INFORMAČNÍ BULLETIN

České statistické společnosti

Ročník 21, číslo 3, srpen 2010

ROBUST 2010

Vybrané práce 16. letní školy JČMF ROBUST 2010,
uspořádané Jednotou českých matematiků a fyziků
za podpory CQR, ČStS a KPMS MFF UK
ve dnech 31. ledna – 5. února 2010 v Králíkách

HORA MATKY BOŽÍ – KLÁŠTER KRÁLÍKY

30. ledna – 5. února 2010

ROBUST 2010 • 16. zimní škola JČMF a ČStS

ROBUST 2010 – PÁR SLOV ÚVODEM

Ve dnech 31. ledna – 5. února 2010 se v areálu kláštera redemptoristů Hora Matky boží v Králíkách uskutečnila šestnáctá zimní škola JČMF ROBUST 2010. Tato akce byla organizována skupinou pro výpočetní statistiku ČMS JČMF za podpory CQR, ČStS a KPMS MFF UK. Tak jako v minulosti, i tentokrát byl ROBUST věnován vybraným trendům matematické statistiky, teorie pravděpodobnosti a analýzy dat. Počet účastníků z čtyř evropských zemí (České republiky, Slovenska, Švýcarska a Velké Británie) přesáhl stovku.

Mezi účastníky bylo k naší velké radosti mnoho mladých tváří. Téměř polovinu účastníků totiž tvořili pregraduální a postgraduální studenti či ti, kteří teprve nedávno obhájili doktorskou práci.

Pozvání přednést přehledné přednášky přijali:

- Prof. RNDr. Jana Jurečková, DrSc., Univerzita Karlova, Praha (CZ).
- Doc. RNDr. Marián Grendár, PhD., Univerzita Mateja Bela, Banská Bystrica (SK).
- RNDr. David Kraus, PhD., Polytechnika, Lausane (CH).
- Dr. Jon McLoone, Wolfram Inc. (UK).
- Doc. RNDr. Ivan Žežula, CSc., Univerzita P. J. Šafárika, Košice (SK).

Vedle toho bylo předneseno 40 delších příspěvků a 31 krátkých sdělení doplněných posterem.

V soutěži o nejlepší práci studentů a doktorandů odborná komise ve složení doc. RNDr. Martin Janžura, CSc., ÚTIA AV ČR, předseda, Ing. Z. Roth, CSc., SZÚ, a doc. RNDr. V. Witkovský, CSc., MFF UKo) ocenila práce následujících doktorandů (v abecedním pořadí):

- Lenka Filová (MFF UK v Bratislavě)
- Jan Kaluža (MFF UK v Praze)
- Stanislav Nagy (MFF UK v Praze)
- Petr Novák (MFF UK v Praze)
- Jakub Petrásek (MFF UK v Praze)
- Jana Timková (MFF UK v Praze)
- Ondřej Vencálek (UPOL Olomouc).

Hodnotné ceny věnovaly společnosti Elkan (`http://www.elkan.cz`) a SAS ČR (`http://www.sas.com/offices/europe/czech/`).

Mnoho času též bylo věnováno diskusím. Pondělní večer byl zasvěcen historii kláštera redemptoristů Hora Matky boží v Králíkách a historii československých opevnění v okolí. Do opevnění Hůrka nás zavedl, díky mimořádné nadílce sněhu až čtvrteční, výlet. Úterní večer byl věnován památce doktora Ivana Saxla a jeho milované historii statistiky a pravděpodobnosti. Ve středu večer vystoupili zástupci firmy *SAS ČR*, kteří předvedli nestandardní možnosti využití jejich programu. Vedle odborných diskusí se též konaly debaty volnější. Za zmínku stojí především čtvrteční večer, který vyplnilo vystoupení skupiny FAB s robustní oporou.

Sborník, s jehož vydáním se původně nepočítalo, vychází tak říkajíc perpartes. Část příspěvků naleznete zde jako třetí číslo letočního Bulletinu ČStS, zbytek pak doufejme jako podzimní číslo časopisu AUC. Z článků, jež byly zaslány do časopise AUC, uveřejňujeme alespoň abstrakty.

ROBUST 2010 by se neuskutečnil a jeho publikace by neexistovaly, nebýt pomoci mnoha lidí. Zvláště bychom chtěli poděkovat všem kteří články recenzovali, paní Haně Bílkové za pomoc s přípravou definitivní verze pro tisk a pracovníkům tiskárny Vězeňské služby v Praze na Pankráci za jeho vytištění. Díky nim vám můžeme popřát příjemné čtení.

V Praze 1. srpna 2010 Jaromír Antoch a Gejza Dohnal

Abstrakty článků, které byly zaslány do časopise AUC

JEDNOFAKTOROVÁ HETEROSKEDASTICKÁ ANOVA - INTERVALY PRE VARIANČNÉ KOMPONENTY

Barbora Arendacká

Kľúčové slová: Zovšeobecnená inferencia, variančné komponenty, heteroskedasticita, nevyvážený ANOVA model.

Abstrakt: V článku sa zaoberáme vlastnosťami a vzájomným porovnaním 3 zovšeobecnených intervalov spoľahlivosti pre medziskupinovú varianciu v heteroskedastickom jednofaktorovom ANOVA modeli s náhodnými efektmi. V krátkosti tiež porovnáme uvažované intervaly s ich homoskedastickými verziami v situácii, keď je analyzovaný model v skutočnosti homoskedastický.

Abstract: The paper focuses on properties and mutual comparison of 3 generalized confidence intervals for the between-group variance in a one-way heteroscedastic ANOVA model with random effects, including a comparison of the considered intervals with their homoscedastic counterpats, when the within-group variances are in fact equal.

1. Úvod

Jednofaktorový heteroskedastický ANOVA model s náhodným efektom sa využíva pri spájaní meraní rovnakej kvantity získaných z viacerých zdrojov/laboratórií, pozri napr. [5]. Variabilita jednotlivých pozorovaní sa potom skladá z variability medzi jednotlivými laboratóriami a z variability v príslušnom laboratóriu, pričom model umožňuje zachytiť často realistický predpoklad, že variability v jednotlivých laboratóriách sú rôzne. Uvažujeme teda model

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad i = 1, ..., k \geq 2, \quad j = 1, ..., n_i \geq 2 \tag{1}$$

kde realizáciou náhodnej premennej Y_{ij} je j-te pozorovanie v i-tom laboratóriu, $\alpha_i \sim N(0, \sigma_A^2)$, $i = 1, ..., k$, sú navzájom nezávislé náhodné efekty a $\epsilon_{ij} \sim N(0, \sigma_i^2)$, $i = 1, ..., k$, $j = 1, ..., n_i$, sú navzájom nezávislé náhodné chyby, nezávislé tiež od náhodných efektov. Parameter $\mu(\in R)$ je neznáma spoločná hodnota a o variančných komponentoch predpokladáme, že $\sigma_A^2 \geq 0$ a $\sigma_i^2 > 0$, $i = 1, ..., k$. V štandardnom maticovom zápise potom pre vektor pozorovaní Y máme

$$Y \sim N(1_n \mu, \sigma_A^2 Z Z^T + diag\{\sigma_1^2 I_{n_1}, ..., \sigma_k^2 I_{n_k}\}) \tag{2}$$

kde $n = \sum_{i=1}^{k} n_i$, 1_n označuje $n \times 1$ vektor jednotiek a $Z = diag\{1_{n_1}, ..., 1_{n_k}\}$. Okrem odhadu spoločnej hodnoty meranej kvantity (μ), môže byť tiež žiaduce odhadnúť veľkosť medzilaboratórnej variability (σ_A^2). V ďalšom sa budeme zaoberať práve intervalovými odhadmi pre σ_A^2, pričom sa zameriame na intervaly odvodené metódou zovšeobecnenej (fiduciálnej) inferencie, pozri [4, 3, 7].

Pri odvodzovaní zovšeobecnených konfidenčných intervalov sa vychádza zo systému štrukturálnych alebo pivotálnych rovníc. Štrukturálne rovnice popisujú mechanizmus generovania dát, t.j. pre náhodný vektor X, ktorého distribúcia závisí na neznámych parametroch θ, majú tvar $X = g(U, \theta)$, kde g je merateľná funkcia a U je náhodný vektor so známou distribúciou nezávislou na θ (pozri [3]). V najjednoduchšom prípade má systém jediné riešenie v θ: $\theta = g^{-1}(X, U)$, ktoré pre dané, napozorované dáta x určuje zovšeobecnené fiduciálne rozdelenie pre θ ako rozdelenie $g^{-1}(x, U^*)$, kde U^* označuje nezávislú kópiu U. Jednotlivé zložky $g^{-1}(X, U^*)$ definujú zovšeobecnené fiduciálne pivoty pre jednotlivé zložky parametra θ. Zovšeobecnený konfidenčný interval pre napr. prvú zložku θ_1 tvorí príslušný dolný a horný kvantil podmieneného rozdelenia $g_1^{-1}(X, U^*)$ pri danom X, kde index 1 označuje prvú zložku $g^{-1}(\cdot, \cdot)$, pozri tiež [4]. V prípade pivotálnych rovníc, $F(X, \theta) = U$, je situácia obdobná. V konkrétnych prípadoch X predstavuje buď priamo napozorované dáta, alebo štatistiky založené na napozorovaných dátach. Druhá možnosť je nevyhnutná, ak chceme na odvodenie intervalov použiť systém rovníc s jediným riešením v neznámych parametroch.

Ak je parametrický priestor ohraničený, napr. $\theta_1 \geq 0$, môže sa stať, že pre niektoré hodnoty X a U^* bude $g_1^{-1}(X, U^*) < 0$, t.j. zovšeobecnené fiduciálne rozdelenie pre θ_1 bude zahŕňať aj hodnoty, ktoré príslušný parameter nemôže nadobúdať. Jednou z možností, ako sa s touto situáciou vyrovnať, je presunúť pravdepodobnosť na záporných číslach na hranicu parametrického priestoru, t.j. do nuly (pozri [3], Remark 9). To zodpovedá tomu, že namiesto $g_1^{-1}(X, U^*)$ uvažujeme $\max(0, g_1^{-1}(X, U^*))$, čo je pri konštrukcii intervalov to isté, ako položiť prípadné záporné hranice rovné nule. S takouto situáciou sa stretneme aj v našom prípade, keďže variančné komponenty sú nezáporné.

Pretože v modeli (2) sú parametre $\sigma_A^2, \sigma_1^2, ..., \sigma_k^2$ invariantné na posunutie v strednej hodnote, môžme najprv situáciu zjednodušiť uplatnením princípu invariancie. Prejdeme tak k modelu

$$\widetilde{Y} = B^T Y \sim N(0, \sigma_A^2 B^T Z Z^T B + B^T diag\{\sigma_1^2 I_{n_1}, ..., \sigma_k^2 I_{n_k}\} B) \tag{3}$$

kde $B^T B = I_{n-1}$, $BB^T = I - 1_n 1_n^T / n$. Následne potrebujeme nájsť (k+1) štrukturálnych (alebo pivotálnych) rovníc, ktoré budú založené na $\widetilde{Y}$ a budú mať jediné riešenie v σ_A^2, σ_1^2, ..., σ_k^2 v parametrickom priestore. (Riešenia mimo parametrického priestoru posunieme do nuly.) V ďalšej časti uvedieme rovnice navrhnuté v [8, 6].

2. Zovšeobecnené pivoty pre σ_A^2

Wimmer, Witkovský [8] navrhli zovšeobecnený pivot pre σ_A^2 založený na nasledujúcom systéme pivotálnych rovníc

$$\text{WS} = Q_0, \;\; S_1^2/\sigma_1^2 = Q_1, ..., S_k^2/\sigma_k^2 = Q_k \tag{4}$$

kde $S_i^2 = \sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_i)^2$, $\bar{Y}_i = \sum_{j=1}^{n_i} Y_{ij}/n_i$, $i = 1, ..., k$, sú úmerné výberovej variancii pozorovaní v i-tom laboratóriu, WS je vážená suma štvorcov

$$(5)\quad \mathrm{WS}(\sigma_A^2, \sigma_1^2, ..., \sigma_k^2, \bar{Y}) = \sum_{i=1}^{k} \frac{1}{\sigma_A^2 + \frac{\sigma_i^2}{n_i}} \left(\bar{Y}_i - \frac{\sum_{j=1}^{k} \bar{Y}_j/(\sigma_A^2 + \sigma_j^2/n_j)}{\sum_{j=1}^{k} 1/(\sigma_A^2 + \sigma_j^2/n_j)} \right)^2$$

$\bar{Y} = (\bar{Y}_1, ..., \bar{Y}_k)^T$ a $Q_0 \sim \chi^2_{k-1}$, $Q_i \sim \chi^2_{n_i-1}$, $i = 1, ..., k$, sú navzájom nezávislé. WS, S_1^2,..., S_k^2 závisia na Y iba cez $\widetilde{Y}$, keďže sú invariantné vzhľadom na posunutie Y v strednej hodnote.

Riešenie $(R_A, R_1, ..., R_k)$ systému (4) v neznámych parametroch s Q_i, $i = 0, ..., k$, nahradenými ich nezávislými kópiami Q_i^* je

$$(6)\quad \mathrm{WS}(R_A, S_1^2/Q_1^*, ..., S_k^2/Q_k^*, \bar{Y}) = Q_0^*,\ R_1 = S_1^2/Q_1^*, ...,\ R_k = S_k^2/Q_k^*$$

kde R_A je dané implicitne. Jednotlivé R_i, $i = 1, ..., k$, (všimnime si, že vždy platí $R_i > 0$) sú zovšeobecnené pivoty pre σ_i^2. Intervaly pomocou nich skonštruované sa zhodujú s klasickými konfidenčnými intervalmi pre σ_i^2 založenými na S_i^2. Čo sa týka zošeobecneného pivotu R_A pre σ_A^2, keďže WS je v σ_A^2 klesajúca na $[0, \infty)$, tak ak $\mathrm{WS}(0, S_1^2/Q_1^*, ..., S_k^2/Q_k^*, \bar{Y}) \geq Q_0^*$, R_A vyhovujúce prvej rovnici v (6) je jediné na $[0, \infty)$. V opačnom prípade je riešenie prvej rovnice v (6) záporné, a teda mimo parametrického priestoru, preto vtedy kladieme R_A rovné nule, viď str. 2. Zovšeobecnený pivot R_A^{WW} pre σ_A^2 je teda definovaný ako nezáporné riešenie alebo nula (ak nezáporné riešenie neexistuje)

$$(7)\quad \sum_{i=1}^{k} \frac{1}{R_A^{WW} + S_i^2/(n_i Q_i^*)} \left(\bar{Y}_i - \frac{\sum_{j=1}^{k} \bar{Y}_j/(R_A^{WW} + S_j^2/(n_j Q_j^*))}{\sum_{j=1}^{k} 1/(R_A^{WW} + S_j^2/(n_j Q_j^*))} \right)^2 = Q_0^*.$$

Li [6] navrhol zovšeobecnený pivot pre σ_A^2 pomocou systému štrukturálnych rovníc

$$(8)\quad S_A^2 = W_0^T W_0 + W_0^T H_2^T diag\{\sigma_i^2/n_i\} H_2 W_0,\ S_1^2 = \sigma_1^2 Q_1, ...,\ S_k^2 = \sigma_k^2 Q_k$$

kde $W_0 \sim N(0, I_{k-1})$, $Q_i \sim \chi^2_{n_i-1}$, $i = 1, ..., k$, sú navzájom nezávislé, $H_2 H_2^T = I - 1_k 1_k^T/k$, $H_2^T H_2 = I_{k-1}$ a $S_A^2 = \sum_{i=1}^{k}(\bar{Y}_i - \sum_{j=1}^{k} \bar{Y}_j/k)^2$ je nevážená suma štvorcov (závisí na Y iba cez $\widetilde{Y}$). Prvá rovnica odráža rozdelenie S_A^2. Opäť, vyriešením systému (8) v neznámych parametroch a nahradením W_0 a $Q_1, ..., Q_k$ ich nezávislými kópiami, dostaneme zovšeobecnený pivot pre σ_A^2 v tvare

$$(9)\quad R_A^{Li} = (S_A^2 - W_0^{*T} H_2^T diag\{S_i^2/(n_i Q_i^*)\} H_2 W_0^*)/W_0^{*T} W_0^*$$

resp. aby sme dostávali iba hodnoty z parametrického priestoru, uvažujeme $\max(0, R_A^{Li})$.

Dá sa ukázať, že $S_A^2 = \sum_{i=1}^{r-1} \widetilde{Y}^T E_i \widetilde{Y}/\lambda_i$, kde λ_i, $i = 1, ..., r-1$, sú navzájom rôzne, nenulové vlastné čísla $B^T Z Z^T B$ a E_i je projektor na podpriestor generovaný vlastnými vektormi prislúchajúcimi k λ_i, pozri [1]. To jednak poukazuje na spojitosť so zovšeobecnenými pivotmi založenými na

sumách $\sum_{i=1}^{r-1} c_i \widetilde{Y}^T E_i \widetilde{Y}$ pre rôzne kladné konštanty c_i v homoskedastickom prípade modelu, a jednak to naznačuje možnosť inej voľby štrukturálnych rovníc (8). Napr. namiesto rovnice s S_A^2, by sme mohli uvažovať rovnicu s $S_I^2 = \sum_{i=1}^{r-1} \widetilde{Y}^T E_i \widetilde{Y} = \widetilde{Y}^T P_{B^T Z Z^T B} \widetilde{Y}$ (P_K označuje projektor na priestor generovaný stĺpcami matice K):

$$S_I^2 = W_0^T \Lambda W_0 + W_0^T B_V^T B^T Z diag\{\sigma_i^2/n_i\} Z^T B B_V W_0 \tag{10}$$

kde stĺpce B_V sú postupne vlastné vektory patriace k $\lambda_1, ..., \lambda_{r-1}$ (a teda $B_V B_V^T = P_{B^T Z Z^T B}$, $B_V^T B_V = I_{k-1}$), $\Lambda = B_V^T B^T Z Z^T B B_V = diag\{\lambda_i 1_{\nu_i}\}$, kde ν_i je násobnosť λ_i, a $W_0 = (B_V^T Var(\widetilde{Y}) B_V)^{-1/2} B_V^T \widetilde{Y} \sim N(0, I_{k-1})$. Využili sme tiež, že $P_{B^T Z Z^T B} B^T (diag\{\sigma_i^2 I_{n_i}\} - Z diag\{\sigma_i^2/n_i\} Z^T) B = 0$. Zovšeobecnený pivot pre σ_A^2 založený na systéme rovníc (8) s prvou rovnicou nahradenou (10) je

$$R_A^I = (S_I^2 - W_0^{*T} B_V^T B^T Z diag\{S_i^2/(n_i Q_i^*)\} Z^T B B_V W_0^*)/W_0^{*T} \Lambda W_0^* \tag{11}$$

resp. $\max(0, R_A^I)$.

Pozn. Dá sa ukázať, že všetky tri uvedené zovšeobecnené pivoty sú založené na minimálnej postačujúcej štatistike pre $(\sigma_A^2, \sigma_1^2, ..., \sigma_k^2)$ v triede rozdelení (3) a tiež, že všetky patria do dvoch širších tried zovšeobecnených pivotov, ktoré sú analogické triedam navrhnutým v zmiešanom lineárnom modeli s dvomi variančnými komponentmi, ktorého príkladom je aj homoskedastická verzia uvažovaného modelu. (Pozri [2].)

3. Vlastnosti

Dôležitou vlastnosťou intervalových odhadov je ich pravdepodobnosť pokrytia. V prípade zovšeobecnených intervalov však nie je garantované, že ich pravdepodobnosť pokrytia bude na nominálnej úrovni. Dokonca platí, že ak je zovšeobecnený interval frekventisticky presný, tak v danom probléme existuje klasický (presný) konfidenčný interval pre parameter záujmu, viď [4], str. 259. Pre praktické použitie ale stačí, ak je pravdepodobnosť pokrytia blízka požadovanej úrovni, čo sa v prípade zovšeobecnených intervalov demonštruje pomocou simulácií. Zároveň sa dá obvykle dokázať, že zovšeobecnené intervaly sú presné aspoň asymptoticky, napr. pre narastajúci počet pozorovaní. Pozri tiež [4, 3].

O intervaloch založených na R_A^{WW}, R_A^{Li}, R_A^I sa dá ukázať (pozri [2]), že ich pravdepodobnosť pokrytia konverguje k nominálnej úrovni $1-\alpha$ ak $\sigma_A^2/\sigma_i^2 \to \infty$, $i = 1, ..., k$, a tiež ak $n_i \to \infty$, $n_i/n \to d_i > 0$, $i = 1, ..., k$ a skutočné σ_A^2 je kladné. V prípade $\sigma_A^2 = 0$, konverguje pravdepodobnosť pokrytia pri narastajúcom počte pozorovaní v skupinách k hodnote vyššej ako $1-\alpha$, keďže všetky intervaly sú konštruované tak, že záporné hranice kladieme rovné nule. To zvyšuje pravdepodobnosť pokrytia nuly, ktorá potom patrí do (niekedy degenerovaného) intervalu, ak je dolná hranica nulová (bez ohľadu na hodnotu hornej hranice). Pre konečné počty pozorovaní v skupinách, v homoskedastickom prípade náprotivky uvažovaných zovšeobecnených

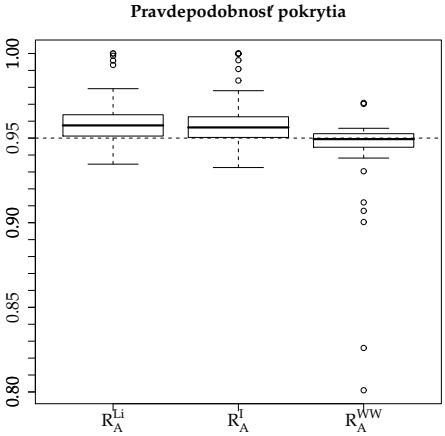
	n_i	σ_i^2
1^*	2, 4, 6, 8, 10, 12, 14, 16, 18, 20	.001, .01, .1, 1, 5, 10, 15, 20, 30, 40
2^*	20, 18, 16, 14, 12, 10, 8, 6, 4, 2	.001, .01, .1, 1, 2, 5, 7, 10, 15, 20
3^*	$n_i = 5$, $i = 1, ..., 20$	.001, .005, .01, .05, .1, .5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 5, 7, 10, 13, 15, 17, 20
4^*	$n_i = 15$, $i = 1, ..., 20$	.001, .01, .1, 1, 2, 3, 4, 5, 7, 10, 13 15, 17, 20, 25, 30, 35, 40, 45, 50
5	50, 60, 70, 80, 90, 50, 60, 70, 80, 90	.001, .01, .1, 1, 5, 10, 20, 30, 50, 90
6^*	20, 19, 18, 17, 16, 15, 14, 13, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, 2	$\sigma_i^2 = 1$, $i = 1, ..., 20$
$7'$	$n_1 = 10$, $n_i = 100$, $i = 2, ..., 10$	$\sigma_i^2 = 1$, $i = 1, ..., 10$

TABUL'KA 1. Dizajny použité v simulačnej štúdii. Symbol $*$ označuje dizajny z [8], symbol $'$ dizajny z [6].

intervalov pokrývajú nulu v súlade s presnými testami o nulovosti σ_A^2 (hoci na hladine významnosti nižšej ako α, a teda s pravdepodobnosťou pokrytia $> 1 - \alpha$). V heteroskedastickom prípade však takáto optimalita nie je zaručená. Už v [8] pri návrhu R_A^{WW} autori upozornili, že intervaly založené na tomto pivote môžu dosahovať výrazne nižšiu ako požadovanú pravdepodobnosť pokrytia pre nulu a malé hodnoty σ_A^2 v niektorých modeloch s vyšším počtom skupín. Intervaly založené na R_A^{Li} boli simulačne skúmané v [6], kde sa ukázalo, že majú uspokojivú pravdepodobnosť pokrytia vo všetkých uvažovaných modeloch a ich priemerná dĺžka bola menšia ako priemerná dĺžka približného intervalu, s ktorým boli porovnávané. V [2] boli v krátkosti ilustrované vlastnosti všetkých troch procedúr v troch konkrétnych prípadoch modelu (1).

V tu prezentovanej simulačnej štúdii ukážeme správanie sa R_A^{WW}, R_A^{Li}, R_A^I v ďalších 7 dizajnoch (pozri Tab. 1), ktoré zahŕňajú nevyvážené a vyvážené heteroskedastické modely (1-5), ako aj modely homoskedastické (6, 7), v prípade ktorých porovnáme R_A^{WW}, R_A^{Li}, R_A^I s ich homoskedastickými náprotivkami (tie získame zo (7), (9), (11), keď S_i^2/Q_i^* nahradíme S^2/Q^*, kde $S^2 = \sum_{i=1}^k S_i^2$ a $Q^* \sim \chi_{n-k}^2$.) Ešte poznamenajme, že vo vyvážených modeloch (3, 4) sa intervaly založené na R_A^I a R_A^{Li} zhodujú, keďže $B^T Z Z^T B$ má len 1 nenulovú vlastnú hodnotu (pozri ich prepis cez sumu kvadratických foriem s maticami E_i na str. 4).

Výsledky, ktoré ďalej uvedieme sú pre každý dizajn a každú hodnotu σ_A^2 (=0, .1, .5, 1, 5, 10) založené na 5000 simulovaných intervaloch. T.j. zakaždým sme nagenerovali 5000 vektorov Y podľa modelu (1) a pre každú realizáciu sme spočítali príslušný zovšeobecnený konfidenčný interval založený na R_A^{WW}, resp. R_A^{Li}, resp. R_A^I. Každý takýto interval pre dané Y sme dostali ako 2.5-tý a 97.5-tý empirický percentil na základe 10000 nasimulovaných hodnôt R_A^{WW}, resp. R_A^{Li}, resp. R_A^I. V prípade R_A^{Li}, resp. R_A^I sme simulovali 10000-krát $W_0^*, Q_1^*, ..., Q_k^*$ a vždy vyčíslili hodnotu zovšeobecneného pivotu



Obrázok 1. Pravdepodobnosť pokrytia pre jednotlivé intervaly vo všetkých uvažovaných dizajnoch a pre všetky uvažované hodnoty σ_A^2.

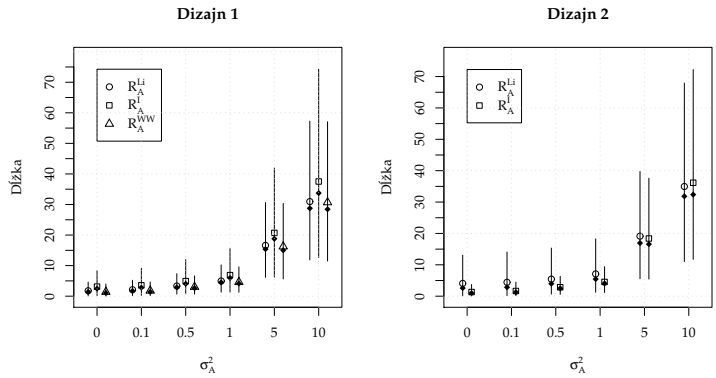
	Dizajn 5					Dizajn 1				
σ_A^2	0.1	0.5	1	5	10	0.1	0.5	1	5	10
R_A^{Li}	0.946	0.597	0.231	0.002	0.000	0.986	0.949	0.814	0.099	0.012
R_A^I	0.952	0.678	0.316	0.004	0.001	0.983	0.958	0.886	0.262	0.062
R_A^{WW}	0.007	0	0	0	0	0.362	0.062	0.017	0.000	0

Tabul'ka 2. Pravdepodobnosť pokrytia nuly jednotlivými intervalmi.

podľa (9), resp. (11). V prípade R_A^{WW} sme pre každé Y generovali 10000-krát $Q_0^*, Q_1^*, ..., Q_k^*$ a hodnotu R_A^{WW} sme našli vyriešením (7) na $[0, \infty)$ pomocou Newtonovej-Raphsonovej metódy s presnosťou 10^{-4}.

Obr. 1 zobrazuje pozorované pravdepodobnosti pokrytia pre jednotlivé intervaly vo všetkých uvažovaných dizajnoch a pre všetky uvažované hodnoty σ_A^2. Kým intervaly založené na R_A^{Li} a R_A^I majú pravdepodobnosť pokrytia vo všetkých situáciách uspokojivú, intervaly založené na R_A^{WW} občas zlyhávajú pre nulové a malé hodnoty σ_A^2 (hodnoty pod 0.94 boli pre R_A^{WW} napozorované v dizajnoch 2, 3, 4, 6 pre $\sigma_A^2 = 0$ a/alebo $\sigma_A^2 = 0.1$, v dizajne 2 aj pre $\sigma_A^2 = 0.5$, pre R_A^{Li} a R_A^I iba v dizajne 6 pre $\sigma_A^2 = 0.1$).

Práve pokrytie nuly je črta, v ktorej sa uvažované intervaly líšia aj v situáciách, keď skutočnú nenulovú hodnotu parametra σ_A^2 pokrývajú na uspokojivej úrovni. Tab. 2 zobrazuje pokrytie nuly, keď je skutočný parameter nenulový, v dizajnoch 5 a 1 (pravdepodobnosti pokrytia skutočnej hodnoty σ_A^2



Obrázok 2. Pozorované dĺžky jednotlivých intervalov. Symboly (◯, □, △) označujú priemerné hodnoty, bodky mediány a úsečky spájajú 5. a 95. percentily.

(vrátane skutočnej nuly) boli pre všetky intervaly > 0.944). Keďže nula predstavuje neprítomnosť medzilaboratórnej variability, môže byť jej vylúčenie z intervalu spoľahlivosti, keď je skutočná medzilaboratórna variabilita nenulová, žiaduce. Obr. 2 sumarizuje dĺžky jednotlivých intervalov v dizajnoch 1 a 2 (v dizajne 2 neuvádzame výsledky pre R_A^{WW}, keďže len pre $\sigma_A^2 = 1, 5, 10$ bola napozorovaná pravdepodobnosť pokrytia > 0.94). Je zrejmé, že z hľadiska dĺžky, nie je medzi jednotlivými intervalmi jednoznačný víťaz.

Na záver sa ešte pozrime na dizajny 6 a 7. Ide o homoskedastické verzie uvažovaného modelu, v ktorých by sme mohli použiť homoskedastcké náprotivky uvažovaných intervalov. Tab. 3 ilustruje vplyv dodatočnej informácie o homoskedasticite na vlastnosti výsledných intervalov. Podľa očakávania, vďaka pomerne vysokým počtom pozorovaní v skupinách v dizajne 7 nepozorujeme veľké rozdiely medzi homoskedastickými a heteroskedastickými verziami. V dizajne 6 vedie dodatočná informácia o homoskedasticite k zlepšeniu pravdepodobnosti pokrytia v prípade R_A^{WW} a k zníženiu pravdepodobnosti pokrytia nuly, keď je skutočná medziskupinová variabilita nenulová, v prípade intervalov R_A^I, R_A^{Li}. Čo sa dĺžky intervalov týka, správne použitie homoskedastických verzií R_A^{Li} a R_A^I viedlo (v dizajne 6, v dizajne 7 neboli rozdiely výrazné) ku skráteniu intervalov pre väčšie hodnoty σ_A^2: postupne pre $\sigma_A^2 = 0, 0.1, 0.5, 1, 5, 10$ a heteroskedastickú (homoskedastickú) verziu R_A^{Li} boli priemerné dĺžky intervalov 0.15(0.19), 0.33(0.39), 1.11(1.04), 2.12(1.80), 10.24(7.97), 20.29(15.75). Pre R_A^I v rovnakom značení boli priemerné dĺžky: 0.09(0.12), 0.29(0.32), 1.16(1.04), 2.28(1.94), 10.74(9.08), 20.72 (18.15). V prípade R_A^{WW} a uvažovaných dizajnov neboli rozdiely v dĺžke veľmi výrazné.

Dizajn 6		Pravdep. pokrytia						Pokrytie 0			
	σ_A^2	0	0.1	0.5	1	5	10	0.1	0.5	1	10
R_A^{WW}	He	0.826	0.938	0.945	0.953	0.956	0.952	0.249	0.002	0	0
	Ho	0.973	0.944	0.944	0.951	0.952	0.952	0.401	0.003	0	0
R_A^{Li}	He	1	0.935	0.940	0.952	0.967	0.968	1	0.999	0.996	0.845
	Ho	0.969	0.946	0.944	0.950	0.952	0.952	0.78	0.045	0.001	0
R_A^{I}	He	1	0.933	0.944	0.955	0.964	0.967	1	0.981	0.949	0.429
	Ho	0.973	0.943	0.942	0.95	0.948	0.956	0.400	0.003	0	0
Dizajn 7											
R_A^{WW}	He	0.970	0.951	0.947	0.951	0.954	0.951	0.008	0	0	0
	Ho	0.972	0.953	0.949	0.950	0.954	0.952	0.008	0	0	0
R_A^{Li}	He	0.976	0.956	0.949	0.952	0.954	0.952	0.298	0.005	0	0
	Ho	0.977	0.952	0.948	0.951	0.954	0.951	0.131	0.001	0	0
R_A^{I}	He	0.984	0.953	0.950	0.951	0.955	0.952	0.012	0	0	0
	Ho	0.973	0.952	0.950	.951	0.954	0.951	0.008	0	0	0

TABUL'KA 3. Pravdepodobnosti pokrytia skutočnej hodnoty σ_A^2 a nuly pre heteroskedastické (He) a homoskedastické (Ho) verzie jednotlivých intervalov.

Literatúra

[1] Arendacká B. (2006) *Approximate confidence intervals on the variance component in a general case of a two-component model.* Sborník prací 14. zimní školy JČMF ROBUST 2006, JČMF, Praha, 9 – 16.

