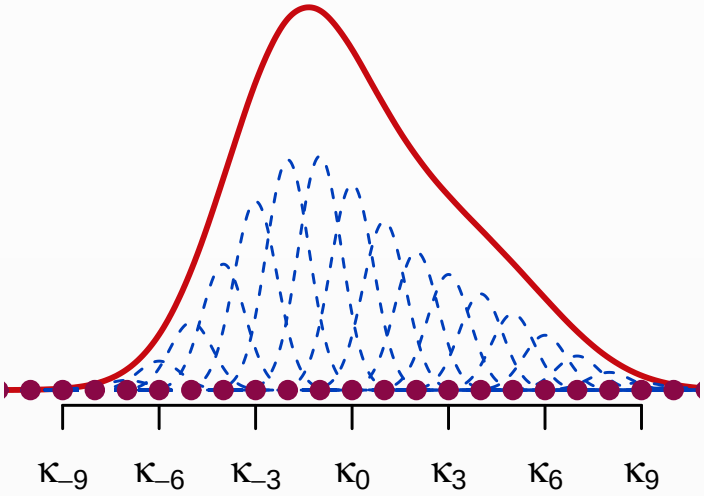
  ➠ flexibilní model pro rozdělení s přibližně nulovou střední hodnotou a jednotkovým rozptylem.
- Regularizace pomocí penalizace diferencí (transformovaných) vah $w_{-M}, \ldots, w_M$.

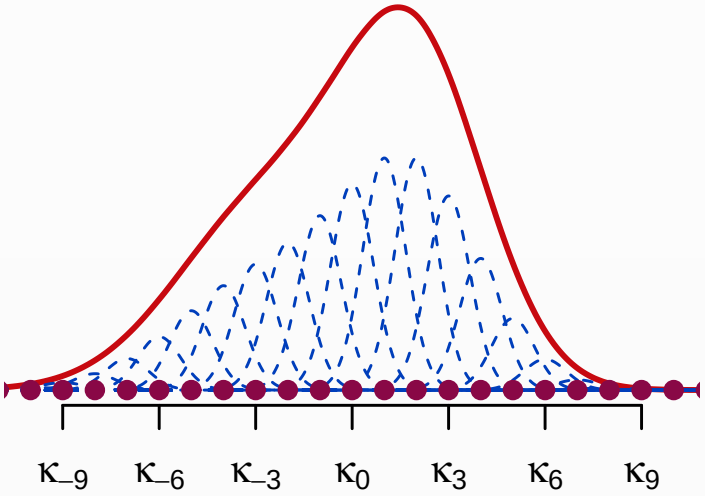
Penalizovaná normální směs

Pro pamětníky: ROBUST 2006 (Lhota nad Rohanovem), ROBUST 2004 (Třešť)









Model pro časy do události

Tvar $p(\boldsymbol{T}_i)$ odvozen z následujícího

- $\log(T_{(i,j)}) = \mathbf{x}_{(i,j)}^\top \boldsymbol{\beta} + b_i + \varepsilon_{(i,j)}, \qquad i = 1, \ldots, N, j = 1, \ldots, J;$
- $\varepsilon_{(1,1)}, \ldots, \varepsilon_{(N,J)}, \; b_1, \ldots, b_N$ nezávislé;
- $\varepsilon_{(i,j)} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$;
- $b_i \sim \mu + \tau \sum_{l=-M}^{M} w_l \mathcal{N}(\kappa_l, \zeta^2)$.

Neznámé parametry modelu

$$\boldsymbol{\theta} = \left(\boldsymbol{\beta}^\top, \mathbf{w}^\top, \mu, \tau^2, \sigma_\varepsilon^2\right)^\top.$$

Rozdělení času do události $T_{(i,j)}$ je (po log-transformaci) konvolucí plně parametrického rozdělení ($\mathcal{N}$) a penalizované normální směsi.

- Penalizovaná normální směs v rozdělení náhodného absolutního členu AFT modelu zajišťuje kromě jiného flexibilní model pro rozdělení času do události $T_{(i,j)}$.