[2] Arendacká B. *A note on fiducial generalized pivots for σ_A^2 in one-way heteroscedastic ANOVA with random effects.* Zaslané na publikovanie.

[3] Hannig J. (2009) *On generalized fiducial inference.* Statistica Sinica, **19**, 491 – 544.

[4] Hannig J., Iyer H., Patterson P. (2006) *Fiducial generalized confidence intervals.* JASA, **101**, 254 – 269.

[5] Iyer H.K., Wang C.M.J., Mathew T. (2004) *Models and confidence intervals for true values in interlaboratory trials.* JASA, **99**, 1060 – 1071.

[6] Li X. (2007) *Comparison of confidence intervals on between group variance in unbalanced heteroscedastic one-way random models.* Commun. in Stat. - Simulation and Computation, **36**, 381 – 390.

[7] Weerahandi S. (1993) *Generalized confidence intervals.* JASA, **88**, 899 – 905.

[8] Wimmer G., Witkovský V. (2003) *Between group variance component interval estimation for the unbalanced heteroscedastic one-way random effects model.* J. of Stat. Computation and Simulation, **73**, 333 – 346.

Poďakovanie: Táto práca bola podporená grantom č. LPP-0388-09 poskytnutým Agentúrou na podporu výskumu a vývoja a čiastočne grantom VEGA 1/0077/09.

Adresa: Ústav merania, SAV, Dúbravská cesta 9, 841 04 Bratislava

E-mail: barendacka@gmail.com

KLASIFIKÁCIA PRE RÔZNE TVARY ŠUMU VSTUPNÝCH DÁT

Katarína Cimermanová

Kľúčové slová: Robustná klasifikačná metóda, zašumené dáta, analýza dychu, fajčiarsky návyk.

Abstrakt: Klasifikácia viacrozmerných pozorovaní do jednej z dvoch tried je dôležitý problém. Existuje niekoľko klasifikačných metód riešiacich daný problém, avšak v reálnom živote sú vektory pozorovaní zašumené. Riešením klasifikácie zašumených dát je robustná formulácia vychádzajúca z metódy oporných bodov. Formulácia je konvexný optimalizačný problém, ktorý je súčasťou problematiky kónického programovania druhého rádu. V robustnej formulácii sa predpokladá elipsoidálny model šumu. Nie je nutný predpoklad typu rozdelenia pozorovaných dát, predpokladá sa len konečnosť momentov druhého rádu.

Robustnú klasifikačnú metódu aplikujeme na analýzu vydychovaných plynov. Klasifikované dáta v sebe zahŕňajú variabilitu opakovane nameraných dát pozorovaných subjektov, označujeme ich ako zašumené dáta. V práci sa venujeme klasifikácii dobrovoľníkov do skupiny fajčiarov a nefajčiarov za predpokladu rôznych typov elipsoidálneho šumu.

Abstract: Classification of multidimensional data into one of two classes is an important issue. There are some classification methods which classify data into one of two classes, but in a real live situation the observation vectors are noisy. Solution to this problem is a robust formulation that stems from the Support Vector Machine method. The formulation is a convex optimization problem; in particular, it is an instance of the Second Order Cone Programming problem. An ellipsoidal uncertainty model is assumed in the robust formulation. It is derived from the worst case consideration and assumes only the existence of the second order moments.

The robust classification method is applied to breath gas analysis. Classified data include variability of repetitive measurements of subjects, noisy data. In this paper we classify volunteers into group of smokers and non-smokers based on assumption of different shapes of noise.

1. Úvod

V súčasnosti sa rozvojom nových analytických techník dá vo vydychovanom vzduchu detegovať 3481 rôznych zlúčenín, z čoho 1753 zlúčenín má pozitívny alveolárny gradient [2], teda koncentrácia zlúčeniny vo vydychovanom vzduchu je vyššia ako vo vdychovanom vzduchu. Na základe tohto faktu sa analýza dychu stáva atraktívnou neinvazívnou diagnostickou metódou.

Koncentrácie prchavých organických zlúčenín VOC (Volatile Organic Compounds) analyzované v tejto práci pochádzajú z pilotnej štúdie vytvorenej

na Lekárskej univerzite v Innsbrucku v rokoch 2006 až 2008 v rámci projektu 6-teho rámcového programu Európskej komisie pod skratkou BAMOD (Breath-Gas Analysis for Molecular-Oriented Detection of Minimal Diseases). Zozbierané vzorky dychu boli analyzované metódou hmotnostnej spektrometrie s protónovou prenosovou reakciou.

Hmotnostná spektrometria s protónovou prenosovou reakciou PTR-MS (Proton-Transfer-Reaction Mass Spectrometry) sa pokladá za ideálny nástroj na analýzu prchavých organických zlúčenín v plynných biologických vzorkách, ako napríklad ľudský dych. Predstavuje mechanizmus schopný detegovať koncentrácie prchavých organických zlúčenín v pomerne krátkom čase s nízkym limitom detekcie (rádovo na úrovni počtu častíc na bilión, ppt) a vysokou senzitivitou merania prchavých organických zlúčenín [1].

V niektorých prípadoch dochádza k tomu, že rôzne zlúčeniny majú tú istú molekulovú hmotnosť. V takomto prípade sú tieto molekuly detegované ako jedna hmotnostná zložka, ozn. m/z (mass-to-charge-ratio). Hmotnostné zložky detegované pomocou PTR-MS sú v rozmedzí od m/z 21 po m/z 230. Hmotnostná zložka je predbežne priradená k tej prchavej organickej zlúčenine, ktorá má najväčšie zastúpenie.

Každá vzorka vydychovaného plynu bola opakovane meraná najmenej 3-krát. Pre niektorých dobrovoľníkov bola odobratá vzorka vydychovaného vzduchu viacej krát. Pred samotnou štatistickou analýzou boli dáta predspracované. Na dosiahnutie nezávislosti medzi meraniami sme ako výslednú hodnotu pre subjekt zobrali medián vypočítaný z mediánov pre opakovane namerané hodnoty koncentrácií m/z jednotlivých vzoriek daného subjektu. Takto spracované dáta v sebe zahŕňajú variabilitu medzi meraniami a označujeme ich ako zašumené dáta.

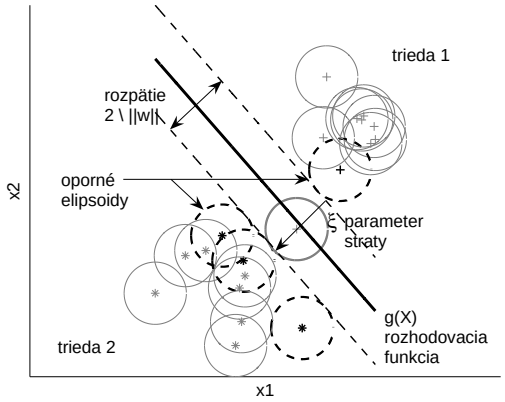
2. **Robustná klasifikačná metóda**

Predpokladajme, že naše namerané dáta $\mathbf{x}_i \in R^N$, $1 \leq i \leq n$, sú zašumené a skutočná hodnota je nejaký bod v špecifikovanom elipsoide, teda predpokladáme elipsoidálny model zašumenia [4]. Nech

$$B(\bar{\mathbf{x}}_i, \mathbf{\Sigma}_i, \gamma_i) = \{\mathbf{x} : (\mathbf{x} - \bar{\mathbf{x}}_i)^{'} \mathbf{\Sigma}_i^{-1} (\mathbf{x} - \bar{\mathbf{x}}_i) \leq \gamma_i^2\}$$

je elipsoid so stredom v bode $\bar{\mathbf{x}}_i$, $\mathbf{\Sigma}_i$ je pozitívna semidefinitná matica, ktorá udáva tvar šumu a parameter $\gamma_i \geq 0$ predstavuje hladinu zašumenia. Pre vstupné dáta nie je nutný predpoklad typu rozdelenia, predpokladáme len konečnosť momentov druhého rádu [4].

Ďalej predpokladajme, že o každom pozorovanom subjekte máme informáciu, do kterej z dvoch tried skutočne patrí. Teda každému subjektu vieme priradiť kategoriálne zatriedenie do tried $y_i \in \{+1, -1\}$, $1 \leq i \leq n$. Kategoriálne zatriedenie y_i platí pre všetky $\mathbf{x} \in B(\bar{\mathbf{x}}_i, \mathbf{\Sigma}_i, \gamma_i)$. Riešením klasifikácie zašumených dát je nájdenie rozhodovacej funkcie $g(\mathbf{x})$ na predikciu y na základe daných elipsoidov B.



Obrázok 1. Schéma riešenia metódy na klasifikáciu zašumených dát do dvoch tried, ktorej riešením je nájdenie parametrov rozhodovacej funkcie na základe dátovej množiny tvoriacej elipsoidy tak, aby rozpätie medzi dvoma paralelnými nadrovinami k hľadanej rozhodovacej funkcii bolo čo najväčšie a v prípade lineárne neseparovateľných dát bola strata čo najmenšia.

Nech náš klasifikátor je nadrovina $\langle \mathbf{w}, \mathbf{x} \rangle + b = 0$, kde úlohou je nájdenie optimálnych parametrov $\mathbf{w}, b$ s pravidlom

$$g(\mathbf{x}) = \operatorname{sign}\left(\langle \mathbf{w}, \mathbf{x} \rangle + b\right).$$

Ak je hodnota rozhodovacej funkcie pozorovaného subjektu kladná, zatriedime subjekt do pozitívnej skupiny $\hat{y}(\mathbf{x}) = +1$. Naopak, ak je hodnota rozhodovacej funkcie pozorovaného subjektu záporná, subjekt zatriedime do negatívnej skupiny $\hat{y}(\mathbf{x}) = -1$.

Optimálne parametre $\mathbf{w}, b$ rozhodovacej funkcie hľadáme tak, aby rozpätie dvoch paralelných nadrovín ku hľadanej nadrovine oddeľujúcej dáta bolo čo najväčšie, teda rovné $2/||\mathbf{w}||$. V prípade lineárne neseparovateľných dát ide o maximalizáciu rozpätia tak, aby bol čo najmenší počet zle klasifikovaných pozorovaní, teda minimálna strata ďalej charakterizovaná voľnými parametrami straty $\xi \geq 0$, obrázok 1.

Elipsoidy, pre ktoré platí

$$y(\langle \mathbf{w}, \mathbf{x} \rangle + b) \geq 1,$$

pre $\forall \mathbf{x} \in B$ a rovnosť platí len v jednom z bodov, teda sa dotýkajú jednej z paralelných nadrovín hrajú rolu tzv. oporných bodov teda ich budeme nazývať oporné elipsoidy (support ellipsoids). Tieto elipsoidy sú postačujúce

pri popise rozhodovacej funkcie $g(\mathbf{x})$, predstavujú len malý zlomok všetkých dát, takže efektívny počet bodov definujúcich rozhodovaciu funkciu $g(\mathbf{x})$ je omnoho menší ako počet subjektov v trénovacej množine.

Riešením klasifikácie zašumených dát je optimalizačná úloha kvadratického programovania

$$\begin{aligned} \min_{\mathbf{w},b,\xi} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \\ \text{s podm.} \quad & y_i(\langle \mathbf{w},\mathbf{x}\rangle + b) \geq 1 - \xi_i \\ & \xi_i \geq 0, \end{aligned}$$

pre $\forall \mathbf{x} \in B(\bar{\mathbf{x}}_i, \mathbf{\Sigma}_i, \gamma_i)$, $1 \leq i \leq n$ a parameter C je regularizačná konštanta, ktorá rieši kompromis medzi maximalizáciou rozpätia a stratou. ξ_i sú voľné parametre straty, predstavujúce vzdialenosť zle klasifikovaného subjektu od prislúchajúcej paralelnej nadroviny, obrázok 1, a v prípade dobre klasifikovaných subjektov $\xi = 0$. Voľné parametre zabezpečujú existenciu riešenia [4]. Regularizačná konštanta C sa volí v rozmedzí $C \in (0, \infty)$. V prípade $C = 0$ sa stráca kontrola nad parametrami straty ξ a riešenie sa nenájde. V prípade $C = \infty$ sa parametre straty nastavia nulové $\xi = 0$ a v prípade, že dáta nie sú lineárne oddeliteľné sa riešenie taktiež nenájde. Rozhodovacie pravidlo teda hľadáme zvyšovaním hodnoty parametra C od dolnej hranice, čím sa zabezpečí nízka strata a nadrovina je definovaná nízkym počtom nenulových prvkov [4]. V prípade vyššie zvolenej hodnoty parametra C síce dochádza k nižšej strate v trénovacej množine, avšak môže dôjsť k pretrénovaniu klasifikačného pravidla na trénovacích dátach a pravdepodobnosť zatriedenia nových subjektov klesá, viac napr. v [5].

Optimalizačná podmienka sa využitím Karush-Kuhn-Tuckerových podmienok dá prepísať na tvar [4]

$$\min_{\mathbf{x}\in B} y_i \langle \mathbf{w}, \mathbf{x}\rangle = y_i \langle \mathbf{w}, \bar{\mathbf{x}}_i\rangle - \gamma_i ||\mathbf{\Sigma}_i^{1/2}\mathbf{w}||.$$

Potom nasledovná robustná formulácia je ekvivalentná s predchádzajúcou optimalizačnou úlohou

$$\begin{aligned} \min_{\mathbf{w},b,\xi} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \\ \text{s podm.} \quad & y_i(\langle \mathbf{w},\bar{\mathbf{x}}_i\rangle + b) \geq 1 - \xi_i + \gamma_i \|\mathbf{\Sigma}_i^{\frac{1}{2}}\mathbf{w}\| \\ & \xi_i \geq 0, \end{aligned}$$

pre $1 \leq i \leq n$, kde robustnou ju robí nelineárny člen $||\mathbf{\Sigma}_i^{\frac{1}{2}}\mathbf{w}||$ nachádzajúci sa v obmedzujúcich podmienkach. Optimalizačná úloha sa rieši ako úloha kónického programovania druhého rádu SOCP (second order cone programing)

$$\begin{aligned} \min_{\mathbf{w},b,\xi} \quad & \sum_{i=1}^{n}\xi_i \\ \text{s podm.} \quad & y_i(\langle \mathbf{w},\bar{\mathbf{x}}_i\rangle + b) \geq 1 - \xi_i + \gamma_i \|\mathbf{\Sigma}_i^{\frac{1}{2}}\mathbf{w}\| \end{aligned}$$

$$||\mathbf{w}|| \leq W$$
$$\xi_i \geq 0,$$

pre $1 \leq i \leq n$, kde člen $||\mathbf{w}||$ je presunutý do podmienky a ohraničený zhora konštantou W, ekvivalentnou s konštantou C. SOCP je založené na základe metódy vnútorného bodu konvexného nelineárneho programovania [4]. Na riešenie využívame programový balík SeDuMi [7].

Stred elipsoidu je ekvivalentný s nameranou hodnotou $\bar{\mathbf{x}}_i \equiv \mathbf{x}_i$. Parametrom γ_i sa znásobuje vplyv šumu, tzv. hladina zašumenia. V prípade $\gamma_i = 0$ pre $\forall i$ sú dáta prezentované bez šumu.

3. Tvar šumu

V robustnej klasifikačnej metóde predpokladáme elipsoidálny model zašumenia, kde matica $\mathbf{\Sigma}_i$ udáva tvar šumu.

Predpokladajme, že máme nezávislé pozorovania. Dôležité je teda odhadnúť iba diagonálne prvky matice $\sigma_{i1}, \ldots, \sigma_{iN}$, kde N je počet charakteristík. Ďalej predpokladajme, že tvar a veľkosť šumu je pre každý pozorovaný subjekt rovnaký. Pre zjednodušenie budeme maticu $\mathbf{\Sigma}_i$ označovať $\mathbf{\Sigma}$.

V prípade, že predpokladáme rovnaký tvar šumu vo všetkých smeroch, tzv. sférický model šumu, potom $\sigma_j = \sigma$. Parameter σ môžeme vypočítať ako

$$\sigma = \sqrt{N} r, \tag{1}$$

kde N je počet pozorovaných charakteristík a r je najmenší napozorovaný rozdiel v charakteristikách danej databázy

$$r = \min_j r_j \qquad r_j = \max_i x_{ij} - \min_i x_{ij},$$

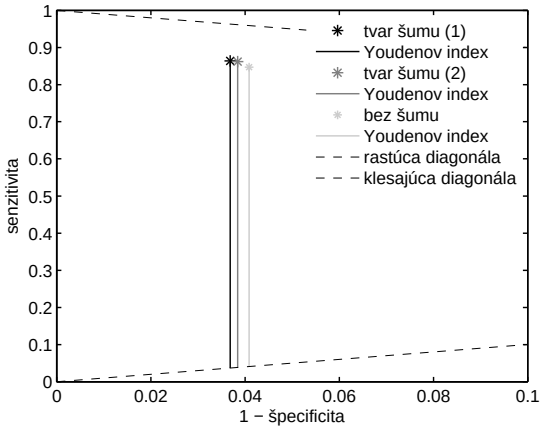
kde $1 \leq j \leq N$, N je počet charakteristík a $1 \leq i \leq n$, n je počet pozorovaných subjektov.

V prípade, že predpokladáme tvar šumu definovaný pre pozorované charakteristiky jednotlivo, potom diagonálne prvky tejto matice σ_j môžeme vypočítať ako výberový rozptyl napozorovaných meraní danej charakteristiky

$$\sigma_j = \frac{1}{n} \sum_{i=1}^{n} (x_{ij})^2 - \left(\frac{1}{n} \sum_{i=1}^{n} x_{ij} \right)^2. \tag{2}$$

4. Analýza dychu

Pilotná štúdia zostavená na Lekárskej Univerzite v Innsbrucku v rokoch 2006 a 2008 obsahuje namerané koncentrácie prchavých organických zložiek pre 54 fajčiarov a 178 nefajčiarov, u ktorých nebola potvrdená diagnóza rakoviny pľúc. Každému subjektu v databáze sa priradila jedna hodnota, ktorá predstavuje medián z mediánov jednotlivých vzoriek vydýchnutého vzduchu, ktorý bol pre presnosť meraný najmenej tri krát. Na základe predchádzajúcich znalostí o metabolizme [6] sme sa zamerali na 12 vybraných VOC ($N = 12$). Medzi vybrané VOC patria molekuly s molekulovou hmotnosťou predbežne



Obrázok 2. Výsledky klasifikácie zašumených dát predstavujúcich koncentrácie 12-tich prchavých organických zlúčenín vydychovaných plynov fajčiarov a nefajčiarov pre rôzne tvary šumu. Tvar šumu (1) prestavuje sférický šum, tvar šumu (2) definovaný pre každú charakteristiku zvlášť.

identifikované ako m/z 28 - kyanovodík, m/z 31 - formaldehyd, m/z 33 - metanol, m/z 42 - acetonitril, m/z 53 - vinylacetylén, m/z 59 - acetón, m/z 61 - kyselina octová, m/z 79 - benzén a m/z 97, m/z 105, m/z 109 a m/z 123.

Klasifikačnú metódu sme aplikovali na naše dáta v simulačnej štúdii na získanie senzitivity a špecificity klasifikácie pozorovaných subjektov do triedy fajčiar vs. nefajčiar. Senzitivita je schopnosť klasifikátora rozpoznať prítomnosť sledovaného znaku, zatiaľ čo špecificita je mierou toho, nakoľko klasifikátor označí tie subjekty, ktoré vlastnosť skutočne nesú [3]. Senzitivita a špecificita sa na základe výsledkov klasifikácie vypočítajú ako

$$Se = \frac{\#\{i, y_i = \hat{y}_i | y_i = +1\}}{\#\{i, y_i = +1\}} \quad \text{a} \quad Sp = \frac{\#\{i, y_i = \hat{y}_i | y_i = -1\}}{\#\{i, y_i = -1\}},$$

$1 \leq i \leq n$, n je počat klasifikovaných subjektov.

Výsledky klasifikácie sme získali zo 100-krát náhodne rozdelenej databázy fajčiarov a nefajčiarov na trénovaciu a testovaciu množinu v pomere 3:2 pre kombinácie parametrov $\gamma = [0 : 0.01 : 0.1]$ a $W = 1$ pri odhadnutom tvare šumu na základe vzorcov (1) a (2). Na obrázku 2 sú znázornené najlepšie výsledky klasifikácie pre rôzne typy tvaru šumu, $\gamma = 0.01$ pre tvar šumu (1) a $\gamma = 0.06$ pre tvar šumu (2) v tzv. ROC grafe. Z ROC grafu vidíme, že klasifikačná metóda má schopnosť klasifikovať subjekty lepšie ako je náhodné chovanie (výsledky sú nad rastúcou diagonálou) a klasifikačné pravidlo

je konzervatívne (výsledky sú pod klesajúcou diagonálou), čo znamená, že pre klasifikačné pravidlo je väčšia chyba zatriedenie negatívneho subjektu do triedy pozitívnych.

Na lepšie porovnanie výsledkov klasifikácie sme odhadli aj Youdenov index. Je mierou efektívnosti zatriedenia subjektov podľa sledovaného znaku. Tento index je ohraničený bodmi 0 a 1, kde hodnota blízka 1 indikuje efektívnu klasifikáciu a hodnota blízka 0 limitovanú efektívnosť. Youdenov index predstavuje vertikálnu vzdialenosť medzi výsledkom klasifikácie v ROC grafe a hlavnou diagonálou. Youdenov index sa vypočíta ako

$$J = Se + Sp - 1.$$

Najefektívnejšia klasifikácie bola dosiahnutá pri tvare šumu podľa vzorca (1).

5. Záver

V práci bola prezentovaná klasifikačná metóda na klasifikáciu zašumených dát. Z výsledkov vyplýva, že pri predpoklade zašumenia dát sa znižuje pravdepodobnosť zlej klasifikácie subjektov pri testovaní. Zlepšenie výsledkov sme dosiahli aj vhodným odhadom tvaru šumu.

Literatúra

[1] Amann, A., Smith, D. (2005) *Breath analysis for clinical diagnosis and therapeutic monitoring*, World Scientific, Singapore.

[2] Bajtarevic, A., et al. (2009) *Noninvasive Detection of Lung Cancer by Analysis of Exhaled Breath*, BMC Cancer, **9**, (348).

[3] Betinec, M. (2006) *Použití ROC křivek pro hodnocení klasifikátorú*, ROBUST'2006, Sborník prací 14. zimní školy JČMF, J. Antoch & G. Dohnal (eds.), Praha, JČMF, 25 – 34.

[4] Bhattacharyya Ch. (2004) *Robust Classification of noisy data using Second Order Cone Programming approach*, In Proceedings International Conference on Intelligent Sensing and Information Processing, 433 – 438.

[5] Cimermanová K. (2008) *Klasifikácia zašumených dát*, ROBUST'2008, Sborník prací 15. zimní školy JČMF, J. Antoch & G. Dohnal (eds.), Praha, JČMF, 41 – 46.

[6] Kushch, I., et al., (2008) *Compounds enhanced in a mass spectrometric profile of smokers' exhaled breath versus non-smokers as determined in a pilot study using PTR-MS*, Journal of Breath Research, **2**, 1 – 26.

[7] Sturm J.F. (1995) *Using SEDUMI 1.02, a Matlab*toolbox for Optimization over symmetric cones*, (Updated for Version 1.05), Optimization Methods and Software, **11**, 625 – 653.

Poďakovanie: Práca bola podporovaná Agentúrou na podporu výskumu a vývoja (APVV): grant SK-AT-0003-08 a Vedeckou grantovou agentúrou Ministerstva školstva SR a Slovenskej akadémie vied (VEGA): grant 1/0077/09 a 2/0019/10.

Adresa: ÚM SAV, Dúbravská cesta 9, 841 04 Bratislava, Slovenská republika

E-mail: katarina.cimermanova@gmail.com

KONZISTENCE NEPARAMETRICKÉHO BAYESOVSKÉHO ODHADU

Michal Friesl

Klíčová slova: Náhodné cenzorování, Koziolův-Greenův model, neparametrické bayesovské odhady, gama proces, konzistence.

Abstrakt: Konzistence bayesovských odhadů nemusí být v případě neparametrických bayesovských odhadů, kdy parametr je nekonečněrozměrný, automaticky zaručena. Připomeneme si, jak je tomu s konzistencí aposteriorního rozdělení a neparametrických bayesovských odhadů funkce spolehlivosti, a podíváme se na konzistenci odhadu v modelu s proporcionálním cenzorováním, prezentovaného na předchozích Robustech.

Abstract: Consistency of nonparametric bayesian estimators is not automatically guaranteed due to infinite number of parameters. In the paper consistency of posterior distribution and of nonparametric bayesian estimators of reliability function is recalled and consistency of an estimator in the proportional censorship model is explored.

1. Úvod

Tento příspěvek je pokračováním posloupnosti mých robustních příspěvků o neparametrickém bayesovském odhadu v Koziolově-Greenově modelu náhodného cenzorování. Pracujeme s modelem náhodného cenzorování, tj. uvažujeme dobu života X — nezápornou náhodnou veličinu s funkcí spolehlivosti

$$S(t) = 1 - F(t) = \mathrm{P}(X > t), \quad t \geqq 0.$$

Pozorování může být zprava cenzorováno časem Y (nezáporná náhodná veličina nezávislá s X), ve skutečnosti pozorujeme první z těchto časů a indikátor, zda jde o pozorování necenzorované,

$$Z = X \wedge Y \quad \text{a} \quad I = I_{[X \leqq Y]}.$$

Cílem je z náhodného výběru dvojic (Z, I) odhadnout funkci spolehlivosti S doby života X.

V tomto obecném modelu náhodného cenzorování uvažujeme navíc dodatečný předpoklad (Koziolův-Greenův, [7]), že rozdělení cenzoru Y souvisí s rozdělením doby života X, a to tak, že pro nějaké $\gamma > 0$ platí

$$S_Y(t) = \big(S(t)\big)^\gamma, \quad t \geqq 0,$$

kde S_Y značí funkci spolehlivosti veličiny Y. Ekvivalentně lze psát

$$\Lambda_Y(t) = \Lambda(t) \cdot \gamma$$

pro odpovídající kumulativní intenzity definované jako

$$\Lambda(t) = -\ln S(t) = \int_0^t \frac{\mathrm{d}F(x)}{S(x)}.$$

Poznámka. *V případě spojitých rozdělení je tato podmínka ekvivalentní nezávislosti veličin I a Z, tj. zda je pozorování cenzorované nezávisí na pozorovaném čase Z. Pravděpodobnost cenzorování je v každém okamžiku stejná a rovna $p = P(I = 1) = 1/(1 + \gamma)$.*

Na minulých Robustech jsem odvodili neparametrický bayesovský odhad funkce spolehlivosti S v tomto modelu, porovnávali ho s jinými odhady funkce spolehlivosti, uvažovali jsme i model se zleva useknutými pozorováními. V tomto příspěvku nejprve odhad připomeneme, poté krátce zmíníme problematiku konzistence neparametrických bayesovských odhadů obecně a nakonec se podíváme na konzistenci našeho odhadu.

2. Odhad

K odhadování funkce spolehlivosti S přistupujeme neparametricky bayesovsky. Tvar funkce S není předem dán, neznámým je nikoli jeden parametr, který by tvar S určil, nýbrž celá $(S(t), t \geqq 0)$. Je třeba zvolit apriorní rozdělení pro proces $S = (S(t), t \geqq 0)$.

Vhodným apriorním rozdělením pro S jsou zprava neutrální procesy, které předpokládají, že pro každé x je $(S(t), t \leqq x)$ nezávislé s $(S(t)/S(x), t > x)$, tj. (relativní) rozložení pravděpodobnosti na intervalu (x, ∞) je nezávislé na tom, jak byla pravděpodobnost rozložena v $(0, x\rangle$ a kolik zbývá na (x, ∞). Ekvivalentním vyjádřením této vlastnosti je, že proces kumulativní intenzity Λ je neklesajícím procesem s nezávislými přírůstky.

Poznámka. *Má-li proces Λ jen náhodnou složku, pak skoro jistě trajektorie procesu S představují funkci spolehlivosti diskrétního rozdělení. V nosiči rozdělení procesu ale mohou být i funkce spolehlivosti všech spojitých rozdělení.*

Naší konkrétní volbou apriorního rozdělení pro parametr Λ je gama proces, tj. apriorně předpokládáme, že zmíněné nezávislé přírůstky procesu Λ mají gama rozdělení,

$$\Lambda(s, t) = \Lambda(t) - \Lambda(s) \sim G(n_0, n_0\Lambda_0(s, t)), \quad 0 \leqq s < t,$$

kde $n_0 > 0$ je společný parametr měřítka, Λ_0 funkce spolehlivosti nějakého pevně zvoleného spojitého rozdělení na $(0, \infty)$ a $\Lambda_0(s, t) = \Lambda_0(t) - \Lambda_0(s)$ značí její přírůstky. Trajektorie tohoto procesu jsou centrovány kolem $\mathrm{E}\,\Lambda(s, t) = \Lambda_0(s, t)$, resp. trajektorie procesu S kolem

$$\mathrm{E}\,S(t) = \mathrm{E}\,\mathrm{e}^{-\Lambda(t)} = (1 + 1/n_0)^{-n_0\Lambda_0(t)} \quad (\approx \mathrm{e}^{-\Lambda_0(t)} \text{ pro velká } n_0),$$

n_0 řídí rozptyl $\operatorname{var}\Lambda(s, t) = \Lambda_0(s, t)/n_0$.

O parametru Koziolova-Greenova modelu γ v apriorním rozdělení předpokládáme, že je nezávislý s procesem Λ a že jeho rozdělení se řídí hustotou $\pi(\gamma)$ vzhledem k nějaké míře μ na $(0, \infty)$.

K popisu aposteriorního rozdělení a odhadů použijeme následující charakteristiky pozorovaného výběru. Pozorované časy $Z_1, \ldots, Z_n$ uspořádáme, ponecháme každou hodnotu jen jedenkrát, po dodefinování $T_0 = 0$, $T_{N+1} = \infty$

dostaneme navzájem různé časy

$$0 = T_0 < T_1 < \cdots < T_N < T_{N+1} = \infty.$$

Pro každý čas T_j dále označíme

$$\begin{aligned} U_j &= \#\text{ necenzorovaných pozorování se } Z_k = T_j, \\ C_j &= \#\text{ cenzorovaných pozorování se } Z_k = T_j, \\ N_j &= \#\text{ pozorování s } Z_k > T_j = n - \sum_{i \leqq j} (U_i + C_i). \end{aligned}$$

Poznámka. *Pro časy pocházející ze spojitého rozdělení je $U_j + C_j = 1$. Dle apriorního modelu se ale (i při spojité Λ_0) časy řídí rozdělením diskrétním a je $U_j + C_j > 1$ s kladnou pravděpodobností.*

Pro aposteriorní rozdělení parametrů Λ a γ pak platí:

- Při daném γ je Λ opět procesem s nezávislými přírůstky, a to:
  Na intervalech mezi pozorovanými časy dochází ke změně měřítka,

$$(\Lambda(s,t) \mid \text{data}, \gamma) \sim \mathrm{G}(M_{j-1}(\gamma), n_0 \Lambda_0(s,t)), \quad (s,t\rangle \subset (T_{j-1}, T_j),$$

  kde $M_j(\gamma) = M_j^0(\gamma)$, $M_j^m(\gamma) = n_0 + N_j(1+\gamma) + m$, $m = 0, 1$, $j = 0, \ldots, N$.
  V okamžicích ukončení pozorování je vždy skok, a to s hustotou (při daném γ)

$$f_{(\Delta\Lambda(T_j)|\text{data},\gamma)}(x) = x^{-1} \mathrm{e}^{-(M_j(\gamma)+C_j)x} (1-\mathrm{e}^{-x})^{U_j} (1-\mathrm{e}^{-\gamma x})^{C_j} / c_j(\gamma), \quad x > 0,$$

  kde $c_j(\gamma) = c_j^0(\gamma)$ je normovací konstanta (viz níže).
- Aposteriorní rozdělení γ má hustotu

$$\pi(\gamma \mid \text{data}) \propto \Big(\prod_{j=1}^{N} q_j(\gamma)\Big) \pi(\gamma),$$

  kde $q_j = q_j^0$,

$$q_j^m(\gamma) = \big(M_{j-1}^m(\gamma)\big)^{-n_0 \Lambda_0(T_{j-1}, T_j)} c_j^m(\gamma), \quad j = 1, \ldots, N,$$

$$c_j^m(\gamma) = \sum_{k=1}^{U_j} \sum_{\ell=1}^{C_j} (-1)^{k+\ell} \binom{U_j}{k} \binom{C_j}{\ell} \ln \frac{M_j^m(\gamma) + C_j}{M_j^m(\gamma) + C_j + k + \ell\gamma}.$$

Odhad funkce spolehlivosti S počítáme jako aposteriorní střední hodnotu. Při jejím výpočtu nejdříve podmíníme hodnotou γ, využijeme nezávislosti přírůstků, takže dostáváme součin přes intervaly a skoky

$$\mathrm{E}(S(t) \mid \text{data}, \gamma) = \mathrm{E}(\mathrm{e}^{-\Lambda(t)} \mid \text{data}, \gamma) = \prod \mathrm{E}(\mathrm{e}^{-\Lambda(\cdot,\cdot)} \mid \text{data}, \gamma).$$

Po zprůměrování přes $\pi(\gamma \mid \text{data})$ pak pro $t \in \langle T_{i-1}, T_i)$ dostáváme odhad

$$\widehat{S}(t) = \frac{\displaystyle\int \Big(\prod_{j<i} q_j^1(\gamma)\Big)\Big(\prod_{j \geq i} q_j(\gamma)\Big)\Big(\frac{M_{i-1}(\gamma)}{M_{i-1}^1(\gamma)}\Big)^{n_0 \Lambda_0(T_{i-1}, s)} \pi(\gamma)\, \mathrm{d}\mu(\gamma)}{\displaystyle\int \Big(\prod_j q_j(\gamma)\Big) \pi(\gamma)\, \mathrm{d}\mu(\gamma)}.$$

V [5] jsem se dotkli kvality tohoto odhadu z bayesovského pohledu prostřednictvím bayesovského rizika $\mathrm{BR}\widehat{S}(t) = \mathrm{E}\big(\widehat{S}(t) - S(t)\big)^2 \to 0$, tedy měřeno apriorním rozdělením. Budeme-li navíc chtít odhad použít i pro data pocházející ze spojitých rozdělení, kterým naše apriorní rozdělení celkově přisuzuje nulovou pravděpodobnost, může nás také zajímat, zda pro konkrétní rozdělení S_* bude odhad fungovat dobře a bude při rostoucím počtu pozorování platit např. $\widehat{S}(t) \to S_*(t)$ (s.j.).

3. **Konzistence obecně**

V tomto odstavci učiníme odbočku od našeho modelu a cenzorování. Nechť pozorování jsou dána jako náhodný výběr $X_1, \ldots, X_n$ z rozdělení s distribuční funkcí $S = S_\theta$, závisející na parametru $\theta \in \Theta$, který neznáme. Předpokládejme pro něj apriorní rozdělení $\pi(\theta)$, příslušné aposteriorní rozdělení (při rozsahu výběru n) označme $\pi_n = \pi(\theta \,|\, X_1, \ldots, X_n)$, resp. bayesovské odhady, které z něj vycházejí, jako $\widehat{\theta}_n = \mathrm{E}(\theta \mid X_1, \ldots, X_n)$, $\widehat{S}_n = \mathrm{E}(S_\theta \mid X_1, \ldots, X_n)$.

Konzistencí odhadu $\widehat{S}_n$, resp. aposteriorního rozdělení π_n, se rozumí, že pokud ve skutečnosti pozorování pocházejí z rozdělení s parametrem $\theta = \theta_*$, bude při rostoucím rozsahu výběru $n \to \infty$ platit $\widehat{S}_n \to S_*$ s.j. (kde $S_* = S_{\theta_*}$), případně $\pi(S \mid \text{data}) \to \delta_{S_*}$ s.j., kde δ_x označuje rozdělení s $\delta_x(\{x\}) = 1$, soustředěné do hodnoty x. Zde s.j. se myslí vzhledem ke skutečnému rozdělení posloupnosti $X_1, X_2, \ldots$.

Na otázku, pro které hodnoty parametru θ_* konzistence nastává, sice existuje obecně platná odpověď vycházející z [2], že pro π-s.v. θ_* (tedy pro skoro všechna vzhledem k apriornímu rozdělení), prakticky ale nevíme, zda parametr θ_*, kterým se v konkrétním případě data řídí, nespadá zrovna do množiny výjimek míry 0, pro které konzistence nenastává. Rádi bychom měli konzistenci zaručenu pro všechna $\theta_* \in \Theta$, resp. všechna z nosiče apriorního rozdělení.

Při našem neparametrickém odhadování rozdělení je parametrem θ funkce spolehlivosti S, kterou můžeme popsat také odpovídající intenzitou poruch Λ, nebo odpovídající pravděpodobnostní mírou P, a oborem Θ jeho hodnot množina představující všechna rozdělení.

Situace je podrobně rozebrána v [3]. Je-li parametr konečněrozměrný (tj. oborem hodnot veličin X_i je daná konečná množina, takže funkce Λ, S, či míra P jsou popsány konečně mnoha parametry-pravděpodobnostmi $p_1, \ldots, p_k$), nastává konzistence pro všechny hodnoty $P = P_* = P_{\theta_*}$ z nosiče apriorního rozdělení.

V nekonečněrozměrném případě však nastávají problémy už při odhadu diskrétních rozdělení, kdy neznámé rozdělení P je popsáno spočetně mnoha pravděpodobnostmi $p_1, p_2, \ldots$. V této situaci pro libovolné rozdělení $\widetilde{P} \neq P_*$ existuje apriorní rozdělení π takové, že jeho nosič obsahuje P_*, ale přitom aposteriorní rozdělení míry P skoro jistě konverguje k $\delta_{\widetilde{P}}$, a nikoli k δ_{P_*}. Např. tedy při "vhodném" apriorním rozdělení pro neznámé P (jeho konkrétní tvar

je v [3] zkonstruován) tak pozorování generovaná z $P_* = \mathrm{Geom}(1/4)$ vedou k odhadu $\widehat{P}_n$ konvergujícímu k $\widetilde{P} = \mathrm{Geom}(3/4)$. Navíc dvojic (θ_*, π), pro které je odhad $\widehat{\theta}_n$ konzistentní, je málo, měřeno topologicky, tvoří množinu 1. kategorie [4]. Volbu apriorního rozdělení je tedy nutno dobře uvážit.

Při praktickém použití požadujeme, aby konzistence nastávala sice pro všechna θ_*, ale stačí pro konkrétní zvolené apriorní rozdělení. Naštěstí pro oblíbené apriorní rozdělení Dirichletovo, či obecněji tail-free, tomu tak je, stejně tak v případě konečné směsi rozdělení, u kterých jednotlivě konzistence platila. V případě apriorních rozdělení, které vzniknou jako nekonečné směsi rozdělení s vlastností konzistence, tomu už tak být nemusí.

Podobně ve spojitém případě při užití základních apriorních rozdělení (Dirichletův proces, beta, gama proces) konzistence nastává. Už ale v třídě procesů neutrálních zprava obecně nikoli, jak ukázali [6] na příkladu zobecněného beta procesu, u kterého bayesovský odhad konverguje k mocnině skutečné funkce spolehlivosti, $\widehat{S}_n \to S_*^\alpha$, kde α je parametr procesu. Konzistence tedy nastává pouze pro $\alpha = 1$, kdy jde přímo o beta proces.

Pro model s cenzorováním pak bylo v [1] ukázáno, že v případě zprava neutrálních procesů konzistence nastává právě tehdy, když v modelu bez cenzorování. Tento výsledek se týká pouze neinformativního cenzorování, tedy nikoli Koziolova-Greenova modelu.

4. **Konzistence odhadu v Koziolově-Greenově modelu**

Předpokládejme, že skutečné parametry modelu jsou γ_* a S_*, přičemž S_* je funkcí spolehlivosti spojitého rozdělení. Pak pozorování nastávají v navzájem různých časech a při daném γ lze psát $\mathrm{E}(S(t) \mid \mathrm{data}, \gamma) = (1) \cdot (2)$, kde první člen vyjadřuje poklesy přes intervaly mezi pozorováními a je

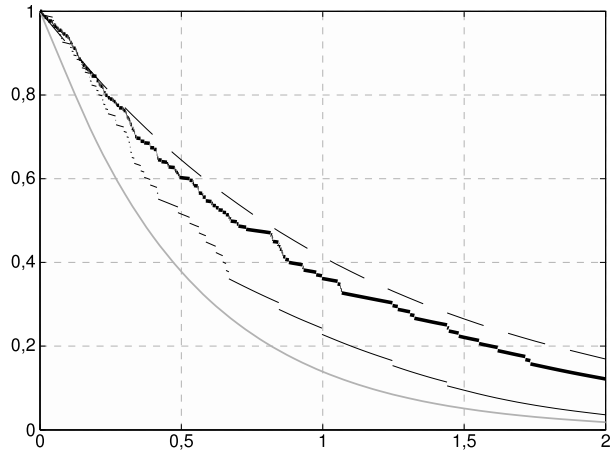
$$(1) = \prod_{j;T_j<t} \Big(1 + \frac{1}{n_{j-1}}\Big)^{-n_0\Lambda_0(T_{j-1},T_j)} \cdot \Big(1 + \frac{1}{n_{i-1}}\Big)^{-n_0\Lambda_0(T_{i-1},t)} \to 1$$

při $n \to \infty$, zatímco druhý, rozhodující, člen vyjadřuje poklesy prostřednictvím skoků v časech pozorování a můžeme ho po chvíli počítání aproximovat

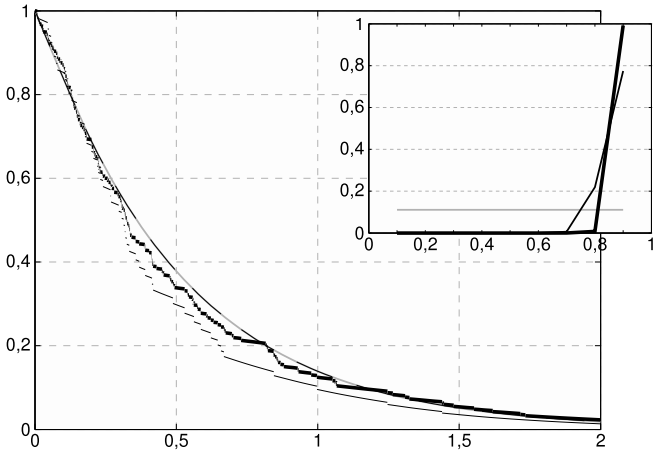
$$(2) = \prod_{j;T_j<t} \frac{\ln\Big(1 + \frac{U_j+\gamma C_j}{n_j+C_j+1}\Big)}{\ln\Big(1 + \frac{U_j+\gamma C_j}{n_j+C_j}\Big)} \approx \prod_{j<i}\Big(1 - \frac{1}{(1+\gamma)N_j}\Big) \to \big(S_*(t)\big)^{\frac{1+\gamma_*}{1+\gamma}},$$

kde $n_j = M_{j-1}(\gamma) = n_0 + N_{j-1}(1+\gamma)$.

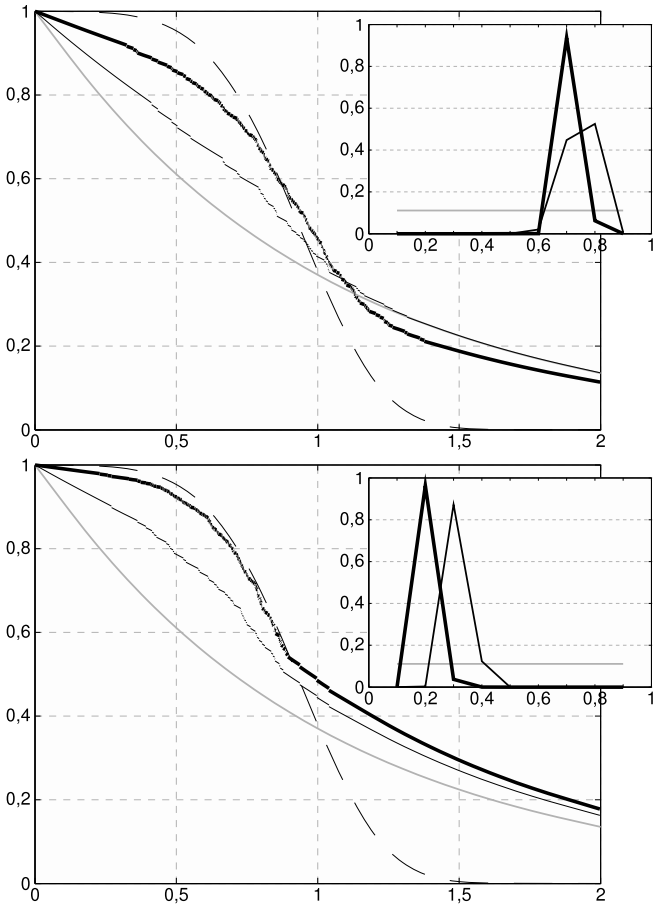
Tuto vlastnost ilustruje obrázek 1 ukazující na základě simulovaných pozorování, jak by dopadl odhad, pokud bychom předpokládali $n_0 = 10$, $\Lambda_0(t) = t$ (odpovídající exponenciálnímu rozdělení) a neuváženě podíl necenzorovaných pozorování $p = 0{,}4$, tj. parametr $\gamma = 1/p - 1 = 1/0{,}4 - 1 = 1{,}5$ (s.j.), zatímco ve skutečnosti data by pocházela sice z $\Lambda_*(t) = \Lambda_0(t)$, ale s podílem necenzorovaných pozorování $p_* = 0{,}9$, tj. parametr modelu by byl $\gamma_* = 1/9$. S rostoucím počtem pozorování se odhad funkce spolehlivosti blíží k hodnotě $S_*^{4/9}(t)$.



Obrázek 1. Odhad při chybném předpokladu o γ. Zdola: šedě apriorní $\exp(-\Lambda_0)$, slabě odhad z $n = 50$ a tlustě z $n = 200$ pozorování, čárkovaně limitní hodnota (v tomto případě chybná).



Obrázek 2. Odhad při "neurčité" informaci o γ. Slabě odhad z $n = 50$ a tlustě z $n = 200$ pozorování, čárkovaně skutečná hodnota $\exp(-\Lambda_*)$. Vpravo graf rozdělení p (šedě apriorní, tence aposteriorní z $n = 50$ a tlustě z $n = 200$ pozorování).



Obrázek 3. Data z Weibullova rozdělení. Křivky zdola: šedě apriorní $\exp(-\Lambda_0)$, slabě odhad z $n = 50$ a tlustě z $n = 200$ pozorování, čárkovaně skutečná hodnota $\exp(-\Lambda_*)$.

Záleží tedy také na správnosti odhadu parametru γ. Po určitých úvahách můžeme nahlédnout, že pro $\gamma \neq \gamma_*$

$$\frac{\pi(\gamma \mid \text{data})}{\pi(\gamma_* \mid \text{data})} \leqq konst\left(\frac{(1+x)^c}{1+cx}\right)^n \frac{\pi(\gamma)}{\pi(\gamma_*)} \to 0,$$

$$\text{kde} \quad x = \frac{\gamma - \gamma_*}{\gamma_*} \quad \text{a} \quad c = \frac{\gamma_*}{1+\gamma_*} = 1 - p_*,$$

a tak aposteriorní hustota parametru γ se soustřeďuje kolem γ_* — pokud tuto hodnotu v apriorním rozdělení připustíme.

Na obr. 2 je znázorněn odhad vycházející z "neurčitého" rovnoměrného apriorního rozdělení γ na 9 hodnotách $\gamma = 1/p - 1$ odpovídajících pravděpodobnostem necenzorovaného pozorování $p = 0{,}1; \ldots; 0{,}9$. V menším obrázku je připojen graf aposteriorních pravděpodobností jednotlivých hodnot p.

Jako další příklad nechť apriorní střední intenzita je jako dosud $\Lambda_0(t) = t$, ale skutečné rozdělení dat nechť se řídí Weibullovým rozdělením $S_*(t) = \exp(-t^5)$, $t > 0$. Navíc zvětšíme vliv apriorní informace na odhad (v porovnání s rozsahy výběru $n = 50$ a $n = 200$) volbou $n_0 = 100$. Na obrázku 3 nahoře je znázorněn odhad, když parametr modelu γ_* byl $1/4$ (odpovídá podílu necenzorování $p_* = 0{,}8$). Spodní obrázek pak zobrazuje situaci, kdy (simulovaná) data pocházejí z modelu s parametrem $\gamma_* = 4$ odpovídajícím extrémně malému podílu necenzorovaných pozorování $p_* = 0{,}2$. V tomto případě parametry vedou k pozorování většího množství menších časů, takže odhad je v levé části blíže skutečné S_*, zatímco v pravé části, kde data chybí, jeho tvar kopíruje tvar $\exp(-\Lambda_0)$. K lepšímu přiblížení by došlo při větším počtu pozorování (nebo při menším n_0). Nepřesnost, že rozdělení parametru p je v horním případě při $n = 200$ soustředěno k hodnotě 0,7, je způsobena menším podílem necenzorovaných pozorování v nagenerovaném výběru oproti nominálním 80 % (a naší hrubou diskrétní volbou možných hodnot p).

Literatura

[1] Dey J., Erickson R.V. and Ramamoorthi R. V. (2003) *Some aspects of neutral to right priors.* Internat. Statist. Rev. **71** (2), 383 – 401.
[2] Doob J.L. (1949) *Application of the theory of martingales.*Coll.Int. du CNRS **13**, 23-27.
[3] Freedman D. (1963) *On the asymptotic behavior of Bayes' estimates in the discrete case.* Ann. Math. Statist. **34** (4), 1386 – 1403.
[4] Freedman D. (1965) *On the asymptotic behavior of Bayes estimates in the discrete case II.* Ann. Math. Statist. **36** (2), 454 – 456.
[5] Friesl M. (2006) *Porovnání neparametrických bayesovských odhadů při cenzorování.* In ROBUST 2006 (Antoch J. a Dohnal G., eds.), JČMF, Praha, pp. 83 – 90.
[6] Kim Y. and Lee J. (2001) *On posterior consistency of survival models.* Ann. Statist. **29** (3), 666 – 686.
[7] Koziol J.A. and Green S.B. (1976) *A Cramér-von Mises statistic for randomly censored data.* Biometrika **63** (3), 465 – 474.

Poděkování: Tato práce byla podporována grantem MSM 4977751301.

Adresa: FAV ZČU, KMA, Univerzitní 22, 306 14 Plzeň
E-mail: `friesl@kma.zcu.cz`

POWER TESSELLATION AS A TOOL FOR ESTIMATING PARAMETERS IN A MODEL OF A RANDOM SET

Kateřina Helisová

Keywords: Boolean model, Gibbs process, interaction process, MCMC maximum likelihood, power tessellation, Quermass-interaction process.

Abstract: Consider a random set $\mathbf{X}$ given by a union of interacting discs with centers randomly scattered in $S \subset \mathbf{R}^2$ and with arbitrary radii. Assume that its probability measure is given by a density with respect to the probability measure of Boolean model, i.e. with respect to a process of discs without any interactions. Next, assume that the density is of the form

$$f_\theta(\mathbf{x}) = \frac{1}{c_\theta} \exp(\theta \cdot T(\mathcal{U}_\mathbf{x}))$$

for any configuration $\mathbf{x} = (x_1, \ldots, x_n)$ of discs $x_1, \ldots, x_n$, where θ is a vector of parameters, $T(\mathcal{U}_\mathbf{x})$ is a vector of geometrical characteristics of the union $\mathcal{U}_\mathbf{x}$ consisting of the discs $\mathbf{x}$ and c_θ denotes a normalizing constant.

In this contribution, we briefly introduce two methods of estimating the parameters θ from data - MCMC maximum likelihood and method based on integral characterization of Gibbs process - and show how can usage of the power tessellation described in this paper make the calculations in both these methods much faster.

Abstrakt: Uvažujme náhodnou množinu $\mathbf{X}$ danou sjednocením kruhů se středy náhodně rozmístěnými v $S \subset \mathbf{R}^2$, libovolnými poloměry a možnými vzájemnými interakcemi. Předpokládejme, že pravděpodobnostní míra této náhodné množiny je daná hustotou vzhledem k pravděpodobnostní míře nějakého Booleovského modelu, tj. náhodného procesu kruhů bez jakýchkoliv interakcí, a že je tato hustota ve tvaru

$$f_\theta(\mathbf{x}) = \frac{1}{c_\theta} \exp(\theta \cdot T(\mathcal{U}_\mathbf{x}))$$

pro libovolnou konfiguraci $\mathbf{x} = (x_1, \ldots, x_n)$ kruhů $x_1, \ldots, x_n$, přičemž θ je vektor parametrů, $T(\mathcal{U}_\mathbf{x})$ je vektor geometrických charakteristik sjednocení $\mathcal{U}_\mathbf{x}$ kruhů z konfigurace $\mathbf{x}$ a c_θ značí normalizační konstantu.

V příspěvku jsou stručně představeny dvě metody odhadu parametru θ z dat, a to metoda maximální věrohodnosti s využitím MCMC simulací a metoda založená na integrální charakterizaci Gibbsova procesu. Obě tyto metody jsou výpočetně náročné, avšak my zde ukážeme, jak využití silové mozaiky popsané v tomto článku činí tyto výpočty znatelně rychlejšími.

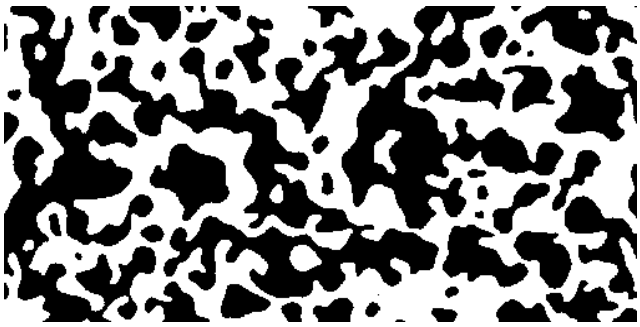


FIGURE 1. Heather dataset first presented by Peter Diggle in 1981. The image shows the presence of heather (indicated by black) in a 10×20 m region at Jädraås in Sweeden.

1. Motivation

In the last years, describing and modeling of random geometrical objects have become very popular. An example of such an object - a random set - which was analyzed by many statisticians is a digital image of a heather grow shown in Figure 1. At the turn of the centuries, many theoretical results were derived (see e.g. [8] or [10]). Some of them concerns also simulation methods and methods for estimating parameters in models of point processes, random tessellations, random sets etc., but possibilities for their applications were very bounded because of their computational complexity. Now, when the possibilities for high-volume computations are wider, many mathematicians working in the field of stochastic geometry aim to that applications. However, in spite of the improved computer speed, they must still keep track on simplifying the algorithms.

One of such simplifications is the power tessellation which is a very useful tool for simulation and consequently for estimation of parameters in a special model for random set introduced e.g. in [6]. This tessellation and its usefulness are described in this paper.

2. Basic definitions and settings

Definition *Consider N the system of locally finite subsets of $\mathbb{R}^d$ with the σ-algebra $\mathcal{N} = \sigma(\{\mathbf{x} \in N : \sharp(\mathbf{x} \cap A) = m\} : A \in \mathcal{B}, m \in \mathbf{N}_0)$, where $\mathcal{B}$ denotes bounded Borel sets and $\mathbf{N}_0$ are the natural numbers including zero. A point process X defined on $\mathbb{R}^d$ is a measurable mapping from some probability space $(\Omega, \mathcal{F}, P)$ to $(N, \mathcal{N})$.*

Definition *A locally finite, diffusion measure μ on $\mathcal{B}$ satisfying $\mu(A) = \mathrm{E}X(A)$ for all $A \in \mathcal{B}$ is called intensity measure. If there exists a function $\rho(x)$ for $x \in \mathbb{R}^d$ such that $\mu(A) = \int_A \rho(x)dx$, then $\rho(x)$ is called intensity*

function. If $\rho(x) = \rho$ is constant then the constant ρ is called intensity and the point process is called stationary process.

Definition *Poisson point process Y is the point process which satisfies:*

- *for any finite collection $\{A_n\}$ of disjoint sets in $\mathbb{R}^d$, the numbers of points in these sets, $Y(A_n)$, are independent random variables,*
- *for each $A \subset \mathbb{R}^d$ such that $\mu(A) < \infty$, $Y(A)$ has Poisson distribution with parameter $\mu(A)$, i.e. $P[Y(A) = k] = \frac{\mu(A)^k}{k!} e^{-k}$, where μ is the intensity measure.*

Definition *Let Y be the Poisson point process with an intensity measure μ and for $F \in \mathcal{N}$, denote $\Pi(F) = P(Y \in F)$. We say that a point process X is given by a density f with respect to the Poisson process Y if*

$$P(X \in F) = \int_F f(\mathbf{x}) \Pi(d\mathbf{x}).$$

3. Model construction

Denote $b = b(u, r)$ a disc with center in $u \in \mathbb{R}^2$ and radius $r \in (0, \infty)$. When we identify b with the point $x = (u, r)$ in $\mathbb{R}^2 \times (0, \infty)$, then the process of discs $\cup b_i = \cup b(u_i, r_i)$ can be identity with a point process in $\mathbb{R}^2 \times (0, \infty)$.

Consider a Poisson point process Y. The corresponding disc process $\mathbf{Y}$ (called Boolean model) plays the role of the reference process. In general, its intensity measure is $\rho(u)\,\mathrm{d}u\,Q(\mathrm{d}r)$ on $\mathbb{R}^2 \times (0, \infty)$, where $\rho(u)$ corresponds to the intensity function of the centers and Q to the probability measure on the radii of the discs.

Our model is then a process of discs $\mathbf{X}$ such that the corresponding point process X is absolutely continuous with respect to the reference Poisson process Y, and it is given by a density $f(\mathbf{x})$ for any finite configuration $\mathbf{x} = \{x_1, \ldots, x_n\}$.

In this paper, we assume for simplicity that X is a finite point process defined on $S \times (0, R)$, where S denotes a given bounded planar region such that $\int_S \rho(u)\,\mathrm{d}u > 0$ and $R < \infty$. Results for unbounded disc radii can be found in [6].

The model density is considered in exponential form

$$f_\theta(\mathbf{x}) = \exp\left(\theta \cdot T(\mathcal{U}_\mathbf{x})\right) / c_\theta, \tag{1}$$

where $T(\mathcal{U}_\mathbf{x})$ is a vector of geometrical characteristics of the union $\mathcal{U}_\mathbf{x}$ consisting of the discs $\mathbf{x}$, θ is a vector of parameters and c_θ denotes a normalizing constant.

In practice, data are usually in the form where we can see the whole union but the structure of the individual discs is unobservable. Therefore it is suitable to choose the geometrical characteristics in (1) so that they can be obtained when having only the picture of the union. Such characteristics are for example $T = (A, L, N_{\mathrm{cc}}, N_{\mathrm{h}}, N_{\mathrm{id}}, N_{\mathrm{bv}})$, where

- $A = A(\mathcal{U}_\mathbf{x})$ is the area,

- $L = L(\mathcal{U}_{\mathbf{x}})$ is the perimeter,
- $N_{\mathrm{cc}} = N_{\mathrm{cc}}(\mathcal{U}_{\mathbf{x}})$ is the number of connected components,
- $N_{\mathrm{h}} = N_{\mathrm{h}}(\mathcal{U}_{\mathbf{x}})$ is the number of holes (i.e. the empty places inside components),
- $N_{\mathrm{id}} = N_{\mathrm{id}}(\mathcal{U}_{\mathbf{x}})$ is the number of isolated discs (i.e. the discs which themselves create connected components of the union),
- $N_{\mathrm{bv}} = N_{\mathrm{bv}}(\mathcal{U}_{\mathbf{x}})$ is the number of boundary vertices (i.e. the points on the boundary of the union which are the intersections of two discs).

Considering these characteristics, the density is of the form

$$f_\theta(\mathbf{x}) = \frac{1}{c_\theta} \exp\left(\theta_1 A(\mathcal{U}_{\mathbf{x}}) + \theta_2 L(\mathcal{U}_{\mathbf{x}}) + \theta_3 N_{\mathrm{cc}}(\mathcal{U}_{\mathbf{x}}) + \theta_4 N_{\mathrm{h}}(\mathcal{U}_{\mathbf{x}}) + \theta_5 N_{\mathrm{id}}(\mathcal{U}_{\mathbf{x}}) + \theta_6 N_{\mathrm{bv}}(\mathcal{U}_{\mathbf{x}})\right). \tag{2}$$

In [6], theoretical results for the density of the form (2) are derived. However later in [7], the authors mention that it is often difficult to observe isolated discs and boundary vertices from data, and for statistical analysis of the image in Figure 1, they reduce the density so that they consider only the first four statistics, i.e. they set $\theta_5 = \theta_6 = 0$.

In some papers (e.g. [5], [1] or [2]), authors concerns a special form of the density (2) where $\theta_5 = \theta_6 = 0$ and $\theta_3 = -\theta_4$, i.e. they work with

$$\begin{aligned} f_\theta(\mathbf{x}) &= \frac{1}{c_\theta} \exp\left(\theta_1 A(\mathcal{U}_{\mathbf{x}}) + \theta_2 L(\mathcal{U}_{\mathbf{x}}) + \theta_3 (N_{\mathrm{cc}}(\mathcal{U}_{\mathbf{x}}) - N_{\mathrm{h}}(\mathcal{U}_{\mathbf{x}}))\right) \\ &= \frac{1}{c_\theta} \exp\left(\theta_1 A(\mathcal{U}_{\mathbf{x}}) + \theta_2 L(\mathcal{U}_{\mathbf{x}}) + \theta_3 \chi(\mathcal{U}_{\mathbf{x}})\right), \end{aligned} \tag{3}$$

where $\chi(\mathcal{U}_{\mathbf{x}}) = N_{\mathrm{cc}}(\mathcal{U}_{\mathbf{x}}) - N_{\mathrm{h}}(\mathcal{U}_{\mathbf{x}})$ is called Euler-Poincaré characteristic of the set $\mathcal{U}_{\mathbf{x}}$ and the model given by the density (3) is called Quermass-interaction process.

4. Simulations

Simulation of the model defined in the previous section is very important for statistical analyses introduced later in Section 6. Before describing the simulating algorithm, define one important term.

Definition *For finite configuration* $\mathbf{x} \subset S \times (0, \infty)$ *and* $y \in S \times (0, \infty) \setminus \mathbf{x}$, *Papangelou conditional intensity is defined as*

$$\lambda_\theta(\mathbf{x}, y) = f_\theta(\mathbf{x} \cup \{y\}) / f_\theta(\mathbf{x}).$$

Using other words, Papangelou conditional intensity says "how much better" is the configuration $\mathbf{x} \cup \{y\}$ than the configuration $\mathbf{x}$. Denoting

$$\begin{aligned} A(\mathbf{x}, y) &= A(\mathbf{x} \cup y) - A(\mathbf{x}), \\ L(\mathbf{x}, y) &= L(\mathbf{x} \cup y) - L(\mathbf{x}), \\ &\vdots \end{aligned}$$

the increments (or decreases) of the considered characteristics, we get

$$\lambda_\theta(\mathbf{x}, y) = \exp\left(\theta_1 A(\mathbf{x}, y) + \theta_2 L(\mathbf{x}, y) + \ldots + \theta_6 N_{bv}(\mathbf{x}, y)\right).$$

To simulate the model (2) we use MCMC methods, especially a simple version of Metropolis-Hastings algorithm (for more details, see e.g. [8]) which runs in the following steps:

(1) Suppose that in time t, we have a configuration $\mathbf{x_t} = \{x_1, \ldots, x_n\}$.
(2) In time $t+1$
  (a) with probability $1/2$, the proposal is $\mathbf{x_t} \cup \{x_{n+1}\}$ and
    (i) we accept it with probability $\min\{1; H(\mathbf{x_t}, x_{n+1})\}$ and set $\mathbf{x_{t+1}} = \mathbf{x_t} \cup \{x_{n+1}\}$,
    (ii) else we refuse it and set $\mathbf{x_{t+1}} = \mathbf{x_t}$,
  (b) else, the proposal is $\mathbf{x_t} \backslash \{x_i\}$ and
    (i) we accept it with probability $\min\{1; 1/H(\mathbf{x_t} \backslash \{x_i\}, x_i)\}$ and set $\mathbf{x_{t+1}} = \mathbf{x_t} \backslash \{x_i\}$,
    (ii) else we refuse it and set $\mathbf{x_{t+1}} = \mathbf{x_t}$,

where the Hastings ratios H are given by

$$H(\mathbf{x_t}, x_{n+1}) = \lambda_\theta(\mathbf{x_t}, x_{n+1}) \frac{|S|}{\rho(x_{n+1}) \cdot (n+1)} \quad \text{and}$$

$$H(\mathbf{x_t} \backslash \{x_i\}, x_i) = \lambda_\theta(\mathbf{x_t} \backslash \{x_i\}, x_i) \frac{|S|}{\rho(x_i) \cdot n}, \quad \text{respectively.}$$

It means that in each iteration, we have to calculate the Papangelou conditional intensity λ_θ, i.e. for each geometrical characteristic, we need to calculate the difference between its value with and without the added or deleted disc. For example for the area or for the perimeter, this calculations are commonly done through the inclusion-exclusion formula which is very complex. Moreover, we usually need many thousands of the iterations. Thus such calculations would be very time consuming unless doing any acceleration. Such an improving is described in the following sections.

5. Power tessellation of a union of discs

Definition *For a disc $b(u, r)$, define the ghost sphere as $s(u, r) = \{a \in \mathbb{R}^3 : \|a - u\| = r\}$, i.e. as the hypersphere in $\mathbb{R}^3$ with center u and radius r. A configuration of discs is said to be in general position if the intersection of any $k+1$ corresponding ghost spheres is either empty or a sphere of dimension $2 - k$, where $k = 1, 2, \ldots$*

The definition of the general position says that any intersection of two boundary circles is not exactly one point but it is either empty or it consists of two points. Moreover, when we joint these two points by a line and do the same for all intersecting discs, then at most three such lines meet in one point.

In [6], there is proved that for any Poisson process, the discs are in general position almost surely. Since we assume that the process $\mathbf{X}$ is absolutely

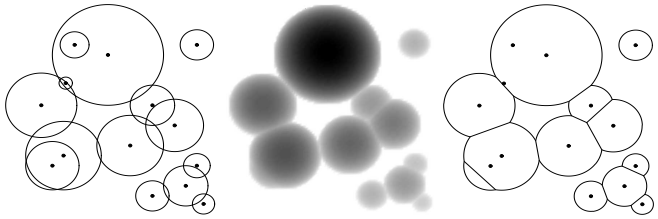


FIGURE 2. Left: A configuration of discs in general position. Middle: The upper hemispheres as seen from above. Right: The power tessellation of the union of discs.

continuous with respect to the reference process, the discs in the model lie in general position almost surely, too.