Část IV

Inference

Celkový model

Věrohodnost

$$p(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N) = \prod_{i=1}^{N} p(\boldsymbol{Y}_i) = \prod_{i=1}^{N} \int_{\mathbb{R}_+^J} p(\boldsymbol{Y}_i, \boldsymbol{T}_i) \mathrm{d}\boldsymbol{T}_i$$

$$= \prod_{i=1}^{N} \int_{\mathbb{R}_+^J} p\left(\boldsymbol{Y}_i \,\middle|\, \boldsymbol{T}_i\right) p(\boldsymbol{T}_i)\, \mathrm{d}\boldsymbol{T}_i .$$

$p\left(\boldsymbol{Y}_i \,\middle|\, \boldsymbol{T}_i\right)$: model pro klasifikační proces

- neznámé parametry: $\boldsymbol{\alpha} = \left(\alpha_{(1,1)}, \ldots, \alpha_{(Q,J)}\right)^\top$, $\boldsymbol{\eta} = \left(\eta_{(1,1)}, \ldots, \eta_{(Q,J)}\right)^\top$: sensitivity a specificity pro jednotlivé examinátory a jednotky.

$p(\boldsymbol{T}_i)$: model pro časy do události

- AFT model s náhodným absolutním členem a penalizovanou normální směsí v jeho rozdělení;
- neznámé parametry: $\boldsymbol{\theta} = \left(\boldsymbol{\beta}^\top, \mathbf{w}^\top, \mu, \tau^2, \sigma_\varepsilon^2\right)^\top$.

Inference

Všechny parametry modelu jsou identifikovatelné, omezíme-li parametrický prostor pro sensitiity a specificity podmínkou

$$\alpha_{(q,j)} + \eta_{(q,j)} > 1, \qquad q = 1, \ldots, Q,\ j = 1, \ldots, J.$$

Výpočet věrohodnosti, respektive její maximalizace (pokud bychom chtěli maximálně věrohodné odhady) je mírně (spíše více) zkomplikována integrálem ve vyjádření $p(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N)$.

Bayesovská inference se slabě informativními apriorními rozděleními je proveditelná s pomocí MCMC:

- detaily opět až po 22. hodině;
- softwarově implementováno v  balíčku `bayesSurv` (od verze 2.3, která zatím není na CRANu, ale časem snad bude).

Část V

Simulační studie

Simulační studie

$J = 4, \quad N = 500,\ 1\,000,\ 2\,000.$

$\log\left(T_{(i,j)}\right) = 2.0 + 0.2\,x_{(i,j),1} - 0.1\,x_{(i,j),2} + b_i + \varepsilon_{(i,j)}.$

$x_{(i,j),1} \sim \mathcal{U}(0,\ 1),\ x_{(i,j),2} \sim \mathcal{A}(0.5).$

$\text{var}(b_i) + \text{var}(\varepsilon_{(i,j)}) = 0.1.$

$\sqrt{\frac{\text{var}(b_i)}{\text{var}(\varepsilon_{(i,j)})}} = \frac{\sigma_b}{\sigma_\varepsilon} = 0.5,\ 1,\ 2,\ 5.$

g_b: (a) $\mathcal{N}$, (b) zřetelně bimodální dvousložková $\mathcal{N}$ směs, (c) Gumbel.

$K_i = 10$ kontrol stavu události (v náhodných intervalech).