Assume a union of discs $\mathcal{U} = \cup b_i$ in the general position. For each disc b_i with ghost sphere s_i, let $s_i^+ = \{(a_1, a_2, a_3) \in s_i : a_3 \geq 0\}$ denote the corresponding upper hypersphere. For $v \in b_i$, let $c_i(v)$ denote the unique point on s_i^+ those orthogonal projection on $\mathbb{R}^2$ is v. Define $C_i = \{c_i(v) : v \in b_i, \|v - c_i(v)\| \geq \|v - c_j(v)\|$ for $v \in b_j\ \forall j\}$. Denote B_i the orthogonal projection of C_i on $\mathbb{R}^2$.

Definition *The system $\mathcal{B}$ of all sets B_i is called power tessellation of a union of discs.*

An example of the power tessellation is shown in Figure 2.

Since the power tessellation provides a subdivision of $\mathcal{U}$ into 2-dimensional convex sets with disjoint interiors, it becomes very useful when calculating values of geometrical characteristics of the union. Some examples of its usefulness are:

- Calculation of $A(\mathcal{U}_\mathbf{x})$: instead of the inclusion-exclusion formula

$$A(\mathcal{U}_\mathbf{x}) = \sum_i A(b_i) - \sum_{\{i_1,i_2\}} A(b_{i_1} \cap b_{i_2}) + \ldots + (-1)^{n+1} \sum_{\{i_1,\ldots,i_n\}} A(b_{i_1} \cap \cdots \cap b_{i_n})$$

  we use

$$A(\mathcal{U}_\mathbf{x}) = \sum_i A(B_i).$$

- Analogously we calculate $L(\mathcal{U}_\mathbf{x})$.
- For Euler-Poincaré characteristic $\chi(\mathcal{U}_\mathbf{x}) = N_{\mathrm{cc}}(\mathcal{U}_\mathbf{x}) - N_{\mathrm{h}}(\mathcal{U}_\mathbf{x})$, we use its equivalent definition (see [10])
  - $\chi(K_i) = 1$ for K_i compact convex set,
  - $\chi(\cup_{i=1}^n K_i) = \sum_{k=1}^n (-1)^{k+1} \sum_{\{i_1,\ldots,i_k\}} \chi(K_{i_1} \cap \cdots \cap K_{i_k})$,

  from which we get

$$\chi(\mathcal{U}_\mathbf{x}) = N_{\mathrm{c}}(\mathcal{U}_\mathbf{x}) - N_{\mathrm{ie}}(\mathcal{U}_\mathbf{x}) + N_{\mathrm{iv}}(\mathcal{U}_\mathbf{x}), \tag{4}$$

where N_c is the number of nonempty cells in the tessellation, N_{ie} the number of interior edges (i.e. lines formed by nonempty intersections of exactly two cells of the tessellation) and N_{iv} the number of interior vertices (i.e. points formed by nonempty intersections of exactly three cells). From the assumption of general position it follows that there are no other addends in (4).

- All the calculations are local in the sense that when we add or delete a disc, only the cells intersected by this disc can be changed, while the rest of the tessellation is unchanged.

The detailed algorithm for construction of the new tessellation in the case of adding a disc or deleting a disc, respectively, when the old tessellation is known, can be found in [6]. Moreover in [4], implementation of this algorithm to the program written in C++ is described.

6. Estimating parameters

6.1. MCMC maximum likelihood

Denote $f_\theta(\mathbf{x}) = h_\theta(\mathbf{x})/c_\theta$ (i.e. $h_\theta(\mathbf{x}) = \exp(\theta \cdot T(\mathcal{U}_\mathbf{x}))$ is the unnormalized density). For an observation $\mathbf{x}$ (or more often $\mathcal{U}_\mathbf{x}$), the log likelihood function is given by

$$l(\theta) = \log h_\theta(\mathbf{x}) - \log c_\theta = \theta \cdot T(\mathcal{U}_\mathbf{x}) - \log c_\theta.$$

The problem is that c_θ has no explicit expression. However, we can work with likelihood ratio instead, since for fixed θ_0, the term c_θ/c_{θ_0} in log likelihood ratio

$$l(\theta) - l(\theta_0) = \log(h_\theta(\mathbf{x})/h_{\theta_0}(\mathbf{x})) - \log(c_\theta/c_{\theta_0})$$

can be approximated by

$$c_\theta/c_{\theta_0} \sim \frac{1}{n}\sum_{m=1}^{n} h_\theta(Y_m)/h_{\theta_0}(Y_m), \tag{5}$$

where Y_m are realizations from $f_{\theta_0}(\mathbf{x})$ obtained from MCMC simulations.

Usually, a large number of simulation is needed for the approximation (5). For example for analyzing the set on Figure 1 described in [7], a few millions of such simulations were used, so the power tessellation provided really significant simplification.

6.2. Integral characterization of Gibbs process

In [1], Dereudre shows a Gibbs property of Quermass-interaction process. The corollary is that for this process, we can use the equation (6) known as integral characterization of Gibbs process (see e.g. [3], [9] or [8]). It says that if S grows to $\mathbf{R}^2$ and the reference process $\mathbf{Y}$ as well as the disc process $\mathbf{X}$ are stationary, intensity of $\mathbf{Y}$ is ρ and denoting B a set of all discs (or equivalently the space $\mathbf{R}^2 \times (0, \infty)$ of the corresponding points) then for an arbitrary measurable function $g : N \times B \to \mathbb{R}$ it holds that

$$\text{(6)} \qquad \mathrm{E}\sum_{x\in\mathbf{X}} g(\mathbf{X}\setminus x, x) = \rho \mathrm{E}\int_{\mathbf{R}^2\times(0,\infty)} g(\mathbf{X}, y)\lambda_\theta(\mathbf{X}, y)\mathrm{d}u\ \mathrm{d}Q(r),$$

where u is the center of the disc y and $Q(r)$ is a probability measure on its radius. Practically, it means that if the observation window W for the data $\mathbf{x}$ is large enough then we can use the approximation

(7)

$$\begin{aligned}\sum_{x\in\mathbf{x}} g(\mathbf{x}\setminus x, x) &= \rho \sum_{u\in W_{grid}, r\in Q_{grid}} g(\mathbf{x}, y)\lambda_\theta(\mathbf{x}, y) \\ &= \rho \sum_{u\in W_{grid}, r\in Q_{grid}} g(\mathbf{x}, y)\exp\left(\theta_1 A(\mathbf{x}, y) + \ldots + \theta_3\chi(\mathbf{x}, y)\right),\end{aligned}$$

where W_{grid} is a discretization of W and Q_{grid} is a discretization of the support of Q multiplied by the corresponding probability weights.

A study of how to choose suitable functions g and solve the equation (7) to obtain estimations of the parameters is in progress and will be presented in [2]. Nevertheless, from (7), it is seen that λ_θ must be calculated many times, and so acceleration provided by the power tessellation is used again.

References

[1] Dereudre D. *Existence of Quermass-Interaction Process for non locally stable interaction and non bounded convex grains.* Advances in Applied Probability **41** (3), 664 – 681.

[2] Dereudre D., Helisová K., Lavancier F. (2010) *Estimating parameters in Quermass-interaction process.* In preparation.

[3] Georgii H.-O. (1976) *Canonical and grand canonical Gibbs states for continuum systems.* Communications of Mathematical Physics **48**, 31 – 51.

[4] Helisová K. (2009) *Models for random union of interacting discs.* Doctoral thesis, Charles University in Prague, Faculty of Mathematics and Physics.

[5] Kendall W.S., van Lieshout M.N.M., Baddeley A.J. (1999) *Quermass-interaction processes: conditions for stability.* Advances in Applied Probability **31**, 315 – 342.

[6] Møller J., Helisová K. (2008) *Power diagrams and interaction processes for unions of discs.* Advances in Applied Probability **40** (2), 321 – 347.

[7] Møller J., Helisová K. (2009) *Likelihood inference for interacting discs.* Scandinavian Journal of Statistics, accepted.

[8] Møller J., Waagepetersen R.P. (2004) *Statistical Inference and Simulation for Spatial Point Processes.* Chapman and Hall/CRC, Boca Raton.

[9] Nguyen X.X., Zessin H. (1979) *Integral and differential characterizations of Gibbs processes.* Mathematische Nachrichten **88**, 105 – 115.

[10] Stoyan D., Kendall W.S., Mecke J. (1995) *Stochastic Geometry and Its Applications.* Wiley, Chichester.

Acknowledgement: This research was supported by Czech Government research program MSM6840770038.

Address: Czech Technical University in Prague, Faculty of Electrical Engineering, Department of Mathematics, Technická 2, 166 27 Prague 6 – Dejvice

E-mail: `helisova@math.feld.cvut.cz`

NEPARAMETRICKÁ KALIBRÁCIA – PREHL'AD

Klára Hornišová

Kľúčové slová: Jednorazová a simultánna kalibrácia, vierohodnostný a aposteriórny prístup, tolerančné oblasti, kopula, neparametrická regresia.

Abstrakt: Článok ponúka porovnanie viacerých neparametrických kalibračných oblastí.

Abstract: Paper offers comparison of selected nonparametric calibration regions.

1. Úvod

Kalibráciou sa v štatistike rozumie odhad neznámych hodnôt náhodnej veličiny (vektora, procesu) $X \in \mathcal{X}$ podľa nameraných (zodpovedajúcich) hodnôt veičiny $Y \in \mathcal{Y}$ s využitím predpokladov či informácií o ich združenom rozdelení (v ďalšom uvažujeme iba 1-rozmerné X a Y). Typicky sa používa, ak sa Y dá odmerať podstatne jednoduchšie ako X, lež nepresnejšie, no špeciálnym prípadom sú aj nezávislé rovnako rozdelené X a Y. Hlavnou úlohou je pre namerané hodnoty Y_j, $j = 1, ...,$ nájsť kalibračné oblasti $\mathcal{C}(Y_j) \subseteq \mathcal{X}$, ktoré by s dostatočne veľkou pravdepodobnosťou $1 - \alpha$ porade pokrývali neznáme zodpovedajúce hodnoty X_j,

$$,, P\left(X_j \in \mathcal{C}(Y_j)\right) = (\geq)\ 1 - \alpha".$$

Nech $\exists$ združená hustota $f(x, y) = f_{X,Y}(x, y)$.

V literatúre sa vyskytujú dva základné postupy, ako sformulovať kalibračnú úlohu. V ideálnom prípade, ak sú pravdepodobnostné rozdelenia celkom známe, možno ich opísať takto:

i) „aposteriórny"postup:

Ak sú známe podmienené hustoty $f(x|y)$ pre $\forall y \in \mathcal{Y}$, tak pre $\forall y \in \mathcal{Y}$

$$P(\mathcal{C}(y)|Y = y) = 1 - \alpha\ ,$$

napríklad

$$\mathcal{C}(y) = \langle f_{\alpha/2}(.|y), f_{1-\alpha/2}(.|y)\rangle.$$

ii) „vierohodnostný"postup:

Ak sú známe podmienené hustoty $f(y|x)$ pre $\forall x \in \mathcal{X}$, tak pre $\forall Y \in \mathcal{Y}$ treba nájsť $\mathcal{C}(Y) \subset \mathcal{X}$;

$$P(x \in \mathcal{C}(Y)|X = x) = 1 - \alpha,\ ; \forall x \in \mathcal{X}\ ,$$

často v inverznom tvare

$$\mathcal{C}(Y) = \{x \in \mathcal{X}; Y \in \mathcal{A}\ (x, 1 - \alpha)\}\ ,$$

kde

$$P(Y \in \mathcal{A}\ (x, 1-\alpha)|X = x) = 1 - \alpha\ ;\ \forall x \in \mathcal{X}$$

teda $\{\mathcal{A}(x, 1-\alpha); x \in \mathcal{X}\}$ je tolerančná oblasť pre Y, napríklad

$$\mathcal{A}(x, 1-\alpha) = \langle f_{\alpha/2}(.|x), f_{1-\alpha/2}(.|x)\rangle.$$

Ak je známa $f(x, y)$, dá sa postupovať podľa i) aj ii), no vo všeobecnosti sú výsledné kalibračné oblasti rôzne:

Pre nezávislé X, Y:

$$\begin{aligned}\mathcal{C}_{ii}(Y) &= \{x; Y \in \langle f_{\alpha/2}(.|x), f_{1-\alpha/2}(.|x)\rangle\} = \{x; Y \in \langle f_{Y,\alpha/2}(.), f_{Y,1-\alpha/2}(.)\rangle\} \\ &= \begin{cases} \emptyset, & \text{ak } Y \notin \langle f_{Y,\alpha/2}(.), f_{Y,1-\alpha/2}(.)\rangle \\ \mathcal{X}, & \text{ak } Y \in \langle f_{Y,\alpha/2}(.), f_{Y,1-\alpha/2}(.)\rangle \end{cases} \\ &\quad \text{(t.j. } \mathcal{C}_{ii} \text{ závisí od } Y) \\ &\neq \langle f_{X,\alpha/2}(.), f_{X,1-\alpha/2}(.)\rangle = \mathcal{C}_i(Y)\end{aligned}$$

Teda aposteriórny postup sa v tomto jednoduchom príklade javí ako prirodzený a správny [14], kým vierohodnostný ako nesprávny, no píše sa o ňom viac.

V skutočnosti $f_{X,Y}(.,.)$, $f(.|x)$, $f(.|y)$, ... nebývajú (v úplnosti) známe, takže sa odhadujú na základe údajov $\mathcal{D}_n := (X_i, Y_i)$, $i = 1, \ldots, n$, z kalibračného experimentu, kde

$$(X_i, Y_i), i = 1, \ldots, n \equiv \begin{cases} \text{náhodný výber z } f_{X,Y}(.,.)\ (\equiv \text{náhodná kalibrácia}) \\ X_i \in \mathcal{X},\ Y_i \sim f(.|X = X_i)\ (\equiv \text{riadená kalibrácia}) \end{cases}$$

(Okrem týchto dvoch možností sa môžu (vo viacrozmerných úlohách) vyskytovať aj iné. Kým napríklad usporiadanie $\dim X = 1$, $\dim Y = 2$, $Y = (Y_I, Y_{II})$, $\mathcal{D} = (X_i, Y_{I,i}, Y_{II,i})$, $i = 1, \ldots, n$, X a Y_{II} sú nezávislé, $X_i \in \mathcal{X}$, $Y_{II,i}$ - náhodný výber z apriórneho rozdelenia $f(y_{II})$, $Y_{I,i} \sim f(.|X = X_i, Y_{II} = Y_{II,i})$ - úlohou je pre budúce dvojice pozorovaní (Y_I, Y_{II}) odhadnúť zodpovedajúce hodnoty veličiny X - ešte zodpovedá riadenej kalibrácii, tak situácia, kde $Y_{II,i} \in \mathcal{Y}_{II}$, nie je ani v jednej z tých dvoch tried.

Ani údaje $\mathcal{D}$ s replikáciami, $Y_{ij} \sim f(.|X = X_i)$, $j = 1, \ldots, m_i$, či už $X_i \sim f(x)$ alebo $X_i \in \mathcal{X}$, nepatria ani do jednej z tých dvoch tried.)

Pri vierohodnostnom postupe sa z $\mathcal{D}$ vypočítajú odhady $\widehat{f}(.|x) = \widehat{f}(.|x, \mathcal{D}_n)$ hustôt $f(.|x)$ pre $\forall x \in \mathcal{X}$, a pre $\forall y \in \mathcal{Y}$ sa nájde kalibračná oblasť $\mathcal{C}(Y, \mathcal{D}_n)$, buď jednorazová

$$P_{\mathcal{D}_n}\left(P_{Y|X}(x \in \mathcal{C}(Y, \mathcal{D}_n)|X = x, \mathcal{D}_n) \geq 1 - \alpha\right) \geq 1 - \delta$$

alebo simultánna

$$P_{\mathcal{D}_n}\left(P_{Y|X}(x \in \mathcal{C}(Y, \mathcal{D}_n)|X = x, \mathcal{D}_n) \geq 1 - \alpha\ ;\ \forall x \in \mathcal{X}\right) \geq 1 - \delta\ ,$$

(pri celkom známych $f_{Y|X}(.|x)$ takéto rozlišovanie nemalo význam)
často v inverznom tvare

$$\mathcal{C}(Y, \mathcal{D}_n) = \{x \in \mathcal{X}; Y \in \mathcal{A}_{\ n}(x, \mathcal{D}_n, 1-\alpha, 1-\delta)\}\,,$$

kde

$$P_{\mathcal{D}_n}\big(P_{Y|X}(Y \in \mathcal{A}\ (x, \mathcal{D}_n, 1-\alpha, 1-\delta)|X = x, \mathcal{D}_n) \geq 1-\alpha\big) \geq 1-\delta,$$

alebo

$$P_{\mathcal{D}_n}\big(P_{Y|X}(Y \in \mathcal{A}\ (x, \mathcal{D}_n, 1-\alpha, 1-\delta)|X = x, \mathcal{D}_n) \geq 1-\alpha; \forall x \in \mathcal{X}\big) \geq 1-\delta,$$

teda $\{\mathcal{A}(x, \mathcal{D}_n, 1-\alpha, 1-\delta); x \in \mathcal{X}\}$ je jednorazová alebo simultánna tolerančná oblasť pre Y.

Pri aposteriórnom postupe sa z $\mathcal{D}$ a apriórnej hustoty $f(x)$ odhadujú prediktívne hustoty $f_{X|Y}(.|y)$ a kalibračná oblasť sa zostrojí ako oblasť najväčších hodnôt hustoty $f_{X|Y}(.|y)$ a s pokrytím $1-\alpha$, a teda na rozdiel od vierohodnostného postupu neuplatňuje hladinu spoľahlivosti $1-\delta$. Pri tomto postupe sa navyše nerozlišuje jednorazová kalibrácia od simultánnej.

Závislosti a pravdepodobnostné rozdelenia veličín X a Y sú navyše často zneprehľadnené ďalšími parametrami, napospol rušivými. Napríklad aposteriórne postupy sa môžu navzájom líšiť podľa toho, či je dané apriórne rozdelenie pre X alebo pre skutočné hodnoty Y, ak sa merajú s chybami, a pre (spoločné) parametre hustôt $f_{Y|X}(.|X = x)$ [8, 10].

Zostrojenie kalibračnej oblasti pre veličinu X (ktorá je zhruba podielom Y a parametra sklonu v regresnom modeli pre $f_{Y|X}$) ako inverzie tolerančnej oblasti pre Y je obmenou Fiellerovej úlohy, takže v prípade slabšie informatívnych údajov $\mathcal{D}$ môže s nenulovou pravdepodobnosťou viesť k neohraničeným intervalom, vrátane $(-\infty, \infty)$. No aj aposteriórne riešenia sú citlivé na apriórne rozdelenie $f(x)$, a menej aj na apriórne rozdelenia rušivých parametrov, ak je nimi $f(x)$ určené len implicitne.

Jestvujú práce o stotožňujúcich apriórnych rozdeleniach (matching priors), pri ktorých sa aposteriórna kalibračná oblasť približne približne zhoduje s vierohodnostnou [3, 4, 5, 21].

2. Niektoré kalibračné metódy

1. Krishnamoorthy a Mathew [13] podľa Mee, Eberhardt a Reeve (1991) inverzia simultánnej tolerančnej oblasti v modeli

$$f(y|x) \sim N(m^\top(x)\beta, \sigma^2); \forall x$$

kde $m(.)$ - známa, $\beta \in \mathbb{R}^p, \sigma^2$ - neznáme.

$$\mathcal{A}\ (x, \mathcal{D}) := \langle m^\top(x)\hat{\beta} - k(x)S, m^\top(x)\hat{\beta} + k(x)S\rangle\ ,$$

kde

$$\hat{\beta} = (M^\top M)^{-1}M^\top D_2, \quad S^{\,2} = (D_2 - M\hat{\beta})^\top(D_2 - M\hat{\beta})/(n-p),$$
$$M_{ij} := m_j(X_i), i = 1, \ldots, n, j = 1, \ldots, p,$$
$$D_2 := (Y_1, \ldots, Y_n)^\top, d^2 := d^2(x) := m^\top(x)(M^\top M)^{-1}m(x).$$

Potom

$$P_{Y|X}(Y \in \mathcal{A}\ (x, \mathcal{D})|X = x, \mathcal{D}) \geq \Phi(dW + k(d)U) - \Phi(dW - k(d)U)\ ,$$

kde $U \sim N(0,1)$, $W^2 \sim \chi^2_p$ sú nezávislé, $k(d) = \lambda\left[z_{1-\alpha/2} + (p+2)^{1/2}d\right]$, a λ je také, že

$$P_{W^2,U^2}\left(\min_d \left(\Phi(dW + k(d)U) - \Phi(dW - k(d)U)\right) \geq 1-\alpha\right) = 1-\delta.$$

2. Witkovský a Chvosteková [20] inverzia simultánnej tolerančnej oblasti pre rovnaký model:

$$\mathcal{A}\,(x,\mathcal{D}) := \langle \inf_{\substack{(\beta,\sigma);\\ \lambda(\mathcal{D},\beta,\sigma)\leq\lambda_{1-\delta}}} (m^\top(x)\beta + u_{\,\alpha_1}\sigma),\quad \sup_{\substack{(\beta,\sigma);\\ \lambda(\mathcal{D},\beta,\sigma)\leq\lambda_{1-\delta}}} (m^\top(x)\beta + u_{\,1-\alpha_2}\sigma)\rangle,$$

kde $\alpha_1 + \alpha_2 = \alpha$ a

$$\lambda(\mathcal{D},\beta,\sigma) := (\mathcal{D}_2 - M\beta)^\top(D_2 - M\beta)/\sigma^2 - n\log((n-p)S^2/(n\sigma^2)) - n\ ,$$
$$\lambda(\mathcal{D},\beta,\sigma) \sim \lambda \sim Q_p + Q_{n-p} - n\log(Q_{n-p}) + n(\log n - 1)\ ,$$

kde $Q_p \sim \chi^2_p$ a $Q_{n-p} \sim \chi^2_{n-p}$ sú nezávislé.

3. Brown [2] aposteriórna bayesovská nesimultánna riadená kalibrácia:

$f(y|x) = N(\mu + \beta x, \sigma^2)$

apriórne rozdelenie: $\pi(\mu,\beta,\sigma^2,x) = \pi(\mu,\beta,\sigma^2)\pi(x)$,

Jeffreysov prior: $\pi(\mu,\beta,\sigma^2) \propto \sigma^{-2}$

$\Rightarrow$ aposteriórne: $(\mu-\hat\mu)|\sigma^2 \sim N(0,\frac{\sigma^2}{n}), (\beta-\hat\beta)|\sigma^2 \sim N(0,\sigma^2(\mathcal{D}_1^\top\mathcal{D}_1)^{-1})$,

$$\frac{\|\mathcal{D}_2 - \hat\mu - \mathcal{D}_1\hat\beta\|^2}{\sigma^2} \sim \chi^2_{n-2}$$

prediktívne rozdelenie $Y|x$:

$$\frac{\sqrt{n-2}}{\|\mathcal{D}_2 - \hat\mu - \mathcal{D}_1\hat\beta\|}\,\frac{Y(x) - \hat\mu - \hat\beta x}{(1 + 1/n + x(\mathcal{D}_1^\top\mathcal{D}_1)^{-1}x)^{1/2}} \sim t_{n-2}$$

$\Rightarrow$ aposteriórne rozdelenie $x|(\mathcal{D}, Y=y)$:

$\pi(x|\mathcal{D}, Y=y) \propto \pi(Y(x)|\mathcal{D},x)\pi(x|\mathcal{D}_1)$

napríklad: $\pi(x|\mathcal{D}_1) \propto (1 + \frac{1}{n} + x(\mathcal{D}_1^\top\mathcal{D}_1)^{-1}x)^{-(n-2)/2}$

4. Gruet [7] neparametrická simultánna vierohodnostná kalibrácia:

$$f(y - Ef(.|x)|x) = \frac{1}{\sigma}\chi\left(\frac{x}{\sigma}\right)\ ,$$

kde $Ef(.|x)$ - neznáma funkcia, rastúca alebo klesajúca v x,

$\chi(.)$ je známa hustota, $\chi(x) = \chi(-x)$.

Bodový kalibračný odhad $\hat x$ neznámej hodnoty x sa určí minimalizáciou vzdialenosti od oblaku dát ako riešenie rovnice

$$H_n(x,Y) := \widehat{Ef(Y|x)} = \frac{1}{n}\sum_{i=1}^n K_{\,h_n}(x - X_i)\,\Psi(Y_i - Y) = 0\ ,\ \text{kde}$$

$K_h(u) = h^{-1}K(h^{-1}u)$, $K(.)$ – jadro s nosičom $\langle -A, A\rangle$, t.j. $K(.) \geq 0$, $K(-x) = K(x)$, $K(-A) = K(A)$, $\int K(u)du = 1$.

Nepárna neklesajúca funkcia $\Psi(.)$ váži vplyv odľahlých pozorovaní, napríklad

$$\Psi(u) = \begin{cases} u \\ \max\{-\kappa, \min(u, \kappa)\} \ , \ \kappa > 0 \end{cases} \quad \text{(Huberova funkcia)}.$$

Ak je riešení $\hat{x}$ viac, vyberie sa to najbližšie k X_i, kde Y_i je najbližšie k Y.

Pre $\Psi(u) = u$:

$$H_n(x, Y) = 0 \ \equiv \ \frac{\frac{1}{n}\sum_{i=1}^{n} K_{\ h_n}(x - X_i)(Y_i - Y)}{\frac{1}{n}\sum_{i=1}^{n} K_{\ h_n}(x - X_i)} = 0 \ \equiv \ \widehat{Ef(.|x)} - Y = 0,$$

teda $\widehat{Ef(.|x)} = \widehat{E_0 f(.|x)}$ je Nadarayov-Watsonov odhad, t.j. lokálne polynomický odhad funkcie $Ef(.|x)$ stupňa 0 [19].

Pre všeobecné $\Psi(.)$:

$$H_n(x, Y) = 0 \ \equiv \ \frac{\frac{1}{n}\sum_{i=1}^{n} K_{\ h_n}(x - X_i)\frac{1}{g_n}L'(\frac{Y_i - Y}{g_n})}{\frac{1}{n}\sum_{i=1}^{n} K_{\ h_n}(x - X_i)} = 0 \ \equiv \ \frac{\partial}{\partial y}\Big|_{y=Y} \widehat{{}_0 f(y|x)}$$

$$\equiv \ \frac{\partial \widehat{f(y|x)}}{\partial y}\Big|_{y=Y} = 0,$$

kde $L(u) := -\int \Psi(u)du$ je jadro, teda formálne by sa odhad $\hat{x}$ určoval tak, aby v $y = Y$ bol stacionárny bod, a často súčasne maximum hustoty $f(y|x)$. $\hat{x}$ by sa mohlo hľadať aj ako riešenie rovnice

$$\frac{\partial \widehat{f(y|x)}}{\partial y}\Big|_{y=Y} = 0,$$

t.j. namiesto derivácie odhadu funkcie $f(y|x)$ by sa použil odhad jej derivácie. Je však známe, že takéto odhady v neparametrickej regresii zaveľa nestoja.

Simultánne intervalové kalibračné odhady neznámych hodnôt x sa hľadajú ako simultánne tolerančné množiny $\mathcal{C}(Y, \mathcal{D}_n) = \mathcal{C}(Y)$ s vlastnosťou

$$\liminf_{n\to\infty} \quad P_{\mathcal{D}_n}\left(P_{Y|X}(x \in \mathcal{C}(Y)|x) \geq 1 - \alpha; \forall x\right) \geq 1 - \delta,$$

a to v tvare $\mathcal{C}(Y) = \{x \in \mathcal{X}; |\widehat{E_0 f(.|x)} - Y| \leq c\}$, $c = c(1 - \alpha, 1 - \delta)$.

Pri istých zjednodušujúcich predpokladoch sa c vypočíta s využitím aproximácie rozdelenia pravdepodobnosti supréma náhodného procesu z článku [1], ktorá však platí len pri predpoklade asymptotickej stacionarity procesu a je už prekonaná napríklad trubicovými metódami [15] - trubicové vylepšenia majú význam najmä vo viacrozmerných prípadoch.

V bodovom odhade $\hat{x}$ možno nahradiť Nadarayov-Watsonov odhad lokálne polynomickým odhadom nepárneho, napríklad 1., stupňa, čo zmenšuje výchylku.

5. Misquitta [16], Misquitta a Ruymgaart [17] riadená neparametrická kalibrácia:

$m(x) := Ef(.|X = x)$ - neznáma funkcia, monotónna v $x \in \mathcal{X}$
$\widetilde{m}(.) : m(\mathcal{X}) \to \mathcal{Y}$ - inverzná funkcia k $m(.)$
Var $f(.|x) = \sigma^2 \in <0, \infty; \forall x$
$y = m(x)$

Bodové kalibračné odhady neznámej hodnoty x:

a)

$$\check{x} = \arg\min_{x \in \mathcal{X}} \frac{1}{n} \sum_{i=1}^{n} K_{\ h}(x - X_i)(y - Y_i)^2$$

b)

$$\widehat{x} = \arg\min_{x \in \mathcal{X}} (y - \widehat{m}(x))^2 \ ,$$

kde

$$\widehat{m}(x) = \frac{\sum_{i=1}^{n} K_{\ h}(x - X_i) Y_i}{\sum_{i=1}^{n} K_{\ h}(x - X_i)}$$

(Nadarayov-Watsonov odhad pre $m(x) = Ef(.|X = x)$)

c)

$$\check{x} = \check{m}(Y) \ , \text{ kde } \check{m}(.) \text{ je odhadom funkcie } \widetilde{m}(.) \ ,$$

$$\check{m}(y) = \frac{\sum_{i=1}^{n} K_{\ h}(y - Y_i) X_i}{\sum_{i=1}^{n} K_{\ h}(y - Y_i)}$$

(Nadarayov-Watsonov odhad pre $Ef(.|Y = y)$)

Prvé dva odhady (analógie klasického bodového kalibračného odhadu z obyčajnej lineárnej regresie) sú konzistentné, tretí (analógia inverznej kalibrácie) je konzistentný, len ak sú v $\mathcal{D}_n$ replikácie pre $\forall X_i; i = 1, \ldots, n$.

6. Na kalibráciu v zmiešaných a nelineárnych parametrických regresných modeloch sa používa bootstrap a trubicové metódy [9, 12, 11].

7. Ďalšie prístupy ku kalibrácii využívajú napríklad Kalmanov filter a pojem hĺbky [22].

3. **Kopuly**

Pri neparametrickom aposteriórnom prístupe by sa dal využiť odhad podmienenej hustoty navrhnutý v článku [6] využívajúci pojem kopuly.

2-rozmerná kopula je funkcia $C : \langle 0, 1\rangle^2 \to \langle 0, 1\rangle$ s vlastnosťami [18]

$C(0, u) = C(u, 0) = 0$, $C(1, u) = C(u, 1) = u$; $\forall u \ \in \langle 0, 1\rangle$

C je 2-rastúca: ak $a < b$ a $c < d$;

$C(b, d) - C(a, d) - C(b, c) + C(a, c) \geq 0.$

Sklarova veta:

$\forall F_{X,Y}(.,.), \exists C : \langle 0, 1\rangle^2 \to \langle 0, 1\rangle; \forall (x, y); F_{X,Y}(x, y) = C(F(x), G(y))$.

Pre spojité $F(.)$, $G(.)$ $\exists! C$; $C(u, v) = F_{X,Y}(F^{-1}(u), G^{-1}(v))$.

$C(.,.)$ je distribučná funkcia na $\langle 0,1\rangle^2$ s rovnomernými marginálnymi rozdeleniami.

$\forall F(.), G(.), \forall C(.,.); H(x,y) := C(F(x), G(y)) \Rightarrow H(.,.)$ - distribučná funkcia s okrajmi F a G.

Ak je (X,Y) spojitý náhodný vektor a $p(.), q(.)$ sú neklesajúce funkcie, tak $\forall$ kopula vektora $(p(X), q(Y))$ sa zhoduje s kopulou vektora (X,Y) na množine $p(\mathcal{X}) \times q(\mathcal{Y})$.

Faugeras [6] navrhuje pre podmienenú hustotu

$$f(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = g(y)c\left(F(x), G(y)\right)$$

namiesto odhadu v tvare podielu

$$\widetilde{f}_n(y|x) = \frac{\hat{f}_{n;X,Y}(x,y)}{\hat{f}_{n;X}(x)} \text{ , kde}$$

$$\hat{f}_{n;X,Y}(x,y) = \frac{1}{n}\sum_{i=1}^{n} K'_{h'}(X_i - x) K_h(Y_i - y)$$

$$\hat{f}_{n;X}(x) = \frac{1}{n}\sum_{i=1}^{n} K'_{h'}(X_i - x) \text{ , a jeho variantov,}$$

odhad v tvare súčinu

$$\hat{f}_n(y|x) = \hat{g}_n(y)\hat{c}_n(F_n(x), G_n(y)) \text{ , kde}$$

$$\hat{g}_n(y) = \frac{1}{nh_n}\sum_{i=1}^{n} K_0\left(\frac{y - Y_i}{h_n}\right) ,$$

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{<X_i,\infty)}(x) \text{ , } G_n(y) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{<Y_i,\infty)}(y) ,$$

$$\hat{c}_n(u,v) = \frac{1}{na_nb_n}\sum_{i=1}^{n} K\left(\frac{u - F_n(X_i)}{a_n}, \frac{v - G_n(Y_i)}{b_n}\right) ,$$

$$a_n = b_n \text{ , } K(u,v) = K_1(u)K_2(v)$$

pri

$$K_0(x) = \frac{3}{4}(1 - x^2)\mathbb{I}_{<-1,1>}(x) \quad \text{(Epanečnikovovo jadro) ,}$$

$$\hat{c}_n(u,v) \rightsquigarrow c_n(u,v) = \frac{1}{n}\sum_{i=1}^{n} K_{u,a_n}(U_i) K_{v,a_n}(V_i) ,$$

$$K_{x,b}(t) = \beta_{\frac{x}{b+1}, \frac{1-x}{b+1}}(t) \quad \text{(beta jadro) , } \beta_{a,b}(t) = \frac{t^{a-1}(1-t)^{b-1}}{B(a,b)}.$$

Pravda, zatiaľ sú známe len lokálne asymptotické vlastnosti odhadu $\hat{f}(y|x)$.

Odhad $\hat{f}(y|x)$ sa dá priamo využiť iba pri náhodnej kalibrácii.

Literatúra

[1] Bickel P.J., Rosenblatt M. (1973) *On some global measures of the deviations of density function estimates*. Ann. St. **1**, 6, 1071 – 1095. Opravy: (1975) Ann. St. **3**, 6, 1370.

[2] Brown P.J. (1993) *Measurement, regression, and calibration*. OUP, Clarendon Press, Oxford.

[3] Eno D.R. (1999) *Noninformative prior bayesian analysis for statistical calibration problems*. PhD Thesis, Virginia polytechnic Institute, Blacksburg.

[4] Eno D.R., Ye K. (2000) *Bayesian reference prior analysis for polynomial calibration models*. Test **9**, 191 – 208.

[5] Eno D.R., Ye K. (2001) *Probability matching priors for an extended statistical calibration model*. Can. J. St. **29**, 19 – 35.

[6] Faugeras O.P. (2009) *A quantile-copula approach to conditional density estimation*. J. of Mult. Analysis **100**, 2083 – 2099.

[7] Gruet M.-A. (1996) *A nonparametric calibration analysis*. Ann. Stat. **24**, 4, 1474 – 1492.

[8] Hoadley B. (1970) *A bayesian look at inverse linear regression*. JASA **65**, 356 – 369.

[9] Huet S. et al. (2004) *Statistical tools for nonlinear regression*. A practical guide with S-Plus and R examples, 2nd ed., Springer, New York.

[10] Hunter W.G., Lamboy W.F. (1981) *A bayesian analysis of the linear calibration problem, with discussion*. Technometrics **23**, 323 – 350. Opravy: (1984) Technometrics **26**, 69.

[11] Choudhary P.K. (2007) *Semiparametric regression for assessing agreement using tolerance bands*. Preprint, Univ. of Texas at Dallas, Richardson.

[12] Choudhary P.K., Ng H.K.T. (2006) *Assessment of agreement under nonstandard conditions using regression models for mean and variance*. Biometrics **62**, 288 – 296.

[13] Krishnamoorthy K., Mathew T. (2009) *Statistical tolerance regions: Theory, Applications and Computation*. J. Wiley, New York.

[14] Lindley D.V. (1972, 1995) *Bayesian statistics, a review*. SIAM, Philadelphia, 1st, 6th printing.

[15] Loader C. (1999) *Local regression and likelihood*. Springer, New York.

[16] Misquitta P.P. (2000) *Some results in non-parametric calibration*. M.S. Thesis, Texas Tech Univ., Lubbock.

[17] Misquitta P., Ruymgaart F.H. (2005) *Some results on nonparametric calibration*. Comm. in St. - Theory and Methods **34**, 1605 – 1616.

[18] Volauf P. (2005) *O asociácii náhodných veličín a kopulách*. Forum statisticum Slovacum **3**, 91 – 98.

[19] Wasserman L. (2006) *All of nonparametric statistics*. Springer, New York.

[20] Witkovský V., Chvosteková M. (2009) *Simultaneous tolerance intervals for the linear regression model*. Measurement 2009.

[21] Yin M. (2000) *Noninformative priors for multivariate linear calibration*. J. Mult. Analysis **73**, 221 – 240.

[22] Zappa D., Salini S. (2003) *Some notes on confidence regions in multivariate calibration*. E.P.N. 118, Univ. Cattolica del S. Cuore, Milano.

Poďakovanie: Na výskum prispela agentúra Vega grantmi 1/0077/09 a 2/0019/10.

Adresa: Ústav merania SAV, Dúbravská 9, 841 01 Bratislava

E-mail: `umerhorn@savba.sk`

ELEMENTY STATISTICKÉ ANALÝZY KOMPOZIČNÍCH DAT

Karel Hron

Klíčová slova: Kompoziční data, relativní informace, Aitchisonova geometrie na simplexu, charakteristiky polohy a variability.

Abstrakt: Mnoho reálných dat v přírodních a společenských vědách i v různých dalších disciplínách jsou ve skutečnosti kompoziční data, protože pouze podíly mezi proměnnými poskytují relevantní informaci. Kompoziční data jsou reprezentována Aitchisonovou geometrií na simplexu. Pro možnost aplikace standardních statistických metod, vytvořených za předpokladu euklidovských vlastností výběrového prostoru, je potřeba vyjádřit kompoziční data jako souřadnice vzhledem k ortonormální bázi či generujícímu systému na simplexu. Úkolem příspěvku je představit základní teoretické aspekty problematiky a populární prostředky popisné statistiky pro tento typ dat. Uveden je též přehled softwarových balíků, které jsou k dispozici v R.

Abstract: Many practical data sets in environmental and social sciences and various other disciplines are in fact compositional data because only the ratios between the variables are informative. Compositional data are represented in the Aitchison geometry on the simplex, and for applying standard statistical methods designed for the Euclidean geometry they need to be expressed in an orthonormal basis or in a generating system of compositions on the simplex. The aim of the paper is to introduce the basic theoretical background as well as the most popular exploratory tools for this kind of data. Also software packages, available in R, are mentioned.

1. Kompoziční data a jejich geometrie

Datové soubory jsou často charakterizovány vícerozměrnými pozorováními s kvantitativně vyjádřenými relativními příspěvky částí na celku. Jako příklady lze uvést koncentrace chemických prvků v hornině, měsíční výdaje domácnosti na různé komodity (jídlo, bydlení, doprava, a podobně) nebo procentuální výskyt různých živočišných druhů ve zkoumané oblasti. Často právě procenta jsou používána k vyjádření zmíněných relativních velikostí složek těchto dat a proto obvykle hovoříme o simplexu jako jejich výběrovém prostoru. Nicméně, situace je obecnější, protože jediná relevantní informace v datech je obsažena v podílech mezi jejich složkami. Z tohoto pohledu reprezentují procenta pouze vhodné vyjádření informace, obsažené v mnohorozměrných pozorováních. Tyto úvahy vedly na počátku osmdesátých let minulého století Johna Aitchisona k zavedení pojmu *kompoziční data* (nebo též zkráceně kompozice) k charakterizování tohoto typu dat a k navržení možností jejich statistické analýzy s využitím tzv. logratio transformací [1].

Geometrie kompozičních dat, označena později jako Aitchisonova, bere jejich výše uvedené charakteristické vlastnosti do úvahy a je založena na speciálních operacích perturbace, mocninná transformace a Aitchisonově skalárním součinu [7]. Podrobněji, pro D-složkové kompozice $\mathbf{x} = (x_1, \ldots, x_D)$, $\mathbf{y} = (y_1, \ldots, y_D)$ a reálné číslo α, takto postupně obdržíme kompozice

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C}(x_1 y_1, \ldots, x_D y_D),\ \alpha \odot \mathbf{x} = \mathcal{C}(x_1^\alpha, \ldots, x_D^\alpha)$$

a reálné číslo

$$\langle \mathbf{x}, \mathbf{y} \rangle_A = \frac{1}{D} \sum_{i=1}^{D-1} \sum_{j=i+1}^{D} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}.$$

S využitím vlastností Hilbertova prostoru vede Aitchisonův skalární součin také k definicím Aitchisonovy normy a vzdálenosti. Přitom symbol $\mathcal{C}$ označuje operaci uzávěru, která transformuje součet složek kompozice na zvolenou konstantu κ (bez ztráty informace). Jak bylo uvedeno výše, za konstantu κ obvykle bereme 1 nebo 100, abychom mohli reprezentovat kompozice na D-složkovém simplexu (dimenze $D-1$),

$$\mathcal{S}^D = \{\mathbf{x} = (x_1, \ldots, x_D),\ x_i > 0,\ \sum_{i=1}^{D} x_i = \kappa\}.$$

Z geometrických vlastností kompozičních dat lze snadno odvodit, že standardní statistické metody jako např. metoda hlavních komponent, faktorová analýza nebo korelační analýza, navržené za předpokladu euklidovských vlastností výběrového prostoru a standardních mnohorozměrných dat s absolutní škálou, mohou vést (a často také vedou) k zavádějícím výsledkům. Toto lze demonstrovat na mnoha příkladech, uvedených v literatuře [1, 8, 9, 13].

Řešením je vyjádřit kompoziční data v souřadnicích vzhledem k (nějaké) ortonormální bázi $\{\mathbf{e}_1, \ldots, \mathbf{e}_{D-1}\}$ na simplexu (s Aitchisonovou geometrií). Potom lze totiž zřejmě každou kompozici $\mathbf{x} \in \mathcal{S}^D$ zapsat jako

$$\mathbf{x} = \langle \mathbf{x}, \mathbf{e}_1 \rangle_A \odot \mathbf{e}_1 \oplus \cdots \oplus \langle \mathbf{x}, \mathbf{e}_{D-1} \rangle_A \odot \mathbf{e}_{D-1}.$$

Pokud označíme

$$\mathbf{x}^* = h(\mathbf{x}) = (x_1^*, \ldots, x_{D-1}^*) = (\langle \mathbf{x}, \mathbf{e}_1 \rangle_A, \ldots, \langle \mathbf{x}, \mathbf{e}_{D-1} \rangle_A),$$

pak $(D-1)$-rozměrný reálný vektor $\mathbf{x}^*$ obsahuje právě souřadnice vzhledem k uvedené ortonormální bázi. Zobrazení h tedy transformuje kompozice izometricky z $\mathcal{S}^D$ do $\mathbf{R}^{D-1}$ a často se též nazývá izometrická logratio (ilr) transformace [5]. V důsledku toho je Aitchisonova geometrie nahrazena euklidovskou a platí následující vztahy (pro reálná čísla α, β),

$$h(\alpha \odot \mathbf{x} \oplus \beta \odot \mathbf{y}) = \alpha h(\mathbf{x}) + \beta h(\mathbf{y}) = \alpha \mathbf{x}^* + \beta \mathbf{y}^*,\ \langle \mathbf{x}, \mathbf{y} \rangle_A = \langle \mathbf{x}^*, \mathbf{y}^* \rangle_E,$$

kde poslední výraz značí obvyklý euklidovský skalární součin. Analogické vztahy by platily i pro normu a vzdálenost kompozic [7].

Volba ortonormální báze na S^D je klíčová pro interpretaci souřadnic. Při její konstrukci je preferován postup, nazývaný postupné binární dělení

(PBD) [6, 7], protože umožňuje interpretaci ve smyslu skupin složek kompozice. Při samotné konstrukci souřadnic (nazývaných v tomto kontextu též bilance neboli rovnováhy) postupujeme následovně. V prvním kroku rozdělíme složky kompozice do dvou skupin; složky první skupiny označíme $+1$ a složky druhé skupiny -1. Takto obdržíme první souřadnici, která vyjadřuje rovnováhu mezi $+1$ a -1 složkami a zastupuje takto vlastně podíly mezi jednotlivými $+1$ složkami na jedné straně a -1 složkami na straně druhé. Ve druhém a následujících krocích je získaná skupina složek ($+1$ či -1) opět rozdělena na dvě nové skupiny, podobně označené pomocí $+1$ a -1, zatímco nezahrnuté složky jsou označeny 0. Získané souřadnice v každém kroku (pro zahrnuté složky) mají analogickou interpretaci jako předtím. Počet kroků, potřebných k dosažení skupin obsahujících pouze jedinou složku, je přesně $D-1$, tedy dimenze $\mathcal{S}^D$. Pokud označíme počet $+1$ a -1 v i-tém kroku jako r_i a s_i, dostaneme nové souřadnice ve tvaru

$$x_i^* = \sqrt{\frac{r_i s_i}{r_i + s_i}} \ln \frac{\prod_+ x_j^{1/r_i}}{\prod_- x_k^{1/s_i}},\ i = 1, \ldots, D-1;$$

přitom součin s indexem $+$ (resp. $-$) probíhá přes složky označené $+1$ (resp. -1) v i-tém kroku. Celá procedura se obvykle zapisuje do tabulky, pro podrobnosti lze odkázat na [7].

Volbě konkrétního PBD obvykle předchází hlubší posouzení studovaného problému, což následně umožní kvalifikovanou interpretaci nových souřadnic a vztahů mezi nimi. Pokud ovšem aplikujeme na datový soubor nějakou objektově orientovanou statistickou metodu jako např. diskriminační analýzu, získané výsledky jsou na volbu báze invariantní (volby dvou různých ortonormálních bází na simplexu se projeví jako ortogonální transformace souřadnic). Jedinou vyjímkou, kdy se jeví použití ortonormálních souřadnic jako méně výhodné, je v případě biplotu pro kompoziční data [2], kdy preferujeme vyjádření kompozic v generujícím systému na simplexu prostřednictvím tzv. centrované logratio (clr) transformace [1], definované pro kompozici $\mathbf{x}$ jako

$$\mathbf{y} = clr(\mathbf{x}) = \left(\ln \frac{x_1}{\sqrt[D]{\prod_{i=1}^D x_i}}, \ldots, \ln \frac{x_D}{\sqrt[D]{\prod_{i=1}^D x_i}} \right).$$

Ačkoliv souřadnice v tomto případě svádějí k interpretaci ve smyslu původních složek kompozice (uvedená transformace je dokonce též izometrická), jsou výsledná data singulární, protože součet složek $\mathbf{y}$ je roven nule. To činí následně obtíže při jejich statistickém zpracování, např. při použití robustních metod [8].

2. **Příprava dat na zpracování - nuly a chybějící hodnoty**

Práce s logaritmy podílů složek implikuje nutnost pouze nenulových hodnot složek kompozic. Nicméně, také nulová pozorování se mohou v reálných datech vyskytovat, a to jako hodnoty pod mezí detekce, nebo vinou úplné

absence příslušné složky v kompozici. První případ se často vyskytuje při geochemických měřeních a v současné době již existuje k nahrazení nul několik přístupů, založených mimo jiné na použití EM-algoritmu a vyjádření kompozic v bázi, která není ortonormální [10, 11]. Druhá situace (tzv. strukturní nuly) se naopak častěji vyskytuje při výběrových šetřeních; je zřejmé, že např. vybrané osoby - nekuřáci budou mít při zjišťování struktury výdajů domácností nulové výdaje za cigarety. Pro diskuzi tohoto problému a návrhy možných řešení lze odkázat na [3].

Druhým častým problémem, který je třeba vyřešit před samotným zahájením statistické analýzy kompozičního datového souboru, je imputace chybějících hodnot. Vzhledem k tomu, že jediná relevantní informace je v tomto případě obsažena v podílech mezi složkami (a s absencí každé složky tak zároveň ztratíme též tyto příslušné podíly), je nutné tomuto faktu samotnou imputaci přizpůsobit. Možné řešení je navrženo v [9], kde je tento problém řešen dvoustupňovým algoritmem. V prvním kroce je provedena úvodní imputace pomocí metody k-nejbližšího souseda (k-nearest neighbor, k-NN, [16]) a v druhé fázi je použito iterativního algoritmu, založeného na regresní imputaci. Přitom v k-NN se obvykle používá k nalezení nejbližších sousedů euklidovské vzdálenosti, která je v tomto případě nahrazena vzdáleností Aitchisonovou,

$$d_A(\mathbf{x},\mathbf{y}) = \sqrt{\frac{1}{D}\sum_{i=1}^{D-1}\sum_{j=i+1}^{D}\left(\ln\frac{x_i}{x_j} - \ln\frac{y_i}{y_j}\right)^2},$$

a modifikována vzhledem k charakteru kompozic [9]. Bohužel, k-NN plně nedokáže zachytit mnohorozměrné vztahy mezi kompozičními složkami, tyto jsou uvažovány pouze nepřímo při hledání k nejbližších sousedů. Z tohoto pohledu je zřejmé, že může být kvalita imputace zlepšena použitím iterativního modelově orientovaného postupu. Pro jeho specifičnost bude tento blíže představen v následujících dvou odstavcích.

V každém kroku iterace této regresní imputace je jedna proměnná jako vysvětlovaná a ostatní slouží jako regresory, tedy mnohorozměrná informace je využita pro imputaci hodnot této proměnné. Protože ovšem pracujeme s kompozičními daty, nemůžeme pro regresi použít původní složky, ale je nutno pracovat v souřadnicích. Nicméně, již pro samotnou konstrukci bilancí je potřeba mít k dispozici datovou matici bez chybějících hodnot. Tento problém lze vyřešit právě inicializováním chybějících hodnot pomocí k-NN imputace, jak bylo zmíněno výše. Dalším problémem je samotná volba bilancí, protože nekvalitní inicializace chybějících hodnot může následně zapříčinit svého druhu šíření chyby. Bilance proto volíme v následujícím tvaru,

$$x_i^* = \sqrt{\frac{D-i}{D-i+1}}\,\ln\frac{\sqrt[D-i]{\prod_{l=i+1}^{D}x_l}}{x_i}, \qquad \text{pro } i = 1,\ldots,D-1,$$

což zaručí nejvyšší možnou stabilitu vzhledem k chybějícím hodnotám. Například, chybějící hodnoty, nahrazené v první složce x_1, ovlivní pouze první bilanci x_1^*, ale nemají již žádný vliv na zbývající ortonormální souřadnice.

Takovou volbou PBD tedy dosáhneme toho, že je chybějícími hodnotami ovlivněn nejmenší možný počet bilancí. Nechť složka x_1 obsahuje největší počet chybějících hodnot, x_2 druhý největší počet, atd., tedy při regresi x_1^* na $x_2^*, \ldots, x_{D-1}^*$ je inicializovanými chybějícími hodnotami v x_1 ovlivněna pouze proměnná x_1^*, která reprezentuje všechny podíly x_1 s ostatními složkami kompozice.

Základní myšlenka metody je pak založena na iterativním zlepšování odhadů chybějících hodnot. Po provedení regrese x_1^* na $x_2^*, \ldots, x_{D-1}^*$ jsou výsledky vyjádřeny ve tvaru původních složek kompozice pomocí vztahů

$$\begin{aligned}
x_1 =& \exp\left(-\frac{\sqrt{D-1}}{\sqrt{D}}x_1^*\right),\\
x_i =& \exp\left(\sum_{l=1}^{i-1}\frac{1}{\sqrt{(D-l+1)(D-l)}}x_l^* - \frac{\sqrt{D-i}}{\sqrt{D-i+1}}x_i^*\right), i = 2, \ldots, D-1,\\
x_D =& \exp\left(\sum_{l=1}^{D-1}\frac{1}{\sqrt{(D-l+1)(D-l)}}x_l^*\right).
\end{aligned}$$

(až na uzávěr) a původně chybějící hodnoty v datové matici jsou aktualizovány. Dále je tentýž postup použit pro složku s druhým největším počtem původně chybějících hodnot, atd. Až jsou takto analyzovány všechny složky, celý proces začíná znovu až do ustálení odhadovaných chybějících hodnot dle zvoleného kritéria. Detailní popis algoritmu je k dispozici v [9].

3. Základní číselné charakteristiky

Standardní nástroje popisné a induktivní statistiky bohužel v případě kompozičních dat neposkytují smysluplné informace. Zejména aritmetický průměr a rozptyl nebo směrodatná odchylka nekorespondují s Aitchisonovou geometrií jako charakteristiky polohy a variability. Tuto skutečnost lze ilustrovat na četných příkladech, pro podrobnosti lze odkázat např. na [1]. Ostatně, problémy s určením korelačního koeficientu [13] iniciovaly zájem o tento typ dat. Z povahy zkoumaných pozorování totiž vyplývají následující vlastnosti, které by měla každá relevantní charakteristika při (nejen) statistické analýze kompozičních dat respektovat:

- *Invariantnost na změnu škály*: Informace obsažená v kompozici nezávisí na jednotkách, ve kterých je tato vyjádřena. Kladné násobky vektoru s kladnými složkami totiž vyjadřují tutéž kompozici (jako třídu ekvivalence). Každá smysluplná charakteristika by tedy měla být invariantní na změnu škály.
- *Invariantnost na permutaci*: Permutace složek nemění informaci, obsaženou v kompozici.
- *Podkompoziční soudržnost*: Informace získaná z kompozice o D složkách by neměla být ve sporu s informací, získanou z podkompozice o d

složkách (vzniklé výběrem složek původní kompozice), $d \leq D$. Speciálně lze přitom zmínit, že každá relevantní charakteristika, která je funkcí složek kompozice, je výhradně funkcí podílů těchto složek. V podkompozici tyto charakteristiky závisí pouze na podílech vybraných složek a nikoli na vynechaných složkách původní kompozice. Toto je potřeba si uvědomit zejména v souvislosti s požadavkem invariantnosti na změnu škály.

Proto je nutné představit příslušné alternativy, centrum, matici rozptylů a celkový rozptyl. Uvažujme datovou matici $\underline{\mathbf{X}}$ o n řádcích a D sloupcích s prvky x_{ik}, obsahující v řádcích pozorované kompozice. Potom charakteristikou polohy pro kompoziční data je uzavřený geometrický průměr, nazývaný též centrum a definovaný jako

$$\mathbf{g} = \mathcal{C}(g_1, \ldots, g_D), \quad g_i = \left(\prod_{k=1}^{n} x_{ik} \right)^{1/n}.$$

Je zřejmé, že centrum již plně vyhovuje z hlediska vlastností, uvedených výše. Disperze v souboru kompozičních dat se nejčastěji popisuje pomocí (normované) matice rozptylů souřadnic jednotlivých podkompozic (x_{ik}, x_{jk}), kde $i, j = 1, \ldots, D, k = 1, \ldots, n$, tedy

$$\mathbf{T}^* = \begin{pmatrix} t^*_{11} & t^*_{12} & \cdots & t^*_{1D} \\ t^*_{21} & t^*_{22} & \cdots & t^*_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ t^*_{D1} & t^*_{D2} & \cdots & t^*_{DD} \end{pmatrix},$$

kde prvky t^*_{ij} představují rozptyl souboru $\{z^k_{ij} = \frac{1}{\sqrt{2}} \ln \frac{x_{ik}}{x_{jk}}, k = 1, \ldots, n\}$. Míra celkové variability souboru, celkový rozptyl, je potom dána vztahem

$$\mathrm{totvar}(\underline{\mathbf{X}}) = \frac{1}{D} \sum_{i=1}^{D} \sum_{j=1}^{D} t^*_{ij}.$$

Interpretace prvků matice $\mathbf{T}^*$ je intuitivní; jestliže je hodnota t^*_{ij} blízká nule, značí to, že podíly mezi i-tými a j-tými složkami kompozic v souboru jsou velmi stabilní. Někdy se též užívá charakteristika $\exp(-t^*_{ij})$, která se realizuje v intervalu $(0, 1)$. Je přitom ovšem potřeba si uvědomit, že ani jedna z nich nenahrazuje korelační koeficient ve smyslu míry těsnosti lineárního vztahu mezi statistickými znaky. Aplikaci těchto charakteristik na reálná data lze nalézt např. v [4].

Z pohledu induktivní statistické analýzy je potom výběrové centrum nejlepším nestranným odhadem centra distribuce náhodné kompozice (vzhledem k Aitchisonově geometrii, resp. celkovému rozptylu odhadu) [12]. Dále, z definice je matice $\mathbf{T}^*$ zřejmě symetrická a s nulami na hlavní diagonále; přitom zřejmě jak její prvky, tak hodnota celkového rozptylu nezávisí na konstantě κ, asociované s výběrovým prostorem $\mathcal{S}^D$, tedy změna škály nemá žádný efekt. Uvedené charakteristiky variability mají navíc další důsledky. Je zřejmé, že

celkový rozptyl shrnuje matici rozptylů v jedinou hodnotu. Tato vlastnost je přitom přirozená, protože všechny složky v kompozici sdílí společnou škálu. Naopak, matice rozptylů vysvětluje, jak je celkový rozptyl rozdělen mezi složky kompozice (resp. mezi logaritmy jejich podílů - logratios).

4. Softwarová podpora v R

Programovací jazyk a softwarové prostředí R [14] dnes představuje zřejmě nejpopulárnější software pro statistickou analýzu dat, volně dostupný z adresy `http://cran.r-project.org`. Pro práci s kompozičními daty jsou k dispozici dvě knihovny, `compositions` [17] a `robCompositions` [15]. První obsahuje ucelený přehled funkcí pro základní statistickou analýzu kompozičních dat (např. též volbu PBD nebo imputaci nulových hodnot v datech podle [10]) a kompletní archiv zkušebních datových souborů z [1]. Druhý balíček je zaměřený na robustní analýzu kompozic včetně detekce odlehlých hodnot, metody hlavních komponent, faktorové analýzy, diskriminační analýzy a imputace chybějících hodnot, společně s odpovídajícími grafickými nástroji.

Stručný informační rozcestník o kompozičních datech (v češtině) je k dispozici na adrese `http://compositions.sweb.cz`.

Literatura

[1] Aitchison J. (1986) The statistical analysis of compositional data. Chapman and Hall, London.

[2] Aitchison J., Greenacre J. (2002) *Biplots of compositional data*. Applied Statistics **51**, 375 – 392.

[3] Bacon-Shone, J. (2003) *Modelling structural zeros in compositional data*. Thió-Henestrosa S., Martín-Fernández J.A., eds., Compositional Data Analysis Workshop – CoDaWork'03, Proceedings. Universitat de Girona, ISBN 84-8458-111-X, http://ima.udg.es/Activitats/CoDaWork03/.

[4] Daunis-i-Estadella J., Barceló-Vidal C., Buccianti A. (2006) *Exploratory compositional data analysis*. In Buccianti A., Mateu-Figueras G., Pawlowsky-Glahn V., eds., Compositional data analysis in the geosciences: From theory to practice. Geological Society, London, Special Publications **264**, 161 – 174.

[5] Egozcue J.J., Pawlowsky-Glahn V., Mateu-Figueraz G., Barceló-Vidal C. (2003) *Isometric logratio transformations for compositional data analysis*. Mathematical Geology **35**, 279 – 300.

[6] Egozcue J.J., Pawlowsky-Glahn V. (2005) *Groups of parts and their balances in compositional data analysis. Mathematical Geology* **37**, 795,- 828.

[7] Egozcue J.J., Pawlowsky-Glahn V (2006) *Simplicial geometry for compositional data*. In Buccianti A., Mateu-Figueras G., Pawlowsky-Glahn V., eds., Compositional data analysis in the geosciences: From theory to practice. Geological Society, London, Special Publications **264**, 145 – 160.

[8] Filzmoser P., Hron K., Reimann C. (2009) *Principal component analysis for compositional data with outliers*. Environmetrics **20**, 621 – 632.

[9] Hron K., Templ M., Filzmoser P. (2010) *Imputation of missing values for compositional data using classical and robust methods*. Computational Statistics and Data Analysis, v tisku.

[10] Martín-Fernández, J.A., Barceló-Vidal, C., Pawlowsky-Glahn, V. (2003) *Dealing with zeros and missing values in compositional data sets using nonparametric imputation.* Mathematical Geology **35** *3*, 253 – 278.

[11] Palarea-Albaladejo, J., Martín-Fernández, J. A. (2008) *A modified EM alr-algorithm for replacing rounded zeros in compositional data sets.* Computer & Geosciences **34**, 902 – 917.

[12] Pawlowsky-Glahn V., Egozcue J.J. (2002) *BLU estimators and compositional data.* Mathematical Geology **34**, 259,- 274.

[13] Pearson K. (1897) *Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurement of organs.* Proceedings of the Royal Society of London **60** (1897), 489 – 502.

[14] R Development Core Team (2010) *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Wien.

[15] Templ M., Hron K., Filzmoser P. (2010) *robCompositions: Robust estimation for compositional data.* Manuál a knihovna, verze 1.3.3.

[16] Troyanskaya O., Cantor M., Sherlock G., Brown P., Hastie T., Tibshirani R., Botstein D., Altman R. (2001) *Missing value estimation methods for DNA microarrays.* Bioinformatics **17**, 520 – 525.

[17] van den Boogaart K.G., Tolosana-Delgado R., Bren M. (2008) *compositions: Compositional data analysis.* Manuál a knihovna, verze 1.01-1.

Poděkování: Tato práce byla podporována grantem MSM 6198959214.

Adresa: Přírodovědecká fakulta Univerzity Palackého, Katedra matematické analýzy a aplikací matematiky, tř. 17. listopadu, 771 46 Olomouc

E-mail: `hronk@seznam.cz`

PHILOSOPHICAL CONCEPTION OF PROBABILITY IN THE WORK OF T. G. MASARYK AND K. VOROVKA

Magdalena Hykšová

Keywords: Philosophical interpretations of probability, inductive logic, history of mathematics.

Abstract: The contribution is devoted to philosophical conception of probability in the work of two inventive Czech thinkers who are well-known in a quite different context: T. G. Masaryk and K. Vorovka. Masaryk dealt with probability in his inaugural lecture at the Philosophical Faculty of the Prague university in 1882, which was published as a separate booklet one year later, and he belonged to enthusiastic proponents of logical interpretation of probability. Vorovka dealt with probability and its philosophical meaning in several treatises published in the years 1912 – 1925, and contrary to Masaryk, he belonged to critics of both logical and subjective interpretations.

Abstrakt: Příspěvek je věnován filosofickému pojetí pravděpodobnosti v díle dvou originálních českých myslitelů, kteří jsou všeobecně známí ve zcela odlišných souvislostech: T. G. Masaryka a K. Vorovky. Masaryk se pravděpodobností zabýval ve své inaugurační přednášce na Filosofické fakultě pražské univerzity v roce 1882, která byla o rok později vydána jako samostatná publikace, a patřil mezi nadšené zastánce logické interpretace pravděpodobnosti. Vorovka se pravděpodobností a jejím filosofickým významem zabýval v několika pojednáních z let 1912 – 1925 a patřil naopak ke kritikům logické i subjektivní interpretace.

1. Introduction

From the strict mathematical point of view, probability can be introduced as a real function over a σ-algebra of sets (modelling events) with values in the interval $[0, 1]$ and satisfying certain axioms, which leads to a nice theory in the sense of Kolmogorov [7]. After this introduction, probability was recognized as an adequate branch of mathematics (in spite of using rather inductive than deductive logic). Nevertheless, such an explanation does not seem satisfactory to philosophers and all other scientists who would like to use probability theory in the real world. Therefore, they have been trying for a long time to find an answer to the seemingly simple question, what the probability really is, how to interpret it. The two main groups of interpretations are usually distinguished, namely *epistemological theory* that identifies probability with the degree of our knowledge or belief or experience and has two branches – *logical* and *subjective*, and *objective theory* (with another two branches – *frequency* and *propensity interpretations*) that considers probability to be

the feature of the objective material world unrelated to human being and its knowledge or belief.

This paper relates to the first group. Logical interpretation, which identifies probability with the degree of rational belief that is equal for all people having the same evidence, is mainly ascribed to J. M. Keynes ([6], 1921). The acknowledged founders of subjective interpretation are F. P. Ramsey ([12], 1931) and Bruno de Finetti ([4], 1937), who considered probability to be a degree of belief that can differ for different people with the same evidence. It is remarkable that the conception of logical interpretation was already developed by B. Bolzano in the 1830's [2],[1] subjective theory was deeply investigated by the Czech priest Václav Šimerka in the 1880's ([16], [17]).

2. Tomáš Garrigue Masaryk (1850 – 1937)

It is certainly not necessary to introduce T. G. Masaryk, the first president of Czechoslovakia. Nevertheless, it is not widely known that he was also interested in probability theory, above all in its philosophical foundations. Let us briefly recall that Masaryk studied philosophy and philology at the university in Vienna. Among his teachers we can find for example Robert Zimmermann (1824 – 1898), the scholar of Bernard Bolzano,[2] and Franz Brentano (1838 – 1917), the founder of analytical metaphysics; Brentano's school of thought appreciated Bolzano's treatise *Wissenschaftslehre* [2] and made it better known. Nevertheless, it concerned above all Bolzano's philosophy of logical realism and much less his own logic or "Wissenschaftstheorie"[scientific theory]. In any case, we can suppose that through his teachers, Masaryk was influenced by Bernard Bolzano, too, at least in the field of philosophy. 1878 Masaryk habilitated at the Vienna university with the sociological treatise *Suicide as a Mass Phenomenon of Modern Civilization*. Four years later he was appointed extraordinary professor of philosophy at the university in Prague, 1897 he earned the full professorship there.

2.1. Hume's scepsis and probability calculus

For his inaugural lecture at the Prague university held on December 16, 1882, Masaryk chose the theme *Hume's Scepsis and Probability Calculus*. He developed the topic further in the treatise [9] published as a separate booklet

[1] In the first half of the 20$^{\text{th}}$ century, Bolzano was often cited and his work was considered important. But later the contributions written in English came into the foreground.

[2] Towards the end of his life Bolzano sought a continuator of his mathematical work. Finally he invested his hopes to young Robert Zimmermann to whom he then willed his mathematical manuscripts. But Zimmermann concentrated only on philosophy and later became professor of philosophy (1852 in Prague, 1861 in Vienna). In 1882 he handed Bolzano's mathematical inheritance over to the Vienna Academy of Sciences and it passed it on, while Zimmermann was still alive, to the manuscript department of the Vienna Court Library, later National Library. Here it lied unnoticed till the end of the World War I; the efforts to organize and publish all manuscripts is still in progress.

in 1883; one year later he published a shortened and slightly modified German variant [10]. Although all other Masaryk's publications were devoted to philosophy, sociology and later politics, and although the character of the mentioned treatises was primarily philosophical, it is remarkable that they display Masaryk's wide acquaintance with the development of probability theory, mainly in the connection with inductive logic.