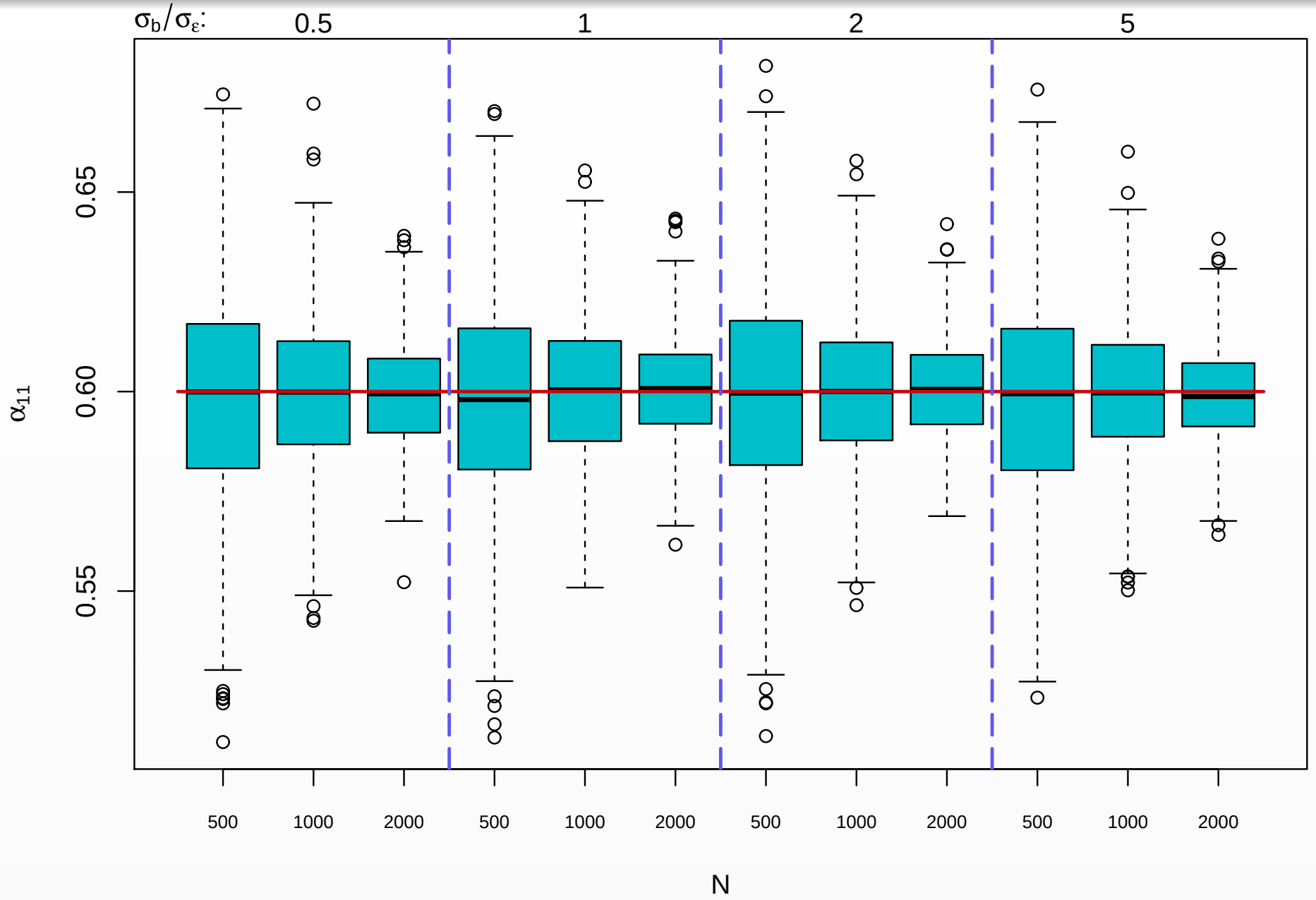
$Q = 5$ examinátorů náhodně přiřazovaných k jednotlivým kontrolám.

Sensitivity a specificity v rozmezí 0.60 – 0.96.

500 datových sad pro každý scénář.

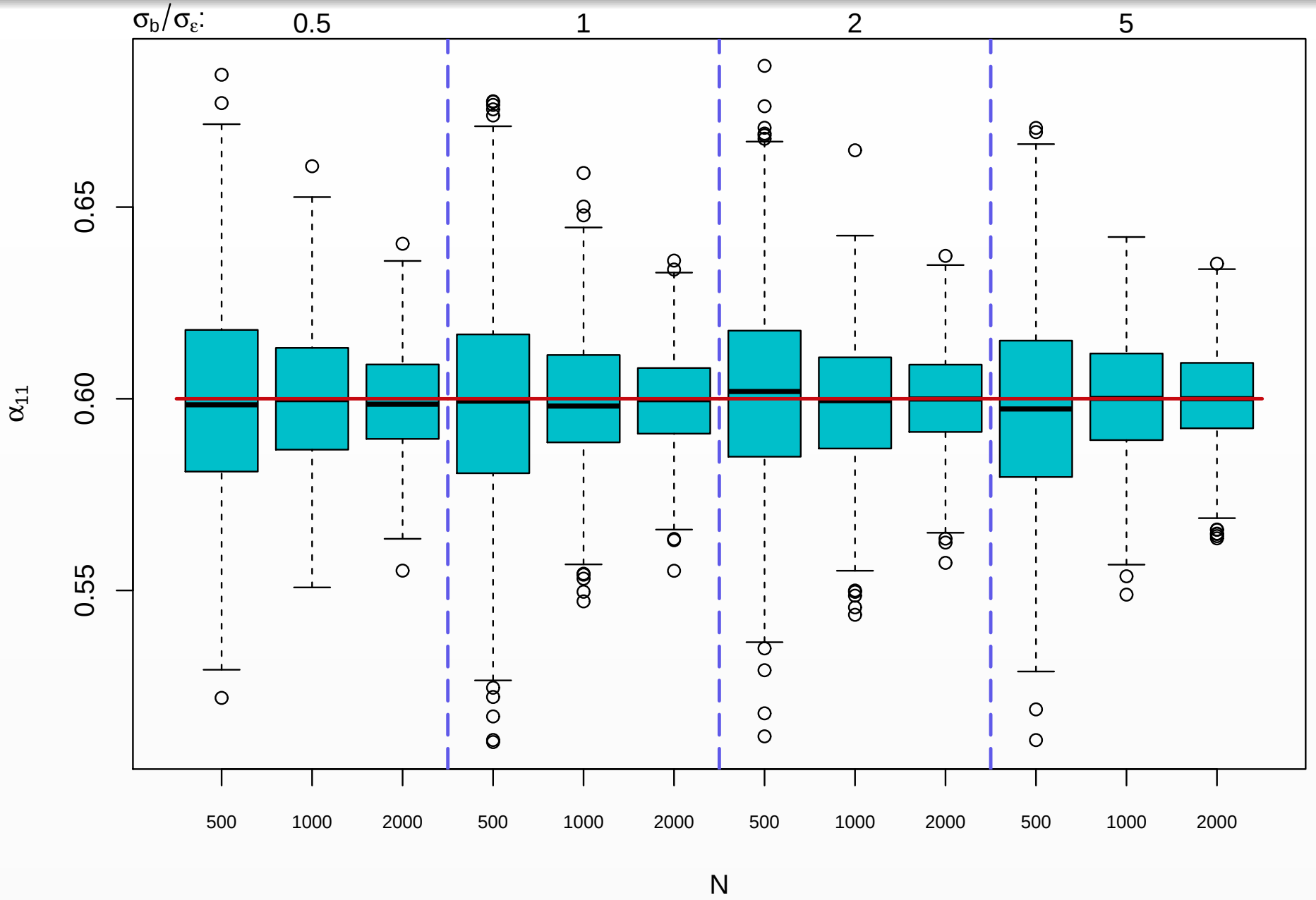
Sensitivita $\alpha_{(1,1)} = 0.60$

g_b: bimodální dvousložková $\mathcal{N}$ směs