In the introduction to [9], Masaryk remarks that he has occupied his mind with this topic for a long time and that he plans to treat it in a more extensive treatise on the theory of inductive logic. But because of doubts whether he will accomplish this project in the near future (and as we know today, he has never accomplished it), he presents at least this essay that explains the basic ideas at the background of the historical development to facilitate understanding and possibly further development by philosophers and other interested people. Masaryk explicitly notes that the treatise demonstrates the logical meaning of probability theory; in today terminology, we can therefore say that it belongs to the field of the *logical interpretation of probability.* The treatise is an answer to Hume's idea that inductive inferences are solely based on habits, and since the concept of causal connection does not correspond to any impression of the external or internal experience, it is completely blank [5]. Masaryk characterizes the principle of Hume's scepsis by the following words: *Only mathematics deserves our confidence, empirical sciences are uncertain, since the recognition of causal connections of facts evades us; because we can gain reliable knowledge only on the basis of an evident relation between the cause and an effect.* ([9], p. 24)

In the main part of his treatise, Masaryk describes the history of philosophical attempts to disprove Hume's scepsis, that he all finds insufficient. He starts with the ideas of philosophers of the Scottish School, Thomas Reid, James Beattie and James Oswald, then he comes to Immanuel Kant and Friedrich Eduard Beneke. He continues with the first attempts to disprove Hume's scepsis with the help of probability theory, namely the contributions of Johann Georg Sulzer, Moses Mendelssohn, Joseph Marie Degérando, Sylvestre-François Lacroix and Siméon Denis Poisson. Then he turns to inductive logic and its history. He discusses the work of Gottfried Wilhelm Leibniz, Jacob Bernoulli, Pierre-Simon Laplace, Adolphe Quetelet, Rudolf Herschel, John Venn etc. He concludes with a remark: *All these recent contributions lack an explicit relation to Hume; hereby they lack, I would say, the real point... Hume himself speaks about probability very often, but it seems that he does not know its mathematical rules, since he cannot distinguish subjective and objective probability, and it is therefore clear how he could have arrived at his sceptical theory of induction.* ([10], p. 14 – 15)

Unfortunately, it seems that Masaryk did not know relevant treatises of Bernard Bolzano yet: in [1] Hume is explicitly cited, in [2] Bolzano systematically builds inductive logic as the extension of deductive logic, based on probability theory.

2.2. **Affair of the manuscripts**

Masaryk's treatise [9] was cited by August Seydler [14] who appreciated it for stressing the importance of probability theory for inductive logic. Before discussing Seydler's paper, let us look at the context in which it was published. In 1817, two seemingly old Czech manuscripts were found; they were named after the places of their discovery: *Königinhof Manuscript* (bellow abbreviated KM) and *Grünberg Manuscript* (GM). The first of them was supposed to be dated from the 13^{th} century, the second one from the 9^{th} or 10^{th} century. Shortly afterwards, doubts about their authenticity arose. Firstly they concerned mainly the GM that would have been the oldest known Czech manuscript at all (there exists a continual series of old Czech manuscripts since the 13^{th} century), later also the KM. Nevertheless, defenders were exceptionally persistent, both manuscripts had an important impact on the literature of the Czech romanticism, and in the time of the Czech National Revival, they represented an important symbol of Czechs. Masaryk was looking for the truth, even though he came in for a lot of hate among Czech nationalists. He invited the opponents of the authenticity to publish their arguments in the journal Athaeneum that he had founded and edited. In the third volume of this journal from the year 1886, we can thus find various philological, historical, sociological, aesthetical or paleographical grounds for the fact that both manuscripts were forgeries. Perhaps the most important was the paper by the philologist and literary historian Jan Gebauer [3] who found many grammatical deviations from the rules of the old Czech grammar (extracted from undoubtedly authentic manuscripts) and coincident occurrence of "suspicious" forms in the manuscripts and in other writings from the beginning of the 19^{th} century, written before the discovery of KM and GM.

As far as the mentioned philological arguments are concerned, the historian Josef Kalousek, one of the most active defenders of the authenticity, claimed that the deviations and coincidences are only accidental. August Seydler, the above mentioned friend of Masaryk and the professor of astronomy and theoretical physics at the university in Prague, decided to calculate the probability that all those suspicious forms are really accidental. The result was published in the couple of papers [14] and [15].[3] Seydler restricted himself to the Königinhof Manuscript that was defended more fiercely, and he arrived at the estimation of the probability that all deviations from the old grammar are accidental: $P_1 < 1/3,48 \cdot 10^9$. Even smaller was the probability of the mentioned coincidences: $P_2 < 1/10^{14}$. Seydler therefore concluded that the deviations and coincidences cannot be attributed to the mere chance and had to be explained. Although these arguments seem to be convincing, Kalousek and other defenders of the authenticity did not admit them and insisted that all oddities were accidental. Nevertheless, towards the end of the 19^{th} century, most scholars inclined to the hypothesis that both manuscripts were forgeries. Definitely it was proved in 1967.

[3] For more details, see the paper [26] by J. Zichová.

3. Karel Vorovka (1879 – 1929)

Karel Vorovka, philosopher of mathematics, opposed Masaryk's optimism, criticized philosophical interpretations of probability as well as the positivistic trend in general. Let us recall that Vorovka studied mathematics and physics at the Philosophical Faculty of Charles University in Prague and then worked as a secondary school professor for 20 years. His doctoral thesis was devoted to integral theory; nevertheless, soon he turned his interest towards philosophy and philosophical problems of mathematics. In 1919 he habilitated for philosophy of natural sciences at the Philosophical Faculty of the Prague university, two years later he was appointed extraordinary professor of philosophy of exact sciences at the recently established Faculty of Science of the same university; in 1924 he became full professor there. In 1920 – 1927 Vorovka participated in publication of the idealistically oriented journal *Ruch filosofický* [Philosophical Action], he took also part in several international philosophical congresses. Among physicists, Vorovka is known as promoter and defender of Einstein's relativity theory.

3.1. Influence of Henry Poincaré

As Vorovka himself often remarked, he was strongly influenced by Henri Poincaré. The first explicit reference can be found in the treatise [18] explaining the principle of Poincaré's *conventionalism*. Another topic influenced by Poincaré concerns the relevance of logical and intuitive elements for mathematics. The concept of an intuition can best be illustrated by the following words: *Both a mere empiricism and a mere logic would be only groping in the dark if they were not associated with the most intellectual, the most internal power of genius, often opposing all senses, a power which moves the whole mechanism of logic and which is perhaps the very true intellect: that is intuition.* ([22],[4] p. 156)

Besides [22], Vorovka dealt with this theme in the book [23], later accepted as the habilitation treatise. The book [23] starts with the discussion of the concept of *logicism* according to which mathematics is the part of logic, all mathematical concepts can explicitly be defined from logic and all mathematical propositions can be deduced from logical axioms and definitions by the mere logical deduction; this field recognizes the concept of *autonomous truth* (e.g., abstract mathematical propositions) that exists out of time and space and out of human consciousness. Vorovka explains the imperfections of logicism and then systematically exposes its opposite called *psychologism*. Later Vorovka turns from psychologism "to the neighbourhood of logicism", but still believes in rational intuition [24].[5]

[4]The paper represents an extended variant of the lecture *H. Poincaré as a Philosopher* held by Vorovka in the Union of Czech Mathematicians and Physicists on January 25, 1913, as one of the series of lectures commemorating the personality of H. Poincaré.

[5]For more details, see the paper [11] by L. Mazliak.

3.2. **Chance, probability, causality**

The influence of Henri Poincaré is apparent in many other Vorovka's treatises. From all of them, let us look at the papers dealing with probability theory and its philosophical meaning. The first [19] was published in 1912 and its character was rather mathematical: it discussed gambler's ruin problem and the history of its solutions, and it proposed the new proof of the fact that if the number of repetitions is not bounded, one of the players will certainly be ruined. Other treatises are more philosophical. For the theme of this paper, the most interesting ones are [20] and [21] where Vorovka criticizes the efforts to base the theory of logical induction on probability theory, challenges to the caution when using probability theory in real situations, and tries to persuade the readers that it cannot solve the problem of causality; he stressed that the concept of cause and effect should be replaced by the concept of correlation. The paper [20] was published in 1913 in the philosophical journal Česká mysl [Czech Thought] under the title *Philosophical Reach of Probability Theory*. Here Vorovka criticizes philosophical interpretations of probability including the contributions of P.-S. Laplace [8], T. G. Masaryk [9] and V. Šimerka [16]. He clarifies the most substantial problem of the logical interpretation, which is the determination of prior probabilities in Bayes' formula for the probability of certain hypothesis, conditioned by an available observation or experience. Unlike Masaryk, Vorovka claims that Hume's objections are justified and they cannot be disproved by probability theory. He insists that probability calculus and Hume's scepsis belong to completely different intellectual areas and it is not possible to bring them into a rational relation. He compares the application of probability calculus to Hume's scepsis to cutting an atom by a knife, and the introduction of Hume's scepsis into probability calculus to sharpening the atoms in the knife.

Similar ideas can be found in the paper *On Probability of Causes* [21] published in 1914 in the journal Časopis pro pěstování mathematiky a fysiky [Journal for Cultivation of Mathematics and Physics; bellow abbreviated ČPMF]. The character of this article is rather popularizing. It was intended mainly for secondary school students; therefore, it contains less philosophy and more mathematics and illustration examples. Vorovka again investigates the possibilities of the use of Bayes' theorem for the proof of the causal connection of certain events, and shows that these possibilities are very limited. He formulates the basic problem in the following way: Certain phenomenon was observed, that must have been caused by one of a finite number n of various events (causes); denote the a priori probability of the k^{th} event by ω_k. Suppose that the events are pairwise excluding and no other possibilities exist, i.e., $\omega_1 + \omega_2 + \cdots + \omega_n = 1$. If the k^{th} of these events comes about, then the observed phenomenon arises with the probability p_k. Using Bayes' theorem, probability that the cause of the observed phenomenon was the k^{th} event (in other words, the k^{th} hypothesis is true) can be expressed in the form

$$h_k = \frac{\omega_k p_k}{\omega_1 p_1 + \omega_2 p_2 + \cdots + \omega_n p_n} \,. \tag{1}$$

Vorovka continues with the discussion of the problem how to assess the prior probabilities ω_k. Using several examples he shows that in some cases it is relatively simple but other times it is more complicated or even impossible. For example, imagine that Peter plays dice with an unknown player; the highest win goes to the player who gets two sixes in the first throw. When the unknown player starts to play, he rolls two sixes. What is the probability that he is a sharpie? If we put $\omega_1 = \omega_2 = 1/2$ (according to Laplace principle), this probability would be 0.97, which contradicts the common sense. We cannot therefore solve this problem without more information. Nevertheless, the conclusion of the paper partly softens the critique: *Yet Bayes' theorem should not be underestimated. After all, it is substantial for probability theory; on one hand, for applications to events ruled by the law of large numbers, on the other for the logical coherence of the whole calculus...* ([21], p. 93)

Let us add that the treatise [20] contains an interesting discussion how to determine prior probabilities in some cases with the help of Poincaré's method of arbitrary functions. More than ten years later Vorovka published a short paper [25] in which he reacted against the efforts to exclude the concept of causality from the scientific research. He concludes: *Functional and conditional thinking will always need to its complementarity the causal thinking.* ([25], p. 115)

4. **Conclusion**

The works of T. G. Masaryk and K. Vorovka represent two completely different conceptions of probability. Masaryk belonged to proponents of its logical interpretation; his treatises are also remarkable for the fact that they manifest good knowledge of probability theory and its history, enthusiasm for inductive logic and a high estimation of mathematics. Thus they show the first Czechoslovak president in the less usual light. On the contrary, Vorovka criticized probability interpretations and claimed that probability theory cannot be applied to philosophical problems but should be independent of philosophy. Let us add that soon it indeed happened: in the 1930's, A. N. Kolmogorov [7] based probability theory on axiomatic foundations, which led to its acceptance as the "real"mathematical discipline. Moreover, in the same time the logical interpretation faced a sharp critique of F. P. Ramsey and B. de Finetti that led to its gradual abandonment in the second half of the 20^{th} century. As we could see above, the core of this critique can already be found in the treatises of K. Vorovka. Nevertheless, serious attempts to bring the logical interpretation to life recently appeared (for more details, see e.g. the paper [13] by I. Saxl).

References

[1] Bolzano B. (1834) *Lehrbuch der Religionswissenschaft.* Sulzbach.
[2] Bolzano B. (1837) *Wissenschaftslehre.* Sulzbach [finished around 1830].
[3] Gebauer J. (1886) *Potřeba dalších zkoušek rukopisu Královédvorského a Zelenohorského* [Necessity of Furter Tests on Königinhof and Grünberg Manuscripts]. Athenaeum **3**, 152 – 164.
[4] Finetti B. de (1937) *La prévision: ses lois logiques, ses sources subjectives.* Annales de l'Institut Henri Poincaré **7**, 1 – 68.
[5] Hume D. (1748) *Philosophical Essays Concerning Human Understanding.* A. Millar, London [later renamed *An Enquiry Concerning Human Understanding*].
[6] Keynes J. M. (1921) *A Treatise on Probability.* Macmillan, London.
[7] Kolmogorov A. N. (1933) *Grundbegriffe der Wahrscheinlichkeitsrechnung.* Springer, Berlin.
[8] Laplace P. S. de (1814) *Essai Philosophique sur les Probabilités.* Paris.
[9] Masaryk T. G. (1883) *Humova skepse a počet pravděpodobnosti* [Hume's Scepsis and Probability Calculus]. J. Otto, Praha.
[10] Masaryk T. G. (1884) *Dav. Hume's Skepsis und die Wahrscheinlichkeitsrechnung.* Carl Konegen, Wien.
[11] Mazliak L. (2007) *An Introduction to Karel Vorovka's Philosophy of Randomness.* Journ@l Electronique d'Histoire des Probabilités et de la Statistique **3**, no. 2, 14 pp.
[12] Ramsey F. P. *The Foundations of Mathematics and Other Logical Essays.* Kegan Paul, Trench, Trübner & Co, London.
[13] Saxl I. (2004) *Filosofické interpretace pravděpodobnosti* [Philosophical Interpretations of Probability]. In: Bečvář J., Fuchs E. (eds): *Matematika v proměnách věků III.* VCDV, Praha, 132 – 155.
[14] Seydler A. (1886a) *Počet pravděpodobnosti v přítomném sporu* [Probability Calculus in the Present Dispute]. Athenaeum **3**: 299 – 308.
[15] Seydler A. (1886b) *Dodatek k mé úvaze o pravděpodobnosti* [Supplement to My Contemplation on Probability]. Athenaeum **3**: 446 – 449.
[16] Šimerka V. (1882) *Síla přesvědčení* [Power of Conviction]. ČPMF **11**, 75 – 111.
[17] Šimerka V. (1883) *Die Kraft der Überzeugung.* Sitzungsberichte der Philos.-Historischen Classe der Kaiserlichen Akad. der Wiss. **104**, 511 – 571.
[18] Vorovka K. (1909) *Konvencionalism* [Conventionalism]. Česká mysl **10**, 217 – 228.
[19] Vorovka K. (1912) *Poznámka k problému ruinování hráčů* [Note on Gambler's Ruin Problem]. ČPMF **41**, 562 – 567.
[20] Vorovka K. (1913) *Filosofický dosah počtu pravděpodobnosti* [Philosophical Reach of Probability Calculus]. Česká mysl **14**, 17 – 30.
[21] Vorovka K. (1914a) *O pravděpodobnosti příčin* [On Prob. of Causes]. ČPMF **43**, 81 – 93.
[22] Vorovka K. (1914b) *Jak soudil H. Poincaré o vztazích mathematiky k logice* [H. Poincaré's Opinions on the Relationships of Mathematics and Logic]. ČPMF **43**, 154 – 162.
[23] Vorovka K. (1917) *Úvahy o názoru v matematice* [Considerations on Opinion in Mathematics]. ČAVU, Praha.
[24] Vorovka K. (1921) *Skepse a gnóse* [Scepticism and Gnosticism]. Gustav Voleský, Praha.
[25] Vorovka K. (1925) *Poznámka o kausálním myšlení* [Remark on Causal Thinking]. Ruch filosofický **5**, 112 – 115.
[26] Zichová J. (2004) *Teorie pravděpodobnosti a rukopisný spor* [Probability Theory and the Affair of the Manuscripts]. PMFA **49**, 95 – 103.

Acknowledgement: The work was supported by the grant GAČR 401/09/1850.

Address: FD ČVUT, Ústav aplikované matematiky, Na Florenci 25, 110 00 Praha 1

E-mail: hyksova@fd.cvut.cz

SIMULTÁNNE OBOJSTRANNÉ TOLERANČNÉ INTERVALY V LINEÁRNOM REGRESNOM MODELI

Martina Chvosteková

Kľúčové slová: Lineárny regresný model, simultánne tolerančné intervaly, tolerančný faktor.

Abstrakt: V príspevku sa budeme zaoberať simultánnymi tolerančnými intervalmi, ktoré sú využívané v mnohých meracích úlohach, najmä pri kalibrácií meracích zariadení v prípade opakovaného dopredu neznámeho počtu meraní na zariadení (pozri [9], [7]). Uvedieme stručný prehľad známych metód na konštruovanie simultánnych tolerančných intervalov v lineárnej regresii s normálnymi chybami. Konkrétne spomenieme Liebermanovu-Millerovu metódu [5], Wilsonovu metódu [10], Limamovu-Thomasovu metódu a Modifikovanú Wilsonovu metódu [6].

Abstract: The simultaneous tolerance intervals are important for many measurement procedures. The most common application for simultaneous tolerance intervals is a multiple-use calibration problem; see e.g. [9], [7]. In this paper we present a brief overview of the methods for constructing simultaneous tolerance intervals in a linear regression with normal errors. In particular, we describe the Lieberman-Miller method [5], the Wilson method [10], the Limam-Thomas method and the modified Wilson method [6].

1. Úvod

Pri analyzovaní biomedicínskych, inžinierskych, či ekonomických úloh vystupuje regresný model ako najlepší prostriedok na vyjadrenie štatistickej závislosti medzi známymi vysvetľujúcimi a pozorovanými odpovedajúcimi premennými pre konkrétny uvažovaný problém. Častou úlohou je stanoviť na základe n nezávislých pozorovaní, ozn. $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ odpovedajúcich k daným vysvetľujúcim premenným $\boldsymbol{x}_i, i = 1, \ldots, n$ hranice pre K budúcich nezávislých pozorovaní $Y_1^*, \ldots, Y_K^*$ odpovedajúcich k daným $\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_K^*$. Pre prípad lineárneho regresného modelu s normálne rozdelennými nezávislými chybami, kde neznáme parametre modelu možno dopočítať metódou najmenších štvorcov, je riešenie pre známu hodnotu K uvedené v [4]. Ak je počet budúcich pozorovaní neznámy a ľubovoľne veľký je úloha riešená pomocou simultánnych tolerančných intervalov.

Tolerančný interval je definovaný *pokrytím* (content) γ, $\gamma \in (0, 1)$ a *úrovňou spoľahlivosti* $1 - \alpha$, $\alpha \in (0, 1)$. Pre populáciu s jednorozmerným rozdelením je tolerančný interval skonštruovaný na základe náhodného výberu tak, aby pokryl aspoň γ časť populácie so spoľahlivosťou $1 - \alpha$.

Na predikciu určeného počtu možných budúcich pozorovaní lineárnej kombinácie $\boldsymbol{x}^T\boldsymbol{\beta} + \epsilon$ pre pevné $\boldsymbol{x}$, kde $\epsilon \sim N(0, \sigma^2)$ a $(\boldsymbol{\beta}, \sigma)$ sú neznáme parametre regresného modelu, nie je možné použiť opakovane predikčný interval skonštruovaný pre jedno budúce pozorovanie. Šírka simultánnych predikčných intervalov, intervalov pokrývajúcich zároveň daný počet budúcich meraní, s zväčšujúcim sa počtom budúcich pozorovaní narastá a aj obtiažnosť získania numerického riešenia narastá. Teda pre prípad, že počet budúcich pozorovaných premenných je neznámy a ľubovoľne veľký, sa využíva tolerančný interval, ktorý vymedzuje interval pokrývajúci aspoň γ časť rozdelenia $Y(\boldsymbol{x})$ so spoľahlivosťou $1 - \alpha$. Za predpokladu, že $\boldsymbol{x}$ je pevné, ide o výber z jednorozmerného rozdelenia a vzťahy na výpočet jednostranných aj obojstranných tzv. nesimultánnych tolerančných intervalov sú uvedené v [3].

V práci sa budeme zaoberať simultánnymi tolerančnými intervalmi (teda $\boldsymbol{x}$ je ľubovoľné) pre model viacrozmernej lineárnej regresie s normálne rozdelenými nezávislými chybami. Skonštruované sú použitím vektora pozorovaní $\boldsymbol{Y}$ tak, aby obsahovali aspoň γ časť budúcich pozorovaní náhodnej premennej $Y(\boldsymbol{x})$ pre každú hodnotu vysvetľujúcej premennej $\boldsymbol{x}$ simultánne s koeficientom spoľahlivosti $1-\alpha$. Predpísaná úroveň spoľahlivosti sa vzťahuje k neistote odhadu neznámych parametrov regresného modelu $(\boldsymbol{\beta}, \sigma)$ z nezávislých pozorovaní $\boldsymbol{Y}$ a pokrytie sa vzťahuje k neistote rozdelenia budúceho pozorovania $Y(\boldsymbol{x})$.

V sekcii 2 zadefinujeme simultánne tolerančné intervaly pre model viacrozmernej lineárnej regresie s normálne rozdelenými chybami. Na ich konštrukciu treba určit tzv. *tolerančný faktor*, ktorého hodnota pre danú vysvetľujúcu premennú závisí od rozdelenia vektora pozorovaní, požadovanej časti pokrytia a úrovne spoľahlivosti. V sekcii 3 popíšeme známe metódy na stanovenie tolerančného faktora, všetky však prevyšujú požadovanú úroveň spoľahlivosti. V diskusii naznačíme možnú metódu na riešenie.

2. Simultánne tolerančné intervaly v lineárnej regresii

Uvažujeme model viacrozmernej lineárnej regresie s náhodnými normálne rozdelenými nezávislými chybami. Maticový zápis modelu

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \sigma\boldsymbol{Z}, \tag{1}$$

kde $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ predstavuje n-rozmerný náhodný vektor meraných hodnôt, $\boldsymbol{X}$ je $n \times q$ známa matica vysvetľujúcich premenných (jej prvky nemajú náhodný charakter) s hodnosťou q a platí $n > q$. Vektor $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{q-1})^T$ je q-rozmerný vektor regresných parametrov, $\boldsymbol{Z}$ je n-rozmerný vektor štandardných nezávislých chýb, tj. $\boldsymbol{Z} \sim N(\boldsymbol{0}, \boldsymbol{I}_n)$ a σ je smerodajná odchýlka, $\sigma > 0$. Poznamenajme, že jednoduchá lineárna regresia je špeciálnym prípadom modelu (1).

Odhady neznámych parametrov modelu $\boldsymbol{\beta}, \sigma^2$ metódou najmenších štvorcov sú dané

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \quad \text{a} \quad S^2 = \frac{(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})}{n-q}. \tag{2}$$

Platí $(n-q)S^2/\sigma^2 \sim \chi^2_{n-q}$, kde χ^2_{n-q} označuje centrálne chi-kvadrát rozdelenie s $n-q$ stupňami voľnosti. Náhodné premenné $\hat{\boldsymbol{\beta}}$ a S^2 sú nezávislé.

Nech $Y(\boldsymbol{x})$ označuje budúce pozorovanie pre dané $\boldsymbol{x}^T = (1, x_1, \ldots, x_{q-1})^T$, potom

$$Y(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta} + \sigma Z, \tag{3}$$

kde $Z \sim N(0,1)$ a $Y(\boldsymbol{x})$ je nezávislé od $\boldsymbol{Y}$ z modelu (1). Pre pevné $\boldsymbol{x}$, $(\gamma, 1-\alpha)$ obojstranný tolerančný interval pre budúce pozorovanie $Y(\boldsymbol{x})$ uvažujeme v tvare

$$\left\langle \boldsymbol{x}^T\hat{\boldsymbol{\beta}} - \lambda(\boldsymbol{x}|\gamma, 1-\alpha, \boldsymbol{Y}, \boldsymbol{X})S, \boldsymbol{x}^T\hat{\boldsymbol{\beta}} + \lambda(\boldsymbol{x}|\gamma, 1-\alpha, \boldsymbol{Y}, \boldsymbol{X})S \right\rangle, \tag{4}$$

ktorý je symetrický okolo odhadu $\boldsymbol{x}^T\boldsymbol{\beta}$ a jeho šírka je $\lambda(\boldsymbol{x}|\gamma, 1-\alpha, \boldsymbol{Y}, \boldsymbol{X})$ - násobkom výberovej smerodajnej odchýlky S, kde $\lambda(\boldsymbol{x}|\gamma, 1-\alpha, \boldsymbol{Y}, \boldsymbol{X})$ je tzv. tolerančný faktor, ktorý je treba určiť tak, aby bola splnená požiadavka na pokrytú časť γ rozdelenia $Y(\boldsymbol{x})$ so spoľahlivosťou $1-\alpha$. Ďalej budeme pre tolerančný faktor v danom $\boldsymbol{x}$ používať pohodlnejší zápis $\lambda = \lambda(\boldsymbol{x}|\gamma, 1-\alpha, \boldsymbol{Y}, \boldsymbol{X})$.

Nech

$$\boldsymbol{b} = \frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\sigma} \sim N(0, (\boldsymbol{X}^T\boldsymbol{X})^{-1}), \quad \text{a} \quad u = \frac{S}{\sigma}, \quad (n-q)u^2 \sim \chi^2_{n-q}, \tag{5}$$

sú nezávislé náhodné premenné, ktorých rozdelenie nezávisí od neznámych parametrov modelu. Pokrytie tolerančného intervalu (4) $P_{Y(\boldsymbol{x})}(\boldsymbol{x}^T\hat{\boldsymbol{\beta}} - \lambda S \leq Y(\boldsymbol{x}) \leq \boldsymbol{x}^T\hat{\boldsymbol{\beta}} + \lambda S |\ \hat{\boldsymbol{\beta}}, S)$ pri danom $\hat{\boldsymbol{\beta}}, S$ môžeme pomocou pivotných premenných $\boldsymbol{b}, u$ zapísať

$$C(\boldsymbol{x}^T\boldsymbol{b}, \lambda u) = \Phi(\boldsymbol{x}^T\boldsymbol{b} + \lambda u) - \Phi(\boldsymbol{x}^T\boldsymbol{b} - \lambda u), \tag{6}$$

kde Φ označuje distribučnú funkciu štandardného normálneho rozdelenia.

Simultánne obojstranné tolerančné intervaly v lineárnom regresnom modeli s normálne rozdelenými nezávislými chybami hľadáme v tvare (4). Skonštruované sú použitím vektora pozorovaní $\boldsymbol{Y}$ z (1) tak, aby obsahovali aspoň γ časť budúcich pozorovaní náhodnej premennej $Y(\boldsymbol{x})$ zároveň pre všetky $\boldsymbol{x} \in \mathcal{R}^{q\times 1}$ s koeficientom spoľahlivosti $1-\alpha$.

Tolerančný faktor musí spĺňať

$$P_{\boldsymbol{b},u}(C(\boldsymbol{x}^T\boldsymbol{b}, \lambda u) \geq \gamma \quad \forall \boldsymbol{x} \in \mathcal{R}^{q\times 1}) = 1-\alpha. \tag{7}$$

Teda $(1-\alpha)\cdot 100\%$ z tolerančných intervalov skonštruovaných na základe rôznych pozorovaní $\boldsymbol{Y}$ bude obsahovať aspoň γ časť z rozdelenia $Y(\boldsymbol{x})$ pre každé $\boldsymbol{x}$.

Nech G označuje množinu pivotov $\boldsymbol{b}, u$ spĺňajúcich (7), ktorú budeme nazývať $(1-\alpha)$-pivotná množina. Oblasť spoľahlivosti pre parametre modelu môže byť vyjadrená pomocou tejto $(1-\alpha)$-pivotnej množiny

$$\{(\boldsymbol{\beta}, \sigma) = (\hat{\boldsymbol{\beta}} - \boldsymbol{b}S/u, S/u) : \ (\boldsymbol{b}, u) \in G\}. \tag{8}$$

Rovnosť (7) môže byť prepísaná do ekvivalentného tvaru

$$P_{\boldsymbol{b},u}(\min_{\boldsymbol{x}} C(\boldsymbol{x}^T\boldsymbol{b}, \lambda u) \geq \gamma) = 1 - \alpha, \tag{9}$$

z ktorého budeme hľadať vyjadrenie pre tolerančný faktor.

3. Metódy na určenie tolerančného faktora

V kapitole popíšeme doteraz známe metódy na výpočet tolerančných faktorov pre simultánne obojstranné tolerančné intervaly (SOTI) v lineárnom regresnom modeli s normálne rozdelenými nezávislými chybami. Konkrétne Liebermanovu-Millerovu metódu (LM), metódy založené na tzv. confidence-set (CS) prístupe tj. Wilsonovu metódu (W), Limamovu-Thomasovu metódu (LT) a modifikovanú Wilsonovu metódu (MW). V metódach založených na CS prístupe jednotlivý autori zadefinovali tvar $(1-\alpha)$-pivotnej oblasti G a tolerančný faktor potom počítali

$$\lambda = \min\{\lambda : C(\boldsymbol{x}^T\boldsymbol{b}, \lambda u) \geq \gamma \ \text{ for all } \ (\boldsymbol{b}, u) \in G\}. \tag{10}$$

3.1. Liebermanova-Millerova metóda

Lieberman a Miller [5] prezentovali simultánne tolerančné intervaly pre prípad jednoduchej lineárnej regresie, špeciálny prípad modelu (1), kedy $q = 2$, $\boldsymbol{X}_1 = (x_1, \ldots, x_n)^T$, $\boldsymbol{x} = (1, \ x)^T$ pre $x \in \mathcal{R}$, bez straty na všeobecnosti uvažovali $\sum x_i/n = 0$, potom $d(x) = \sqrt{\boldsymbol{x}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}} = \sqrt{1/n + x^2/\sum_i x_i^2}$. Tolerančný faktor vyjadrili v tvare

$$\lambda = \lambda^* \cdot d(x). \tag{11}$$

Označme $C^*(b_0, b_1, u) = \min_x C(b_0 + b_1 x, \lambda^* d(x)u)$, potom vzťah (9) pre $\boldsymbol{b} = (b_0, b_1)$ možeme prepísať na tvar

$$E_{\boldsymbol{b}}[P_u\{C^*(b_0, b_1, u) \geq \gamma | b_0, b_1\}] = 1 - \alpha, \tag{12}$$

pričom po vyjadrení výrazu v strednej hodnote do Taylorovho radu v $b_0 = 0$, $b_1 = 0$ odvodili aproximáciu

$$E_{\boldsymbol{b}}[P_u\{C^*(b_0, b_1, S) \geq \gamma | b_0, b_1\}] \approx$$
$$P_u\{C^*(b_0, b_1, S) \geq \gamma | b_0 = 1/\sqrt{n}, b_1 = \sqrt{1/\textstyle\sum_i x_i^2}\}.$$

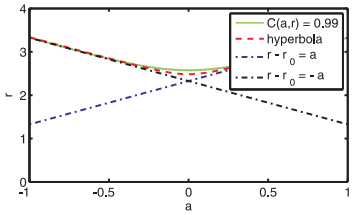
Funkcia $\min_x C(1/\sqrt{n} + x/\sqrt{\sum x_i^2}, hd(x))$ je neklesajúcou funkciou v h, preto existuje konštanta h_0, ktorá spĺňa

$$\min_x C(1/\sqrt{n} + x/\sqrt{\textstyle\sum_i x_i^2}, h_0 d(x)) = \gamma,$$

pričom minimum sa dosahuje v x^* spĺňajúcom

$$(1/\sqrt{\textstyle\sum_i x_i^2} + hd(x)^{-1}x/\textstyle\sum_i x_i^2) f_{N(0,1)}(1/\sqrt{n} + x/\sqrt{\textstyle\sum_i x_i^2} + hd(x))$$
$$-(1/\sqrt{\textstyle\sum_i x_i^2} - hd(x)^{-1}x/\textstyle\sum_i x_i^2) f_{N(0,1)}(1/\sqrt{n} + x/\sqrt{\textstyle\sum_i x_i^2} - hd(x)) = 0,$$

kde $f_{N(0,1)}$ je hustota štandardného normálneho rozdelenia.



OBRÁZOK 1. Ilustrácia ohraničenia krivky $C(a,r) = 0.99$ hornou časťou hyperboly $(r-r_0)^2-a^2 = h^2$, kde $h^2 = 0.0244$ a asymptot $r - r_0 = \pm a$ $(r \geq 0)$ s $r_0 = \Phi^{-1}(0.99)$.

Pre $\lambda^* u > h_0$ platí, že $\min_x C(1/\sqrt{n}+x/\sqrt{\sum_i x_i^2}, \lambda^* d(x)u) > \gamma$, neznámu konštantu λ^* určíme z rovnosti

$$P(u > h_0/\lambda^*) = 1 - \alpha. \tag{13}$$

Vzhľadom na rozdelenie $(n-q)u^2 \sim \chi^2_{n-q}$ dostávame

$$\lambda^* = h_0\sqrt{(n-q)/\chi^2_{n-q}(\alpha)}, \tag{14}$$

kde h_0 je numericky dopočítané vyššie uvedeným postupom.