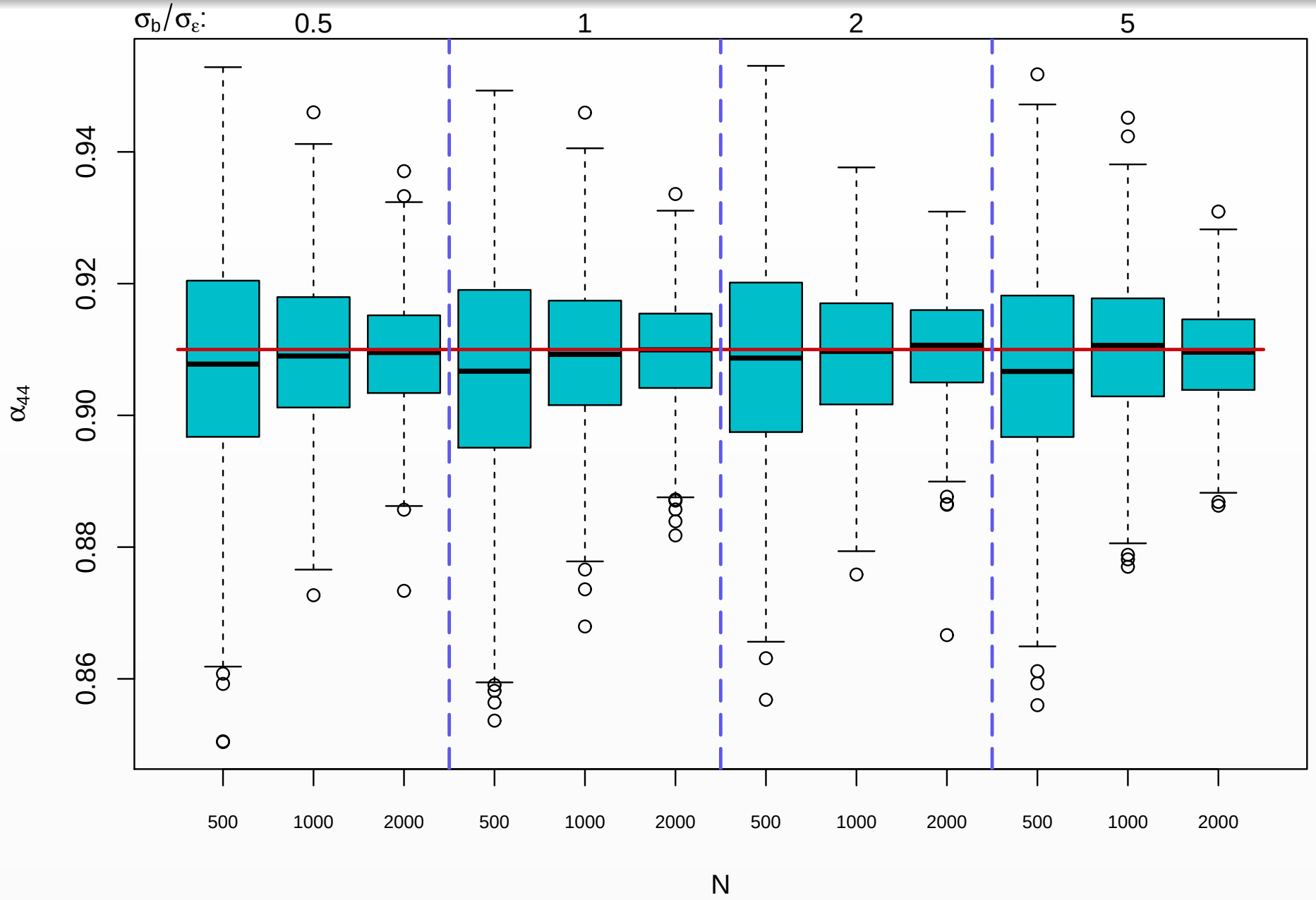
Sensitivita $\alpha_{(1,1)} = 0.60$

g_b: Gumbel



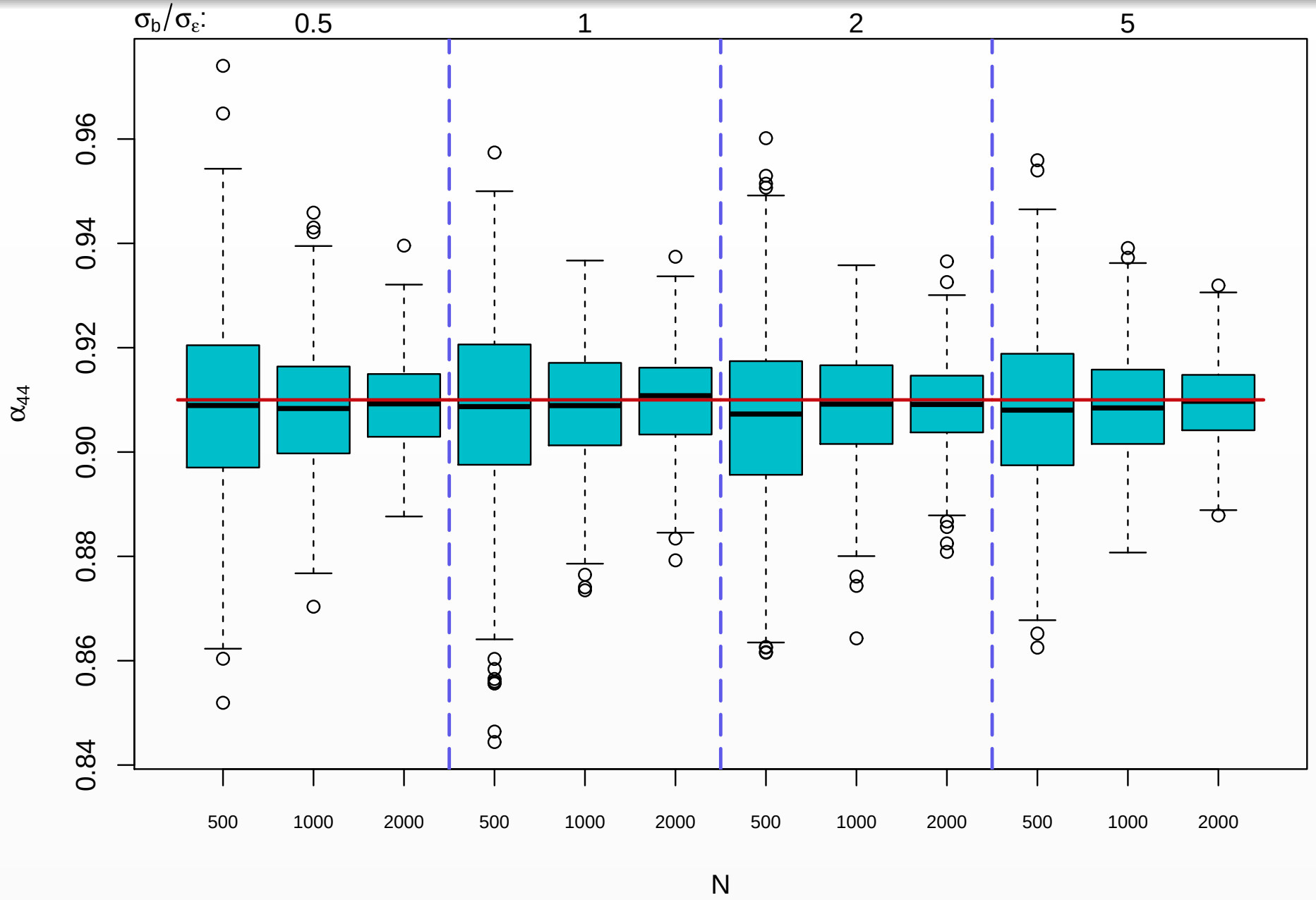