3.2. **Wilsonova metóda**

Wilson [10] zadefinoval tvar pivotnej oblasti ozn. G_W ako $(q+1)$-rozmerný elipsoid. Na jeho konštrukciu využil aproximáciu $(2\chi^2_{n-q})^{1/2} \sim N([2(n-q)-1]^{1/2}, 1)$ [1], z ktorej určil približné rozdelenie pivota u, platí $(n-q)u^2 \sim \chi^2_{n-q}$. Nech $v = n-q$, potom $u \sim N(k, 1/(2v))$, kde $k = \sqrt{(2v-1)/(2v)}$ a platí $\boldsymbol{b}^T(\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{b} \sim \chi^2_q$. Wilsonova oblasť spoľahlivosti odpovedá pivotnej množine s približnou $1-\alpha$ spoľahlivosťou tvaru

$$G_W = \{(\boldsymbol{b}, u) : \boldsymbol{b}^T(\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{b} + 2v\ (u\ -k)^2 \leq c\}, \tag{15}$$

kde $c = \chi^2_{q+1}(1-\alpha)$.

Matica $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ je kladne definitná a symetrická, potom na základe Schwarzovej nerovnosti pre ľubovoľný vektor $\boldsymbol{b}$ platí

$$\max_{\boldsymbol{x}} \frac{(\boldsymbol{x}^T\boldsymbol{b})^2}{\boldsymbol{x}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}} = \boldsymbol{b}^T(\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{b}. \tag{16}$$

Označme $\delta(\boldsymbol{x}) = \sqrt{\boldsymbol{x}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}}$, teda $(\boldsymbol{x}^T\boldsymbol{b})^2 \leq \boldsymbol{b}^T(\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{b}\ \delta^2(\boldsymbol{x})$ pre každé $\boldsymbol{x}$ a spolu so vzťahom (15) platí

$$|\boldsymbol{x}^T\boldsymbol{b}| \leq \sqrt{c - 2v\ (u\ -k)^2}\delta(\boldsymbol{x}) \quad \text{pre každé} \quad \boldsymbol{x}. \tag{17}$$

Funkcia $A_{\boldsymbol{x}}(u) = \sqrt{c - 2v(u-k)^2}\delta(\boldsymbol{x})$ je definovaná na intervale $u \in [k - \sqrt{c/2v}, k + \sqrt{c/2v}]$. Wilson ukázal, že ak $a = \boldsymbol{x}^T\boldsymbol{b}$ a $r = \lambda u$ potom

$$G_W \subset H(\boldsymbol{x}, \lambda) = \{(a,r) : a^2/\delta^2(\boldsymbol{x}) + 2v\ (r/\lambda - k)^2 \leq c\} \tag{18}$$

pre každé $\boldsymbol{x}$ a $\lambda > 0$.

Pre pevné γ Wilson navrhol ohraničiť množinu $S_\gamma = \{(a,r) : C(a,r) = \Phi(a+r)-\Phi(a-r) \geq \gamma\}$ definovanú v R^2 hornou hranicou hyperboly $(r-r_0)^2 - a^2 = h^2$, kde $r_0 = \Phi^{-1}(\gamma)$ a hodnoty h^2 získal aproximačne pre vybrané hodnoty γ, napr. pre $\gamma = 0.99$ je $h^2 = 0.0244$. Optimálny tolerančný faktor založený na pivotnej množine G_W leží na prieniku hyperboly a množiny $H(\boldsymbol{x}, \lambda)$, je to najmenšie λ také, že žiaden bod $H(\boldsymbol{x}, \lambda)$ neleží pod hyperbolou $(r-r_0)^2 - a^2 = h^2$ pre dané α, γ. Dosadením $a^2 = (r-r_0)^2 - h^2$ do (18) a nahradením znamienka nerovnosti znamienkom rovnosti získal kvadratickú rovnicu v r

$$(r - r_0)^2 - h^2 - [c - 2v\ (r/\lambda - k)^2]\delta^2(\boldsymbol{x}) = 0. \tag{19}$$

Diskriminant je kvadratickou funkciou v λ, ktorá má tvar

$$(4\delta^2(\boldsymbol{x})c - 8\delta^2(\boldsymbol{x})k^2v + 4h^2)\lambda^2 + 16r_0v\ \delta(\boldsymbol{x})^2k\ \lambda + 8v\ \delta(\boldsymbol{x})^4c + 8v\ \delta(\boldsymbol{x})^2h^2 - 8v\ \delta^2(\boldsymbol{x})r_0^2 = 0 \tag{20}$$

Riešenie tejto kvadratickej rovnice je

$$\lambda_{1,2} = \frac{-16r_0v\ \delta^2(\boldsymbol{x})k\ \pm\sqrt{D}}{2(4\delta^2(\boldsymbol{x})c - 8\delta^2(\boldsymbol{x})k^2v\ + 4h^2)}, \tag{21}$$

kde $D = 128h^2v\delta^2(\boldsymbol{x})r_0^2 - 128h^4v\delta^2(\boldsymbol{x}) - 128h^2v\delta^4(\boldsymbol{x})c + 128\delta^4(\boldsymbol{x})cvr_0^2 - 128\delta^6(\boldsymbol{x})c^2v + 256\delta^4(\boldsymbol{x})k^2v^2h^2 + 256\delta^6(\boldsymbol{x})k^2v^2c$. Oba korene sú reálne a hľadaný tolerančný faktor je väčší z nich.

3.3. **Limamova-Thomasova metóda**

Limam a Thomas odvodili $(1-\alpha)$-pivotnú množinu ozn. G_{LT} z množinového súčinu $(1-\alpha/2)$-konfidenčných oblastí pre regresné parametre $\boldsymbol{\beta}$ a σ. Pre parameter $\boldsymbol{\beta}$ použili oblasť spoľahlivosti v tvare q-rozmerného elipsoidu a pre neznámy parameter modelu σ zhora ohraničený interval, potom

$$G_{LT} = \{(\boldsymbol{b}, u) : \boldsymbol{b}(\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{b} \leq u^2k_1^2 \quad \text{a} \quad u\ \geq k_2\}, \tag{22}$$

kde $k_1^2 = qF_{q,n-q}(1-\alpha/2)$ a $F_{q,n-q}(1-\alpha/2)$ je $(1-\alpha/2)$-kvantil F-rozdelenia so stupňami voľnosti q a $n-q$ a $k_2 = \sqrt{\chi^2_{n-q}(\alpha/2)/(n-q)}$. Na základe Bonferroniho nerovnosti platí $P((\boldsymbol{b}, u) \in G_{LT}) \geq 1-\alpha$.

Podobne ako Wilson využili Scheffého výsledok (16) na ohraničenie lineárnej kombinácie

$$|\boldsymbol{x}^T\boldsymbol{b}| \leq u\ k_1\delta(\boldsymbol{x}) \quad \text{pre každé} \quad \boldsymbol{x}, \quad (\boldsymbol{b}, u) \in G_{LT}. \tag{23}$$

Pokrytie definované (6) je párna funkcia v $\boldsymbol{x}$ a z priebehu normálneho rozdelenia je zrejme, že je klesajúca pre $|\boldsymbol{x}^T\boldsymbol{b}|$ pri pevnom λu, teda platí

$$C(\boldsymbol{x}^T\boldsymbol{b}, \lambda u) \geq C(uk_1\delta(\boldsymbol{x}), \lambda u) \quad \text{pre každé} \quad (\boldsymbol{b}, u) \in G_{LT}. \tag{24}$$

Funkcia $C(uk_1\delta(\boldsymbol{x}), \lambda u)$ je rastúca v u, keď interval $[k_1\delta(\boldsymbol{x}) - \lambda, k_1\delta(\boldsymbol{x}) + \lambda]$ obsahuje nulu. Podmienka je splnená, ak uvažujeme

$$C(uk_1\delta(\boldsymbol{x}), \lambda u) = \Phi[u(k_1\delta(\boldsymbol{x}) + \lambda)] - \Phi[u(k_1\delta(\boldsymbol{x}) - \lambda)] \geq 1/2. \tag{25}$$

Potom

$$C(uk_1\delta(\boldsymbol{x}), \lambda u) \geq C(k_2k_1\delta(\boldsymbol{x}), k_2\lambda) \quad \text{pre} \quad u \geq k_2. \tag{26}$$

Z úvahy vyplýva ohraničenie pre pokrytie $\gamma \geq 1/2$, čo je však vo väčšine aplikácií akceptovateľné.

Tolerančný faktor λ je vyjadrený v tvare

$$\lambda = \frac{r_\gamma[k_2k_1\delta(\boldsymbol{x})]}{k_2}, \tag{27}$$

kde $r = r_\gamma(a)$ je koreň rovnice $C(a, r) = \Phi(a + r) - \Phi(a - r) = \gamma$. Pre numerický výpočet je stanovený bod z hyperboly $r^0_\gamma(a) = \Phi^{-1}(\gamma)+\{(\Phi^{-1}[(\gamma+1)/2] - \Phi^{-1}[\gamma])^2 + a^2\}^{1/2}$ ako štartovacia hodnota pre dané γ.

3.4. **Modifikovaná Wilsonova metóda**

Výraz na ľavej strane nerovnosti v (15) nemá kvôli použitej aproximácií pre rozdelenie pivota u chi-kvadrát rozdelenie s $q+1$ stupňami voľnosti, preto hodnota konštanty $c = \chi^2_{q+1}(1-\alpha)$ na pravej strane nerovnosti ohraničuje len približne $(1-\alpha)$-pivotnú množinu. Limam a Thomas ohraničili Wilsonov elipsoid spoľahlivosti upravenou hodnotou c, ozn. c_m, tak aby platilo $P((\boldsymbol{b}, u) \in G_W) = 1 - \alpha$. Modifikovaná konštanta c_m je menšia, čo má za následok zmenšenie hodnoty tolerančného faktora λ. Na výpočet upravnej hodnoty c_m použili výsledok (17) a z vlastnosti funkcie pokrytie vyplýva $C(\boldsymbol{x}^T\boldsymbol{b}, \lambda u) \geq C(A_{\boldsymbol{x}}(u), \lambda u)$ pre $(\boldsymbol{b}, u) \in G_W$. Funkcia $A_{\boldsymbol{x}}(u)$ klesá a λu rastie na intervale $u \in [k, k + \sqrt{c/2v}]$, preto $C(A_{\boldsymbol{x}}(u), \lambda u)$ ako funkcia v u je rastúca na tomto intervale. Na určenie tolerančného faktora λ stačí uvažovať len podmnožinu G_W odpovedajúcu intervalu $[k - \sqrt{c/2v}, k]$. Limam a Thomas zadefinovali pivotnú oblasť $G_{MW} = G_{MW1} \cup G_{MW2}$, kde G_{MW1} je tvaru (15) pre $u \in [k - \sqrt{c_m/2v}, k]$ a oblasť G_{MW2} je rovnakého tvaru ako pivotná oblasť (22) skonštruovaná tak, aby mala priesečník s G_{MW1} v $u = k$

$$G_{MW2} = \{(\boldsymbol{b}, u) : \boldsymbol{b}^T(\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{b} \leq u^2c_m/k^2 \ \text{ a } \ u \geq k\}. \tag{28}$$

Ak $c = c_m$, platí $G_W \subset G_{MW}$ a teda $P(G_W) < P(G_{MW})$. Potom $P(G_{MW}) = P(G_W)$ implikuje, že $c_m < c$.

Koeficienty c_m sú dopočítané iteračným postupom ako riešenie rovnice $P(G_{MW}) = 1-\alpha$ na dosiahnutie požadovanej úrovne spoľahlivosti. Pre $q = 2$, $n = 15$ konštanta $c_m = 9.656$, ak $\alpha = 0.01$ a $c_m = 6.432$, ak $\alpha = 0.05$. Tolerančný faktor sa vypočíta použitím postupu z Wilsonovej metódy (21) s modifikovanou hodnotou c_m.

4. **Diskusia**

V sekcii 3 sme popísali známe metódy na konštrukciu simultánnych obojstranných tolerančných intervalov. Tolerančné faktory, potrebné na stanovenie tolerančných intervalov, dopočítané uvedenými metódami vedú pre použité známe vzťahy, Bonferroniho nerovnosť, Scheffého výsledok založený na

Schwarzovej nerovnosti a viaceré aproximácie (Fisherovu pre chi-kvarát rozdelenie atď.) pri ich odvodení, len k približným SOTI v lineárnom regresnom modeli s normálnymi chybami. Tolerančné faktory, ktorým je úmerná šírka intervalov, sú počítané pre každý bod regresnej krivky a ich hodnota okrem daného pokrytia γ a úrovne spoľahlivosti $1-\alpha$ závisí aj od matice plánu $\boldsymbol{X}$ a teda všeobecne rozhodnúť, ktorú metódu použiť pre konkrétny problém je predmetom štúdia. Pri kalibrácii meracích zariadení v prípade opakovaného dopredu neznámeho počtu meraní na zariadení sa uvažuje s ohraničenou množinou možných vysvetľujúcich premenných a v tomto prípade najužšie SOTI dosiahli Mee a kol. v [7].

Presný test pomerom vierohodnosti pre testovanie nulovej hypotézy $H_0 : (\boldsymbol{\beta}, \sigma) = (\beta_0, \sigma_0)$ proti alternatíve $H_1 : (\boldsymbol{\beta}, \sigma) \neq (\boldsymbol{\beta}_0, \sigma_0)$ môže byť využitý na definovanie oblasti spoľahlivosti pre všetky parametre regresného modelu zároveň. Tvar presnej $(1-\alpha)$-oblasti spoľahlivosti je daný

$$\mathcal{C}_{1-\alpha}(\boldsymbol{Y} \mid \boldsymbol{X}) = \{(\boldsymbol{\beta}, \sigma) : \lambda(\boldsymbol{Y} \mid \boldsymbol{X}) \leq \lambda_{1-\alpha}\}\,, \tag{29}$$

kde $\lambda(\boldsymbol{Y} \mid \boldsymbol{X})$ je testovacia štatistika testu pomerom vierohodnosti, ktorej rozdelenie závisí od počtu pozorovaní a od počtu komponentov vektora $\boldsymbol{\beta}$. Kritické hodnoty $\lambda_{1-\alpha}$ pre test pomerom vierohodnosti sú uvedené v tabuľkách priložených v [2] pre rôzne počty komponentov vysvetľujúcej premennej $q = 1, \ldots, 10$, pre vybrané počty pozorovaných meraní $n = q+1 : (1) : 40, n = 45 : (5) : 100$ a ∞ a pre zvyčajné hladiny významnosti $\alpha = \{0.1, 0.05, 0.01\}$.

Literatúra

[1] Fisher R.A. (1928) *Statistical Methods for Research Workers*. 2nd Edition, 96 – 97.
[2] Chvosteková M., Witkovský V. (2009) *Exact Likelihood Ratio Test for the Parameters of the Linear Regression Model with Normal Errors*. Measurement Science Review **1**, 9, 1 – 8.
[3] Krishnamoorthy K., Mathew T. (2009) *Statistical Tolerance Regions: Theory, Applications, and Computation*, Wiley.
[4] Lieberman G.J. (1961) *Prediction Regions for Several Predictions from a Single Regression Line*. Technometrics **1**, 3, 21 – 27.
[5] Lieberman G.J., Miller R.G., Jr. (1963) *Simultaneous Tolerance Intervals in Regression*. Biometrika **1/2**, 50, 155 – 168.
[6] Limam M.M.T., Thomas, R. (1988) *Simultaneous Tolerance Intervals for the Linear Regression Model*. Journal of the American Statistical Association **403**, 83, 801 – 804.
[7] Mee R.W., Eberhardt K.R., Reeve C.P. (1991) *Calibration and Simultaneous Tolerance Intervals for Regression*. Technometrics **2**, 33, 211 – 219.
[8] Rao C.R. (1979) *Lineární statistické metody a jejich aplikace*, Academia, Praha.
[9] Scheffé H. (1973) *A Statistical Theory of Calibration*. The Annals of Statistics **1**, 1, 1 – 37.
[10] Wilson A.L. (1967) *An Approach to Simultaneous Tolerance Intervals in Regression*. The Annals of Mathematical Statistics **38**, 1536 – 1540.

Poďakovanie: Práca bola podporená VEGA grantmi 1/0077/09, 2/0019/10 a APVV grantom SK-AT-0003-08.

Adresa: Ústav merania SAV, Dúbravská cesta 9, 841 04 Bratislava
E-mail: chvosta@gmail.com

INTERLABORATORY COMPARISON UNDER HETEROSCEDASTIC ANOVA MODEL FOR THE OBSERVED DATA

Mária Janková

Keywords: Interlaboratory comparison, common mean, heteroscedastic ANOVA model, confidence interval, metrological approach, generalized pivotal approach.

Abstract: In metrology, one frequently encounters the so-called common mean problem. In practice, the aim is to find the most exact estimate of the true value of measured physical quantity, where this estimate is often called the key comparison reference value (KCRV). For this assessment data are available from several laboratories.

The article deals with interval estimators of the common mean. Heteroscedastic ANOVA model is considered for the data, where a single measurement error consists of a so-called laboratory error, which is the same for all observations from a single laboratory, and from a so-called measurement error. Two methods of interval estimation are compared: method based on metrological approach proposed by Witkovský and Wimmer in [4] and generalized confidence intervals (GCI) proposed by Wang and Iyer in [2]. The results are also compared for normal and uniform distribution of the laboratory error.

Abstrakt: V metrológii sa často stretávame s problémom stanovenia spoločnej strednej hodnoty. V praxi ide o stanovenie čo najpresnejšieho odhadu skutočnej hodnoty meranej veličiny, pričom tento odhad sa nazýva kľúčová porovnávacia referenčná hodnota (KCRV - key comparison reference value). Pre jej určenie sú k dispozícií dáta z viacerých laboratórií.

V tomto príspevku sa budeme zaoberať intervalovými odhadmi spoločnej strednej hodnoty. Pre dáta budeme uvažovať heteroskedastický ANOVA model, pričom chyba každého pozorovania pozostáva z tzv. laboratórnej chyby, ktorá je pre všetky pozorovania z jedného laboratória rovnaká, a z chyby jednotlivých meraní. Porovnáme dve metódy intervalového odhadu: metódu založenú na metrologickom prístupe navrhnutú Witkovským a Wimmerom v [4] a zovšeobecnené intervaly (GCI - generalized confidence intervals) navrhnuté Wangom a Iyerom v [2]. Tiež porovnáme výsledky pre normálne a rovnomerné rozdelenie laboratórnej chyby.

1. Introduction

Consider a situation where multiple measurements of one (identical) physical quantity are performed by two or more laboratories. To determine the true value of the measurand, the provided measurements from each laboratory should be combined in an appropriate way, so that the resulting estimate

is a sufficient approximation. In case the laboratories provide the sample mean and sample standard deviation as output, the result of Graybill and Deal (1959) is significant. In [1] Graybill and Deal proved that under particular conditions on the number of measurements provided by each laboratory, the estimate constructed as weighted sum of sample means, where weights are inversely proportional to sample deviations and proportional to sample sizes, disposes of smaller variance than any of the single estimates itself. Graybill and Deal deal with a simple model where random samples are drawn from normal distributions with same mean and possibly different variances.

In this article, we will consider a more complex model. In particular, the model discussed will be one - way heteroscedastic ANOVA model. Formally, the considered model can be represented as follows:

$$Y_{ij} = \mu + b_i + \varepsilon_{ij}, \tag{1}$$

for $i = 1, \dots k$ representing the number of laboratories and $j = 1, \dots n_i$ representing the number of measurements by the $i-th$ laboratory. Using metrological concepts each measurement Y_{ij} of the true value of measurand μ is biased by the measurement error ε_{ij} and by characteristic laboratory error b_i. Distribution of random variable ε_{ij} under this model is $N(0, \sigma_{A,i})$, $\sigma_{A,i}$ unknown. Distribution of b_i is fully known, as well as the mean β_i and variance $\sigma_{B,i}$. The resulting task of estimating μ under this model is known as the common mean problem; it can also be referred to as the problem of finding the key comparison reference value.

Various methods have been considered for estimating μ from model (1), of which we will more closely look at interval estimation, particularly at the approaches proposed by Wang and Iyer [2] and Witkovský and Wimmer [4]. The Wang and Iyer approach utilizes the general pivotal quantities. More on introduction of generalized confidence intervals can be found in Weerahandi [3]. Article [2] provides construction details of the confidence interval for the common mean μ, yet the frequentist properties study is left out. Witkovský and Wimmer approach is a specific approach described in more detail in [4], based on a partially Bayesian approach. In [4] a simulation study is carried out, considering different values of all input parameters, including different distributions of b_i, in order to gain the empirical coverage probabilities of confidence intervals constructed by this method.

We will compare the frequentist properties of both methods with respect to the length of the intervals, as well as the empirical coverage probability property. We will also analyze the parameter change sensitivity of the resulting length of the confidence interval provided by each method. Moreover, we will look at the differences in performance of the two methods when different distributions of b_i are considered. For the purpose of this article, we have chosen to compare patterns with normally distributed b_i and uniformly distributed b_i.

2. Compared methods

Denote the sample mean $\bar{Y}_i = \frac{1}{n}\sum_{j=1}^{n_i} Y_{ij}$ and sample standard deviation $S_i^2 = \frac{1}{n_i-1}\sum_{i=1}^{n_i}(Y_{ij}-\bar{Y}_i)^2$ and their realizations $\bar{y}_i$ and s_i^2. Notice that both methods assign weights to sample means or involved random variables. While in Witkovský and Wimmer approach, after the data Y_{ij} have been collected, the weights are deterministic, the weights in Wang and Iyer approach are stochastic.

2.1. Witkovský and Wimmer approach

Metrological approach proposed in [4] (further on referred to as WW approach) constructs the confidence intervals as follows. Consider random variable $\tilde{\tilde{\mu}}$ given by:

$$\tilde{\tilde{\mu}} = \sum_{i=1}^{k} w_i \bar{y}_i - \sum_{i=1}^{k} w_i \sqrt{\frac{s_i^2}{n_i}} T_i - \sum_{i=1}^{k} w_i B_i,$$

where $T_i \sim t_{n_i-1}$ and w_i are chosen in the following way:

$$w_i = \frac{\left(1/\left(\sqrt{\frac{s_i^2}{n_i}}\sqrt{\frac{s_p^2}{n_i}\frac{n_i-1}{n_i-3}} + \sigma_{(B),i}^2\right)\right)}{\sum_{l=1}^{k}\left(1/\left(\sqrt{\frac{s_l^2}{n_l}}\sqrt{\frac{s_p^2}{n_l}\frac{n_l-1}{n_l-3}} + \sigma_{(B),l}^2\right)\right)},$$

where $s_p^2 = \sum_{i=1}^{k}(n_i-1)s_i^2 / \sum_{i=1}^{k}(n_i-k)$. Then we take $(\mu_{KCRV} + q_{\alpha/2}, \mu_{KCRV} + q_{1-\alpha/2})$, where μ_{KCRV} is the mean value of random variable $\tilde{\tilde{\mu}}$ and q_β is $\beta\%$ quantile of random variable $\tilde{\tilde{\mu}}$, as the estimate of $(1-\alpha)\times 100\%$ confidence interval for μ.

2.2. Wang and Iyer approach

Generalized confidence intervals proposed by Wang and Iyer (further on referred to as WI confidence intervals) can be constructed as follows. As the lower and upper boundary of $(1-\alpha)\times 100\%$ confidence interval we take $q_{\alpha/2}$ and $q_{1-\alpha/2}$ quantiles of random variable R_μ, where R_μ is given by:

$$R_\mu = \frac{\sum_{i=1}^{k}(\bar{y}_i - B_i)n_i W_i/[(n_i-1)s_i^2]}{\sum_{i=1}^{k} n_i W_i/[(n_i-1)s_i^2]} - Z\sqrt{\frac{1}{\sum_{i=1}^{k} n_i W_i/[(n_i-1)s_i^2]}},$$

where $W_i \sim \chi^2_{n_i-1}$ and $Z \sim N(0,1)$. Here, the previously mentioned stochastic weights u_i are represented by $u_i = \frac{n_i W_i/[(n_i-1)s_i^2]}{\sum_{i=1}^{k} n_i W_i/[(n_i-1)s_i^2]}$.

3. Methodology of comparison

We compared the empirical coverage probabilities and relative lengths of intervals gained by each method. This was done on the basis of data artificially generated from model (1), for different parameter combinations of model (1).

For each of these different designs we generated 10000 confidence intervals. The empirical coverage probability was computed as number of times the constructed confidence interval covered the true value of μ. Without loss of generality we set μ=0.

The relative lengths of confidence intervals were computed as ratio of the length of the interval constructed by either method to length of a reference confidence interval. This reference confidence interval was constructed under the assumption that all model parameters are known. The reference confidence interval was constructed using the generalized least squares estimator of μ. In case of normal distribution of b_i, the generalized least squares estimator is the MVUE. Exploiting the property of normality of this estimator, we construct exact $(1-\alpha)\times 100\%$ confidence interval for μ. When uniform distribution of b_i is considered, the distribution of generalized least squares estimator is a weighted sum of uniform and normal distributions, quantiles of which can be computed using relevant packages, e.g. $t-dist$ package in Matlab.

4. **Parameter selection**

Testing is performed at significance level $\alpha = 0.05$. Distribution of laboratory error b_i is either $N(0,\sigma^2_{b,i})$ or $U(-\sqrt{3}\sigma_{B,i},\sqrt{3}\sigma_{B,i})$. The number of participating laboratories is either $5, 10$ or 15, i.e. $k \in \{5,10,15\}$. As for number of observations in i^{th} laboratory, $n_i = 5$, $n_i = 10$, $n_i = 15$ or $\sigma_{B,i} \in \{1,2,3,4,5,1,2,3,4,5,1,2,3,4,5\}, i = 1,\dots k$. Designs chosen for $\sigma_{B,i}$ are denoted subsequently: $\sigma_{B,i} = 1$ denoted a, $\sigma_{B,i} = 5$ denoted b, $\sigma_{B,i} \in \{1,2,3,4,5,1,2,3,4,5,1,2,3,4,5\}, i = 1,\dots k$ denoted c, $\sigma_{B,i} = 0$ denoted d. There were three different designs of $\sigma_{A,i}$ chosen: $\sigma_{A,i} = 1$,$\sigma_{A,i} = 5$,$\sigma_{B,i} \in \{1,2,3,4,5,1,2,3,4,5,1,2,3,4,5\}, i = 1,\dots k$.

In graphical representation of simulation results (Figure 1 and 2), we will not focus on the difference between single $\sigma_{A,i}$ choice, but difference in performance of WI and WW method. Following this aim, results computed for WI are denoted by an empty circle ($\circ$), for WW by a small black sphere ($\bullet$).

5. **Results. Parameter change sensitivity analysis**

In case of normal distribution of b_i we can see that WW outperforms WI method in both empirical coverage probability and the length of the confidence interval. Let us remark, that the dominance of WW over WI method holds for small number of observations provided by participating laboratories, particularly up to number 15, with increased number of observations the differences gradually vanish.

Figures representing the performance of methods for uniform distribution of b_i are not given due to minor differences in empirical coverage probabilities and relative lengths of confidence intervals. Let r_{length} be the measure of difference between the lengths of confidence intervals for designs with

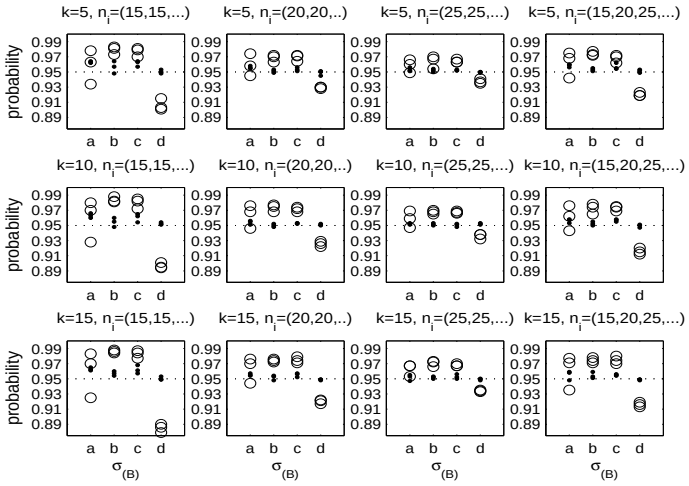


FIGURE 1. Empirical coverage probabilities of $(1-\alpha)\times 100\%$ confidence interval estimates for μ, where $\alpha = 0.05$. Comparison of WW ($\bullet$) and WI ($\circ$) method for $b_i \sim N(0, \sigma_{B,i})$.

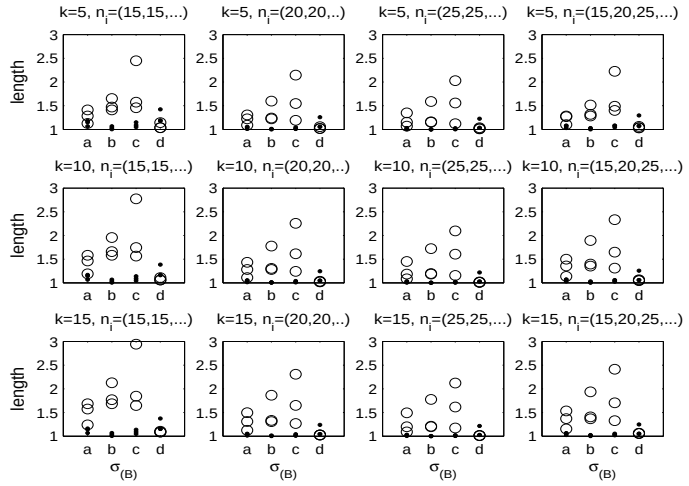


FIGURE 2. Relative lengths of $(1-\alpha)\times 100\%$ confidence interval estimates for μ, where $\alpha = 0.05$. Comparison of WW ($\bullet$) and WI ($\circ$) method for $b_i \sim N(0, \sigma_{B,i})$.

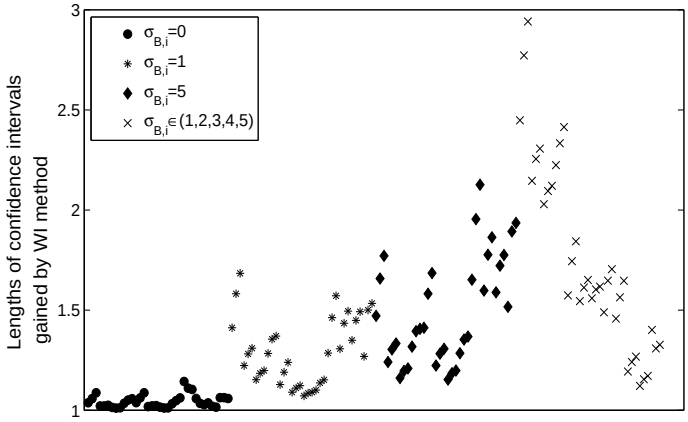


FIGURE 3. Dependence of relative lengths of confidence intervals for WI method on different combination of input parameters ordered by $\sigma_{B,i}$. $b_i \sim N(0, \sigma^2_{B,i})$.

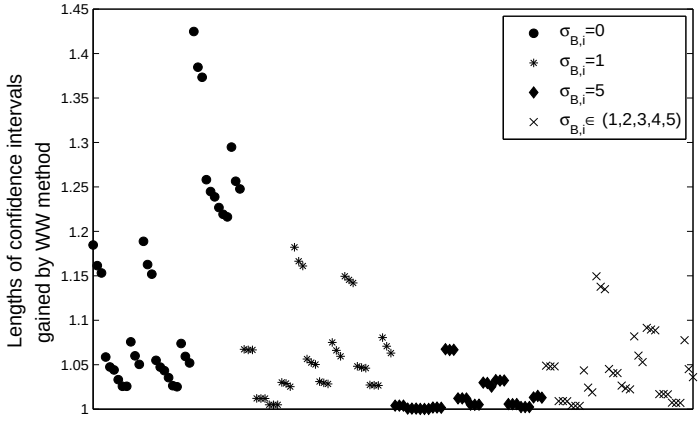


FIGURE 4. Dependence of relative lengths of confidence intervals for WW method on different combination of input parameters ordered by $\sigma_{B,i}$. $b_i \sim N(0, \sigma^2_{B,i})$.

$b_i \sim N(0, \sigma^2_{B,i})$ and lengths for designs with $b_i \sim U(-\sqrt{3}\sigma_{B,i}, \sqrt{3}\sigma_{B,i})$, such that r_{length} is the absolute value of maximum over differences between the lengths of confidence intervals for designs with $b_i \sim U(-\sqrt{3}\sigma_{B,i}, \sqrt{3}\sigma_{B,i})$ and $b_i \sim N(0, \sigma^2_{B,i})$ when all other parameters are identical. Then $r_{length,WW} = 0.017$ and $r_{length,WI} = 0.14$. Similarly, let $r_{coverage}$ be the absolute value of maximum over difference of empirical coverage probabilities with $b_i \sim U(-\sqrt{3}\sigma_{B,i}, \sqrt{3}\sigma_{B,i})$ and $b_i \sim N(0, \sigma^2_{B,i})$ when all other parameters are identical. So defined measure of difference between empirical coverage probabilities of designs with normal and uniform distribution gives the following numerical results: $r_{coverage,WW} = 0.079$ and $r_{coverage,WI} = 0.098$. Thus, the distribution of b_i had small impact on the results in our simulation study.