Sensitivita $\alpha_{(4,4)} = 0.91$

g_b: bimodální dvousložková $\mathcal{N}$ směs



Sensitivita $\alpha_{(4,4)} = 0.91$

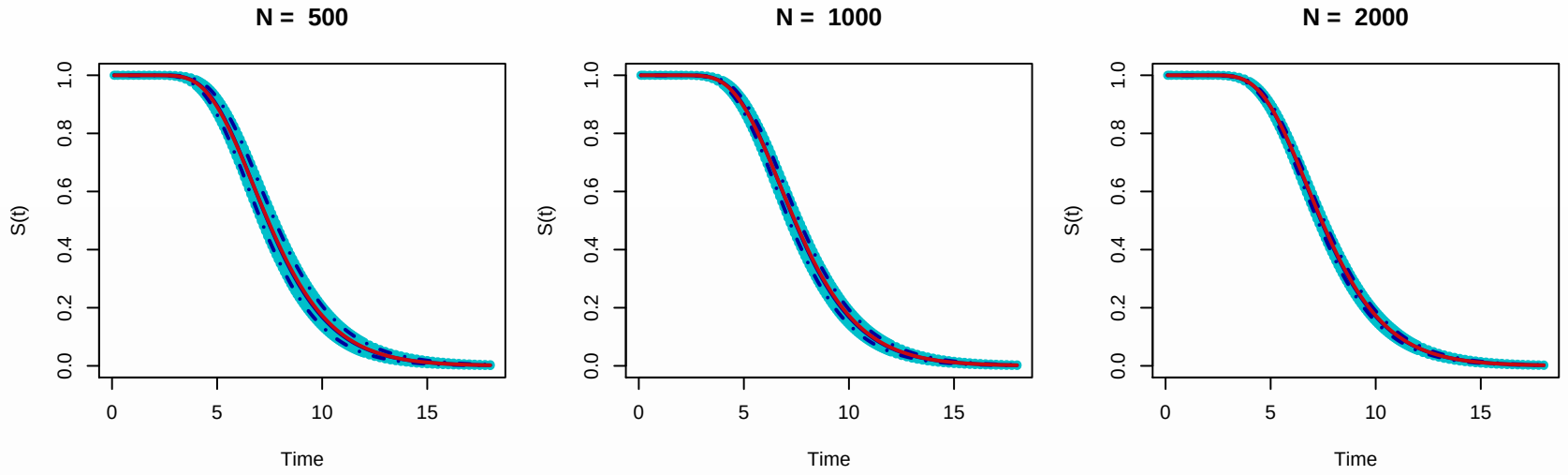
g_b: Gumbel



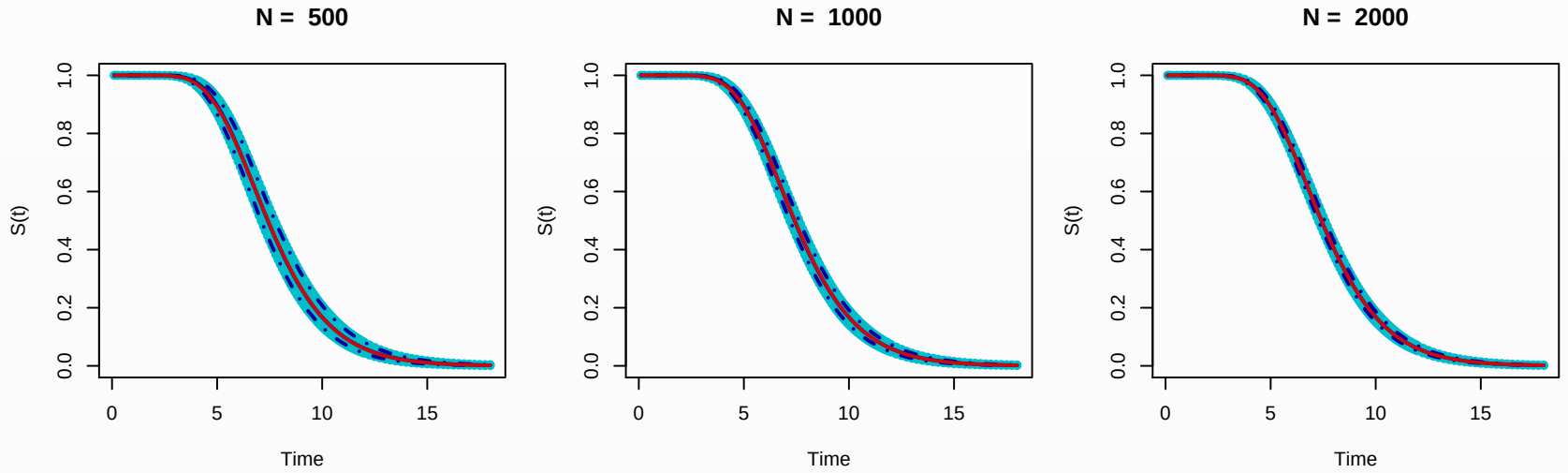
Funkce přežití pro $x_{(i,j),1} = 0.5$, $x_{(i,j),2} = 0$

$\sigma_b/\sigma_\varepsilon = 0.5$

g_b : bimodální dvousložková $\mathcal{N}$ směs



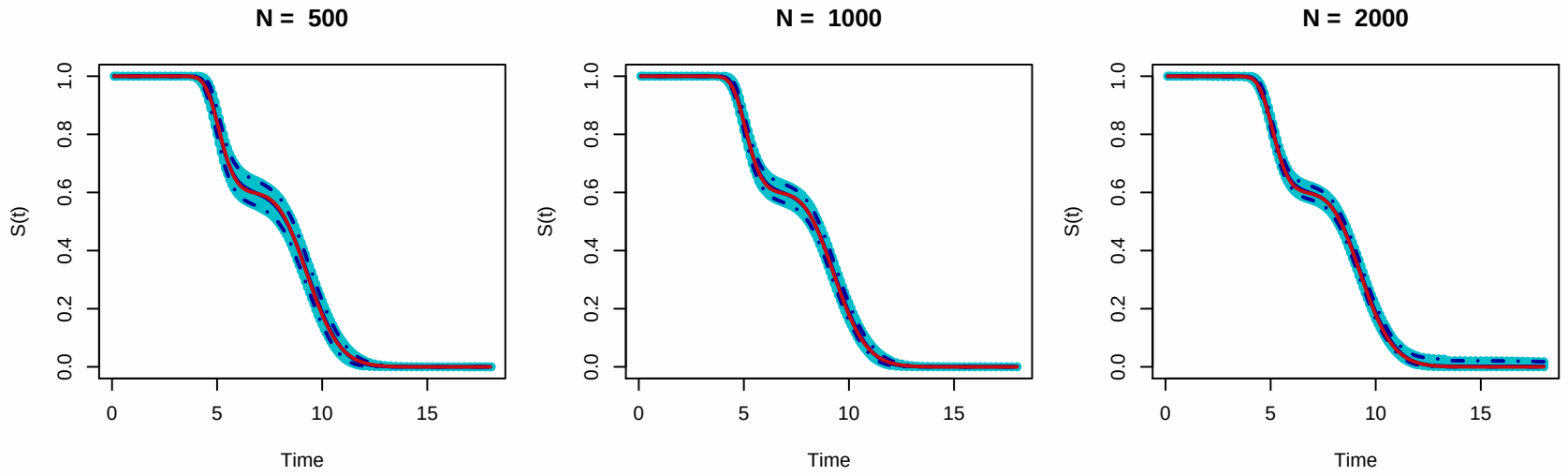
g_b : Gumbel



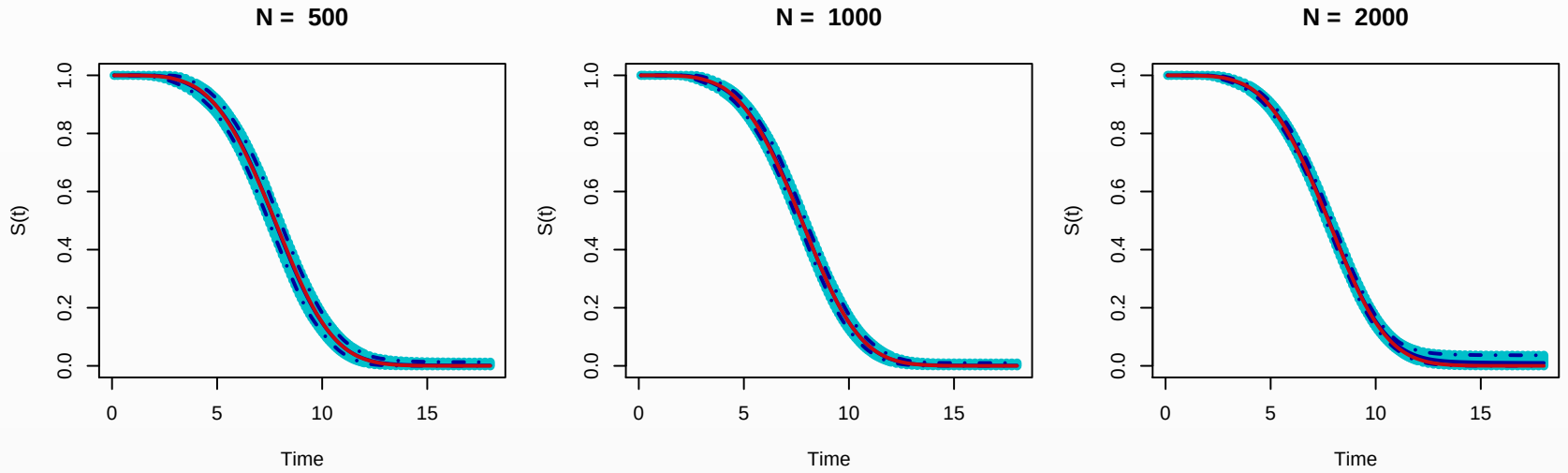
Funkce přežití pro $x_{(i,j),1} = 0.5$, $x_{(i,j),2} = 0$

$\sigma_b/\sigma_\varepsilon = 5$

g_b : **bimodální dvousložková $\mathcal{N}$ směs**



g_b : **Gumbel**



DĚKUJI ZA POZORNOST.

IWSM
International Workshop on Statistical Modelling

SMS
Statistical Modelling Society

29th International Workshop on Statistical Modelling

14. – 18. července/júla 2014

Göttingen

- dlouholeté působiště Carla Friedricha Gausse,
- 388 km po silnici z Jetřichovic.

Termín pro zaslání příspěvků: až 31. ledna/januára 2014
(a stejně bude ještě o minimálně týden posunut).

Zvaní řečníci

- Antoine de Falguerolles *(Université de Toulouse)*
- Alejandro Jara *(Pontificia Universidad Católica de Chile)*
- Sophia Rabe-Hesketh *(University of California, Berkeley)*
- Gerhard Tutz *(Ludwig-Maximilians-Universität München)*
- Simon Wood *(University of Bath)*