Figures 1-2 suggest parameter change sensitivity of the average relative length of confidence intervals and empirical coverage probability. This is mostly apparent for the dependence of relative lengths of confidence intervals on the choice of parameter $\sigma_{B,i}$. Graphical representation of this dependence is given in Figure 3 for WI method, in Figure 4 for WW method. Within groups (these groups defined by different $\sigma_{B,i}$ combination) the data are ordered first by number of observation, then by $\sigma_{A,i}$. Average length of confidence intervals gained by WW method is the shortest for designs with $\sigma_{B,i} \in \{1, 2, 3, 4, 5\}$ and the longest for $\sigma_{B,i} = 0$. For the WI method the worst results are provided for different $\sigma_{B,i}$, best performance is achieved when the laboratory error is not present in the model. The figures hint that WW method better copes with bigger values of the laboratory error than with smaller values and that WI method is more adequate for smaller values.

References

[1] Graybill F.A., Deal R.B. (1959) *Combining unbiased estimators.* Biometrics, **15**, 4, 543 – 550.

[2] Wang C.M., Iyer H.K.(2006) *A generalized confidence interval for a measurand in the presence of type-A and type-B uncertainties.* Measurement, **39**, 9, 856 – 863.

[3] Weerahandi S. (1993) *Generalized confidence intervals.* Journal of the American Statistical Association, **88**, 899 – 905.

[4] Witkovský V., Wimmer G. (2007) *Confidence interval for common mean in interlaboratory comparisons with systematic laboratory biases.* Measurement Science Review, **7**, Section 1, No. 6.

Acknowledgement: This work was supported by grants VEGA 1/0077/09, VEGA 2/0019/10 and APVV grant SK-AT-0003-07.

Address: Institute of Measurement Science, Slovak Academy of Sciences, Dúbravská cesta 9, 841 04, Bratislava, Slovakia

E-mail: majka.jankova@gmail.com

JOSEPH BERTRAND

Anna Kalousová

Klíčová slova: Joseph Bertrand, pedagogická činnost, teorie pravděpodobnosti, geometrická pravděpodobnost.

Abstrakt: Joseph Bertrand patří k nejznámějším francouzským matematikům 19. století. V článku připomeneme jeho život, pedagogickou činnost a vědecké práce. Zaměříme se na práce o teorii pravděpodobnosti, zeména na části týkající se pravděpodobnosti geometrické.

Abstract: Joseph Bertrand is ranked among the best known French mathematicians of the 19th century. In the article, we recall his life, pedagogical activities and scientific work. We concentrate on his work on probability theory, especially, on the parts concerning the geometrical probability.

1. Úvod

Výuku pravděpodobnosti na začátku 20. století silně ovlivnila kniha o pravděpodobnostním počtu [8] francouzského matematika Josepha Bertranda. Ačkoli byla mnohými význačnými matematiky (Darboux, Poincaré, Borel) vysoce ceněna, byla po válce kritizována kvůli příliš literárnímu stylu a značnému omezování role matematické analýzy.

Teorie pravděpodobnosti byla pro Bertranda nejoblíbenější částí matematiky. Uvědomoval si, jak je důležité, aby základy teorie pravděpodobnosti byly pochopitelné i lidem, kteří nejsou dostatečně matematicky vzděláni. O to se snažil i v této knize. V úvodu píše, že se všichni shodují na tom, že nelze rozumět pravděpodobnostnímu počtu bez přečtení Laplaceovy knihy [13] a že Laplaceovu knihu nelze číst bez hlubokého studia matematiky. Bertrand se naproti tomu snaží používat jazyk, který bude srozumitelnýný všem.

V tomto článku se budeme věnovat životu Josepha Bertranda a jeho dílu. Podrobný rozbor jeho díla z oblasti teorie pravděpodobnosti lze nalézt v [15]. My se zaměříme na ty části, které se týkají geometrické pravděpodobnosti.

2. Rodina

Joseph-Louis-François Bertrand se narodil 11. března 1822 v Paříži jako druhý syn Alexandra a Marie-Caroline Bertrandových. Oba jeho rodiče však pocházeli z Rennes. Matčin otec, Joseph Blin (1764–1834), byl ředitelem pošt. V roce 1792 se jako kapitán dobrovolníků z Rennes podílel na obraně Champagne napadené Pruskem. Ač byl přesvědčeným republikánem (a kapitánem granátníků v Národní gardě), vystupoval proti prokonsulovi Jean-Baptiste Carrierovi a zachránil 300–400 osob před deportací do Nantes a následným

utopením[1]. Byl poslancem v *Conseil des Cinq-Cents* za *Ille-et-Vilaine*, po roce 1815 byl zvolen presidentem pěti bretaňských departementů.

Otec Alexandre Bertrand (1795–1831) se při studiích na lyceu v Rennes spřátelil s Jean-Marie Duhamelem[2] a Pierrem Leroux[3]. Spolu s nimi začal v roce 1814 studovat na *École polytechnique*. Ačkoli projevoval značné nadání pro matematiku, rozhodl se školu opustit a věnovat se studiu medicíny. V roce 1825 začal v časopise *le Globe* vydávat *comptes rendus* ze zasedání *Académie des Sciences*[4]. Věnoval se popularizaci vědy (*Lettres sur la Physique, Lettres sur la Révolutions du Globe*), napsal studii o náměsíčnictví, byl stoupencem biomagnetismu (*magnétisme animal*).

3. Dětství, mládí a léta studií

Rodiče Josepha Bertranda byli vzdělaní lidé. Přesto u svého syna potlačovali touhu po vzdělání. Jak sám uvedl, nikdo nevěřil, že se dožije dospělosti, proto jakékoli vzdělávání bylo v jeho případě považováno za ztrátu času a dokonce za něco životu nebezpečného. Joseph se naučil číst sám v necelých pěti letech během dlouhé nemoci, kdy za jeho starším bratrem docházel učitel. Znal už písmena, ale neuměl je spojovat do slov. Poslouchal bratrovo slabikování a ukládal si vše do paměti. Když mu bylo lépe a rodiče mu přinesli knihu o přírodě, aby si prohlížel obrázky, začal k velkému údivu rodičů číst komentáře k obrázkům. Od té chvíle se vzdělávání svého syna věnoval Alexandre Bertrand sám. Bral ho všude s sebou, povídal si s ním o různých námětech, vždy latinsky. Zemřel, když bylo Josephovi 9 let. V té době bydleli v Paříži u Alexandrovy sestry a jejího manžela J.-M. Duhamela, který vedl přípravnou třídu pro studium na *École polytechnique*. Joseph se přátelil se studenty,

[1]V roce 1793 byl J.-B. Carrier (1756–1794) poslán do Nantes, aby všemi prostředky potlačil vzpouru v této oblasti (povstání ve Vendéee (1793–1796)). V Nantes bylo ve vězení soustředěno mnoho zajatých povstalců a díky špatným hygienickým podmínkám a nedostatku jídla se mezi nimi začaly šířit nemoci. Carrier se rozhodl k radikálnímu řešení - popravám. Někteří vězni byli zastřeleni, jiní popraveni gilotinou, další utopeni v Loiře při tzv. marriages républicains, kdy byli vězni spoutáni po dvou (nejlépe osoby opačného pohlaví) a potom naloženi do lodi, odvezeni doprostřed řeky a hozeni do vody. Během jediného roku tak bylo zabito přibližně 10 000 osob, včetně malých dětí.

[2]J.-M. Duhamel (1797–1872) byl významným francouzským matematikem. Působil nejprve na středních školách (*institution Massin, collège Sainte-Barbe, lycée Louis-le-Grand*), od roku 1831 učil na *École polytechnique*. V roce 1834 byl zvolen profesorem na katedře analýzy, později působil na katedře mechaniky, od roku 1851 opět na katedře analýzy. Oženil se s Virginií Bertrand, sestrou Alexandra Bertranda.

[3]P. Leroux (1797–1871) byl vydavatelem, politikem a filosofem, stoupencem hnutí saint-simonismu. Pocházel z velmi chudé rodiny, v Rennes studoval díky státnímu stipendiu. Po smrti otce odešel z *École polytechnique* a vyučil se tiskařem. Byl jedním ze zakladatelů časopisu *le Globe*.

[4]V té době mohli být na zasedání Akademie přítomni jen někteří (akademiky schválení) vědci (A. Bertrand patřil mezi ně). Snahy seznámit širší veřejnost s tím, co se v Akademii děje, narážely na odpor některých členů. *Comptes rendus* vycházely nejprve v *le Globe*, později v *le Temps* (Desiré Roulin) a teprve roku 1835 (kdy se François Arago (1786–1853) stal stálým tajemníkem Akademie) začaly vycházet pod hlavičkou Akademie.

kteří byli mnohem starší než on, a záhy s nimi začal navštěvovat výuku. Profesoři ho nechávali sedět ve třídě, nevšímali si ho a studenti brzy postřehli, že všemu velmi dobře rozumí.

Po manželově smrti opustila Josephova matka Paříž a vrátila se do Rennes. S ní odešel i starší syn Alexandre[5]. Joseph zůstal v Paříži u strýce a tety. V deseti letech pravidelně navštěvoval strýcův kurz speciální matematiky a patřil mezi nejlepší studenty. Když při zkoušení některý student neuměl odpovědět, vyzval Duhamel celou třídu, aby se pokusila odpověď najít. Pokud ani ta neuspěla, obrátil se na Josepha. Většinou odpověď znal. O rok později Duhamel Josephovi vyjednal povolení navštěvovat přednášky na *École polytechnique*. Musel ale podstoupit zkoušku. Při ní byl hodnocen jako druhý nejlepší. Od té doby si své vzdělávání mohl řídit sám. Navštěvoval také přednášky na *Sorbonne*, v *Collège de France*, v *Jardin des plantes*.

Ve Francii v té době bylo potřeba, aby se úspěšný mladý muž honosil také nějakými tituly. Duhamel usoudil, že nadešel čas, aby jeho synovec složil potřebné zkoušky a získal odpovídající tituly. Dal mu k dispozici všechny knihy, které mohl potřebovat, a nechal ho studovat. V roce 1838 během šesti týdnů složil šestnáctiletý Bertrand potřebné zkoušky a získal tituly *bachelier ès lettres* (20.3.), *bachelier ès sciences* (10.4.) a *licencié ès sciences* (4.5.). O rok později sepsal doktorskou práci o termodynamice a po složení zkoušek (9.4. a 22.6.) se stal doktorem přírodních věd (*doctor ès sciences*). V tomtéž roce byl přijat na *École polytechnique*, u zkoušek dosáhl nejlepšího výsledku. V roce 1840 složil státní zkoušku pro výuku matematiky na vysokých školách (*agrégation des Facultés*). Požadavkem u zkoušky byl věk alespoň 25 let, takže Bertrand musel požádat o výjimku. Ta mu byla udělena. O rok později dokončil studia na *École polytechnique*. Protože byl v matematice mnohem silnější než v kreslení, skončil jako šestý, což mu umožnilo pokračovat ve studiu na prestižní *École des mines*. V tomtéž roce složil státní zkoušku pro výuku matematiky na středních školách (*agrégation des Collèges*). Zkoušku skládal také Charles Briot[6], který měl být Bertrandovým silným konkurentem. Mladíci se ale spřátelili (podíl na tom měla možná i madame Duhamel, která oběma na povzbuzení nabídla sklenku malagy) a u zkoušky si dokonce pomáhali. Nakonec získali oba první místo ex-aequo.

V květnu 1842 se Joseph, jeho bratr Alexandre a přítel Marcel Aclocque vydali na výlet do Versailles. Zpátky se vraceli vlakem, který měl 18 vagonů a byl tažen dvěma lokomotivami. Z neznámých příčin přední náprava první lokomotivy praskla. Obě lokomotivy se převrhly a zastavily vlak. Od rozpáleného koksu v topeništi druhé lokomotivy začalo hořet prvních pět vagonů. V té době bylo zvykem zamykat cestující v kupé na klíč, aby byli chráněni před následky své neopatrnosti. Kvůli tomu během krátké doby zemřelo

[5]Alexandre Bertrand (1820–1902) byl významný francouzský archeolog, průkopník galské a galo-románské archeologie, zakladatel a první ředitel *Musée des antiquités nationales*.

[6]C. Briot (1817–1872) byl francouzským matematikem a fyzikem. Učil zpočátku na středních školách, od roku 1855 přednášel na *École normale supérieure*, v roce 1870 nahradil G. Lamé na katedře matematické fyziky na *Sorbonne*.

41 osob v plamenech. Bertrand a jeho přátelé byli vážně popáleni, ale přežili. Do Paříže se vrátili až po deseti dnech. O dva roky později se Joseph oženil s Louise Aclocque, sestrou svého přítele. Z jejich dětí je nejznámější nejstarší syn Marcel (1847–1902), významný geolog. V roce 1844 byl Bertrand jmenován profesorem elementární matematiky v *collège Saint-Louis* a také repetitorem analýzy na *École polytechnique*. To mu trochu komplikovalo studia na *École des mines*. Přesto všechny zkoušky řádně složil a školu dokončil. Nikdy ale jako *ingénieur des mines* nepracoval.

4. **Dospělost, léta učitelská**

V *collège Saint-Louis* byl Bertrand jen o málo starší než jeho žáci. V roce 1848 ze školy odešel, protože mu přibyly povinnosti na *École polytechnique*, kde se stal *examinateur d'admission*, a v *Collège de France*, kde byl pověřen zastupováním Jean-Baptiste Biota[7]. V revolučním roce 1848 byl také zvolen kapitánem Národní gardy. Jeho vojáci o něm však říkali, že nemá vojenského ducha. Přesto, když byl vydán rozkaz, aby spolu se svými muži dobyl jistou barikádu, nezaváhal a navzdory kulkám, které svištěly okolo, se vydal k barikádě. Až u ní zjistil, že ho nikdo nedoprovází. Ostatní se zalekli.

Když bylo v roce 1852 zřízeno (druhé) císařství, proběhla reorganizace výuky na francouzských středních školách. Na nejdůležitější katedry byli povoláni nejzkušenější profesoři. Bertrandovi byla nabídnuta katedra speciální matematiky na *lycée Napoléon* (bývalá *collège Henri IV*). Opustil svou funkci na *École polytechnique*, aby se mohl plně věnovat svému novému pověření. Působil zde jen tři roky, potom definitivně opustil výuku na středních školách.

V roce 1856 se stal profesrem analýzy na *École polytechnique* (nahradil Charlese Sturma[8]) a v roce 1857 začal přednášet na *École normale supérieure*, kde působil pět let do roku 1862. Druhou katedru analýzy na *École polytechnique* vedl od roku 1851 J.-M. Duhamel; strýc a synovec spolupracovali při výuce infinitezimálního počtu až do roku 1869, kdy Duhamel odešel do důchodu. Na jeho místo nastoupil Charles Hermite[9], manžel Bertrandovy sestry Louise. Bertrand zůstal na katedře až do roku 1895, kdy dosáhl věkové hranice pro odchod do důchodu. Na škole působil 51 let, na katedře analýzy 40 let. Také *Collège de France* byl Bertrand věrný. Biotovým zástupcem byl až do roku 1862, kdy získal katedru matematické fyziky, na které působil až

[7]J.-B. Biot (1774–1862) byl fyzikem, astronomem a matematikem, zabýval se studiem polarizace a vztahy mezi elektrickým proudem a magnetismem. V roce 1804 dokončil balón plněný vzduchem a s Gay-Lussacem podnikl výstup do výšky 5 km, aby prozkoumali atmosféru Země. Byl blízkým přítelem Louise Pasteura.

[8]C. Sturm (1803–1855) byl francouzský matematik německého původu. Byl členem Akademie věd, profesorem na *École polytechnique*, následníkem Denise Poissona na katedře mechaniky na fakultě přírodních věd pařížské *Sorbonne*.

[9]C. Hermite (1822–1901) byl významným francouzským matematikem, zabýval se především teorií čísel a algebrou. Jako první dokázal, že Eulerovo číslo je transcendentní. Působil na *École polytechnique*, *École normale supérieure* a *Sorbonne*. Byl členem *Académie des Sciences* a velkodůstojníkem (*grand officier*) Čestné legie.

do své smrti. V roce 1856 byl zvolen členem *Académie des sciences*, od roku 1874 byl stálým tajemníkem (*secrétaire perpétuel*) matematické sekce. V roce 1884 byl zvolen do *Académie française*.

Bertrandovy přednášky byly mezi studenty velmi oblíbené. V [12] na ně vzpomíná Gaston Darboux a tvrdí, že i když ho učilo mnoho výborných profesorů, žádný z nich v něm nezanechal takové vzpomínky jako Bertrand. I těžké důkazy uměl podat formou, která byla pro posluchače přitažlivá. Svým studentům se věnoval s velkým nasazením. Připomeňme alespoň jednoho z nich, Joseph-Émile Barbiera (1839–1889). Studoval na *École normale supérieure*, kde ho učil Joseph Bertrand. Studia ukončil v roce 1860, kdy také publikoval svůj článek [1]. V článku uvádí, že mu s napsáním některých částí pomáhal jeho učitel. Když se u Barbiera v roce 1865 rozvinula duševní nemoc, zmizel z Paříže a zpřetrhal svazky se všemi spolupracovníky, byl to Bertrand, kdo ho vyhledal v ústavu v Charenton-St-Maurice a povzbudil k další matematické práci. Přimluvil se, aby Barbier získal Francoeurovu cenu, a pomohl mu tak k nevelkému příjmu, který umožňoval vést v Paříži život v přijatelných podmínkách.

V roce 1870 vypukla prusko-francouzská válka. Po porážce Francie v bitvě u Sedanu (1. září 1870) bylo svrženo císařství a vyhlášena (třetí) republika. Následně byla Paříž obležena pruskými vojsky, na obraně se podíleli i členové Akademie včetně Bertranda. Zapojili se i jeho synové. Po skončení obléhání byla *École polytechnique* přemístěna do Tours, Bertrand tam musel také odejít, aby dostál svým povinnostem profesora. Tam se také dozvěděl, že při požárech zažehnutých za dnů Pařížské komuny (1871) byl zničen jeho pařížský dům včetně cenné knihovny, rukopisu o termodynamice, který byl připraven k tisku, a také materiálů ke třetímu dílu jeho *Traité de calcul différentiel et de calcul intégral*. Bertrand, zbavený svého domova, se přestěhoval do vily v Sèvres, po jejím vydrancování se usadil ve Virollay.

V roce 1878 se rozhodl přenechat přednášky v *Collège de France* svému zástupci Edmondu Laguerre[10]. V roce 1886 však Laguerre znovu vážně onemocněl a odešel do Bar-le-Duc, kde zanedlouho zemřel. Bertrand se vrátil do školy zastupovat svého zástupce. V následujících letech napsal podle svých přednášek tři učebnice [7], [8] a [9]. Zemřel v Paříži 3. dubna 1900.

5. Dílo

První články napsal Bertrand po přijetí na *École polytechnique*, týkaly se problému rozvodu elektřiny a prokázaly, že je napsal opravdový geometr. Také další práce byly z oblasti geometrie, ale také analýzy, matematické fyziky a mechaniky. Byly publikovány v *Jornal de mathématiques pures et appliquées*, *Journal de l'École polytechnique*, později i v *Comptes rendus de l'Académie des sciences*.

[10]E. Laguerre (1834–1886) byl francouzským matematikem, zabýval se především geometrií a komplexní analýzou. Měl velmi chatrné zdraví.

Své zkušenosti z výuky na střední škole (*collège Saint-Louis*) využil k napsání dvou středoškolských učebnic, jedna byla o aritmetice [2], druhá o algebře [3]. Obě byly velmi dobře napsány a měly velký vliv na výuku matematiky na francouzských lyceích. Vzbuzovaly ve studentech zájem o matematiku a dávaly jim chuť k dalšímu bádání.

Přednášky na *Collège de France* inspirovaly napsání učebnice diferenciálního a integrálního počtu [4], jejíž první dva díly vyšly v roce 1864 a 1870. Rozpracovaný třetí díl shořel ve dnech Pařížské komuny a Bertrand ho již znovu nezpracoval. Předmluva obsahuje historii diferenciálního a integrálního počtu, která byla na jeho kursech také vyučována. V dalších částech Bertrand vedle známých výsledků prezentuje také výsledky své, které byly uveřejněny v různých článcích v předchozích letech.

Postupně se Bertrand začal zajímat také o historii vědy. Souviselo to s tím, že nekolikrát přijal nabídku mluvit jménem *Académie des sciences* na výročních zasedáních *Institutu*. První historickou prací bylo [5]. Bertrand zde zachytil život a práci nejznámějších astronomů, jejich objevy popisoval způsobem, který byl srozumitelný i lidem, kteří neměli vyšší matematické vzdělání. O čtyři roky později vyšla kniha [6] věnovaná historii *Académie des sciences* od jejího založení v roce 1666 do roku 1793. Toto téma je jistě velmi obsáhlé, Bertrand uvádí, že se chce věnovat především změnám v organizační struktuře, průběhu zasedání a také vztahům mezi členy navzájem. V první části popisuje historii, druhá část je věnována jednotlivým akademikům. Popisuje jejich životy i charaktery, hodnotí jejich dílo. Byl velkým obdivovatelem Jean Le Rond d'Alemberta, jehož životu a dílu věnoval studii v *Collection des grands écrivains français*. Jako stálý tajemník matematické sekce *Académie des sciences* pronesl *éloges* při úmrtí devatenácti akademiků, mezi nimi byl třeba August Cauchy, Gabriel Lamé, Victor Puiseux a Urbain Le Verrier. Historii věnoval také články v *Journal des savants*.

6. **Bertrand a geometrická pravděpodobnost**

Úlohy geometrické pravděpodobnosti nalezneme v [8] a [4]. V úvodu [8] Bertrand popisuje historii teorie pravděpodobnosti a uvádí známé příklady (jako třeba *Petrohradský paradox*). V první kapitole pak definuje pravděpodobnost jako poměr počtu příznivých jevů ku počtu jevů možných a na několika příkladech ukazuje, jak pravděpodobnost spočítat. Jak ale postupovat v případě, že je náhodných jevů nekonečně mnoho? Bertrand ukazuje, že v tomto případě je možné pojem *náhodně vybrat* chápat více způsoby a vypočtené pravděpodobnosti se pak liší. Ve čtvrtém odstavci chce spočítat pravděpodobnost, že vybereme-li *náhodně* číslo od jedné do sta, bude toto číslo větší než 50. Odpověď se zdá být zřejmá - 1/2. Když ale místo čísla budeme uvažovat jeho druhou mocninu, bude pravděpodobnost, že je číslo větší než 50 (a tedy jeho mocnina větší než 2500) rovna 3/4.

V pátém odstavci je uveden příklad známý jako *Bertrandův paradox*: Vybíráme *náhodně* tětivu kružnice a ptáme se, jaká je pravděpodobnost, že tato

tětiva bude menší než strana rovnostranného trojúhelníka vepsaného dané kružnici. Bertrand předkládá tři různá řešení. Nejprve pevně zvolí jeden koncový bod tětivy a náhodně vybírá její směr. Pravděpodobnost, že je tětiva delší[11] než strana rovnoramenného trojúhelníka, je 1/3. V druhém řešení je pevně dán směr tětivy a náhodně vybírána vzdálenost tětivy od středu kružnice. Pravděpodobnost, že je tětiva delší než strana trojúhelníka, je 1/2. V třetím řešení je náhodně vybírán střed tětivy a takto vypočtená pravděpodobnost je 1/4. Tento rozpor přirozeně vzbudil pochybnosti o výsledcích dosažených v rozvíjející se geometrické pravděpodobnosti. A také snahu matematiků tento rozpor vysvětlit. Ve druhém vydání [14] Poincaré ukázal, že v takto zadané úloze je správné druhé řešení, protože je invariantní vzhledem k posunutí, rotaci a reflexi. Totéž uvádí i Borel v [10].

V šestém odstavci je dán další příklad: Vyberme *náhodně* rovinu v prostoru. Jaká je pravděpodobnost, že svírá s horizontem úhel menší než $\pi/4$? První řešení počítá s tím, že úhel nabývá hodnot mezi 0 a $\pi/2$. Pravděpodobnost je tedy 1/2. V druhém řešení uvažujeme přímku (paprsek) kolmou k (vybrané) rovině a procházející středem koule. Vybrat náhodně rovinu je totéž jako vybrat náhodně průsečík odpovídajícího paprsku s povrchem koule. Sevřený úhel bude menší než $\pi/4$ právě tehdy, když průsečík leží v oblasti, jejíž povrch je roven $4\pi R^2 \sin^2(\pi/8)$ (povrch vrchlíku). Hledaná pravděpodobnost je $2\sin^2(\pi/8)$, tedy přibližně 0,29.

A v sedmém odstavci je tento příklad: Vyberme *náhodně* dva body na povrchu koule. Jaká je pravděpodobnost, že jejich vzdálenost je menší než 10 minut? V prvním řešení je kružnice, která spojuje tyto dva body, rozdělěna na 2160 dílů po 10 minutách. Hledaná pravděpodobnost je 2/2160=1/1080. V druhém řešení je dán jeden z těch bodů. Druhý musí ležet v oblasti, jejíž povrch je $4\pi R^2 \sin^2(\pi/2160)$ (povrch vrchlíku). Hledaná pravděpodobnost je 1/236 362.

Ve 43. odstavci ve třetí kapitole je uvedena Buffonova úloha o jehle: Na neomezenou plochu jsou ve stejných vzdálenostech narýsovány rovnoběžky. Jehla je náhodně házena na tuto plochu. Pierre dostane 1 frank, když jehla protne nějakou rovnoběžku. Jaká je Pierrova očekávaná výhra?[12] Bertrand píše, že očekávaná výhra je závislá jen na délce jehly, nikoli jejím tvaru. Ukazuje také rozdíl, který vznikne, když úsečku (jehlu) nahradíme křivkou. Úsečka, která je kratší než vzdálenost mezi rovnoběžkami, může protnout nejvýše jednu rovnoběžku, zatímco křivka stejné délky může mít průsečíků více. Když je tedy na plochu s rovnoběžkami házena jehla délky $l \leq a$, kde a je vzdálenost mezi rovnoběžkami, je pravděpodobnost protnutí stejná jako očekávaná výhra. Když ale budeme na plochu házet kružnici o poloměru

[11] Opravdu je v zadání tětiva *menší* a v řešeních *delší*.

[12] On trace sur un plan idéfini des lignes parallèles équidistantes. Une aiguille est lancée au hasard sur le plan. Pierre recevra 1 fr par rencontre de l'aiguille avec une des parallèles. Quelle est l'espérance mathématique de Pierre?

$R \leq a/2$, je pravděpodobnost protnutí $2R/a$, zatímco očekávaná výhra je $4R/a$, protože pokud kružnice protne nějakou rovnoběžku, protne ji dvakrát.

Geometrické pravděpodobnosti se Bertrand věnuje také v [4]. Poslední část páté kapitoly je věnována Croftonově větě (poprvé publikovaná v [11]). Bertrand nejprve ukazuje Barbierův výsledek [1], potom uvádí Croftonovu větu a dokazuje ji právě s využitím Barbierových úvah.

7. Závěr

Joseph Bertrand se vždy snažil předat látku svým posluchačům a čtenářům co nejsrozumitelněji. Studenti tuto jeho snahu oceňovali, byl velmi oblíbený. V roce 1895 se Bertrandovi kolegové a také studenti rozhodli oslavit padesát let jeho působení na *École polytechnique* a při té příležitosti nechali pro něj vyrobit krásnou medaili u známého rytce Jules-Clément Chaplaina. To byla pocta, které se nedostalo ani Cauchymu nebo Lamému.

Literatura

[1] Barbier J.-É (1860) *Note sur le problème de l'aiguille et le jeu du joint couvert.* Journal de mathématiques pures et appliquées **5**, 273 – 286.
[2] Bertrand J. (1849) Traité d'arithmétique. Librairie Hachette, Paris.
[3] Bertrand J. (1850) Traité d'algèbre. Librairie Hachette, Paris.
[4] Bertrand J. (1864–1870) Traité de calcul différentiel et de calcul intégral. Gauthier-Villars, Paris.
[5] Bertrand J. (1865) Les fondateurs de l'astronomie moderne: Copernic, Tycho Brahé, Képler, Galilée, Newton. J. Hetzel, Paris.
[6] Bertrand J. (1869) L'Académie des sciences et les académiciens de 1666 à 1793. J. Hetzel, Paris.
[7] Bertrand J. (1887) Thermodynamique. Gauthier-Villars, Paris.
[8] Bertrand J. (1889) Calcul des probabilités. Gauthier-Villars et fils, Paris.
[9] Bertrand J. (1890) Leçons sur la théorie mathématique de l'électricité. Gauthier-Villars et fils, Paris.
[10] Borel É. (1909) Élements de la théorie des probabilités. Hermann, Paris.
[11] Crofton M.W. (1868) *On the theory of local probability, applied to straight lines drawn at random in a plane, the methods used being also extendedto the proof of certain new theorems in the integral calculus.* Philosophical transaction of the Royal society of London, **158**, 181 – 199.
[12] Darboux G. (1902) *Éloge historique de J.-L.-F.Bertrand.* Éloges académiques, nouvelle série, Librairie Hachette, Paris, VII – LI.
[13] Laplace P.-S. de (1812) Théorie analytique des probabilités. Imprimerie Royale, Paris.
[14] Poincaré J.H. (1896) Calcul des probabilités. Gauthier-Villars, Paris, (2. vydání Carré, Paris, 1912).
[15] Sheynin O.B. (1994) *Bertrand's work on probability.* Arch. Hist. Exact. Sci **48**(2), 155 – 199.

Adresa: FEL ČVUT, kat. matematiky, Technická 2, 166 27 Praha 6 – Dejvice

E-mail: kalous@math.feld.cvut.cz

APLIKACE BODOVÝCH PROCESŮ PŘI ANALÝZE VEŘEJNÉ SPRÁVY V ČR

Radka Lechnerová, Tomáš Lechner

Klíčová slova: Bodové procesy, sumární statistiky, e-Government, veřejná správa v České republice.

Abstrakt: Prostorové rozmístění poskytovatelů veřejných služeb je důležitým indikátorem při zjišťování ukazatelů efektivnosti a účinnosti výkonu veřejné správy v ČR. Jestliže danou kategorii veřejné služby poskytují pouze některé z daných orgánů veřejné moci, lze prostorové rozmístění těch orgánů, které vybranou službu poskytují, účinně zkoumat za pomoci sumárních statistik bodových procesů. Použitá metoda navíc není závislá na konkrétním rozmístění orgánů v rámci území, tj. dává spolehlivé odpovědi, i když orgány nejsou rovnoměrně náhodně rozmístěné v rámci území ČR, což je realitou. Zkoumali jsme tak prostorové rozmístění a vzájemné interakce úřadů územních samosprávných celků, které provozují některé z následujících služeb: elektronické podatelny, své datové schránky si aktivovali dobrovolně dříve, Czech POINT, výkon přenesené působnosti na úseku stavebního úřadu nebo poskytují další služby v rámci výkonu přenesené působnosti. Získané výsledky podstatným způsobem doplňují obraz výkonu veřejné správy v České republice.

Abstract: Spatial distribution of public services providers is an important indicator in determining the effectiveness and efficiency of Public Administration in the Czech Republic. If some category of public services is only provided by chosen local authority, we can spatial distribution of authority, providing the selected service, effectively examined by means of summary statistics of point processes. Moreover, the used method does not depend on the specific distribution of local authority within the territory of the Czech Republic, i.e. it gives a reliable answer, even if local authorities are not uniformly random distributed within the territory, which is a reality. We examined the spatial distribution and interactions of municipalities, which provide some of following services: electronic registry, their data boxes were activated voluntary earlier, service called "Czech POINT", exercise delegated powers in the field of Building Authority or provide other services in the exercise of delegated powers. The obtained results significantly complement the image of Public Administration in the Czech Republic.

1. Úvod

Veřejná správa v České republice funguje na základě zákona a v jeho mezích. Z toho také nutně plyne, že veškeré poskytované veřejné služby musí být podloženy příslušnými právními předpisy. V některých případech je poskytování určité služby nařízeno (např. provozování elektronické podatelny), v jiných

je definováno, kdo danou službu smí poskytovat, nicméně konkrétní realizace nařízena není (např. kontaktní místa veřejné správy, tzv. Czech POINT). Od 90. let minulého století probíhá ve veřejné správě v České republice proces implementace informačních a komunikačních technologií s cílem zvýšení efektivity jejího výkonu. Tento proces bývá označován jako e-Government.

Mezi orgány veřejné moci, které vykonávají veřejnou správu v České republice, se obecně řadí státní orgány, orgány územních samosprávných celků, Pozemkový fond České republiky a jiné státní fondy, zdravotní pojišťovny, Český rozhlas, Česká televize, samosprávné komory zřízené zákonem, notáři a soudních exekutoři, celkem asi 8 234 subjektů [6]. V rámci našeho výzkumu se zabýváme pouze vybranou částí veřejné správy, tedy orgány územních samosprávných celků (ÚSC), kterých je 6 248 [5]. Oblast těchto orgánů se vyznačuje stejnorodostí pravidel a předpisů, jimiž se tyto orgány musí řídit, a je tedy vhodná k celkovému statistickému zpracování.

Postupně jsme se zabývali elektronickými podatelnami, které všechny orgány mají mít zřízeny od října 2001, ale tuto povinnost plnilo v roce 2008 jen 15 % z nich [2]. Dále kontaktními místy veřejné správy, která může provozovat ten úřad územního samosprávného celku, který zároveň vykonává přenesenou působnost na úseku matrik, anebo požádal Ministerstvo vnitra o zápis do seznamu kontaktních míst, jež je zveřejňován v podobě vyhlášky ve Sbírce zákonů. Po té datovými schránkami, a to konkrétně v rámci přechodného období, kdy si je mohly orgány veřejné moci dobrovolně aktivovat. A konečně výkonem přenesené působnosti úřady územních samosprávných celků, jmenovitě matričními úřady, stavebními úřady a živnostenskými úřady. Výsledky z oblasti elektronických podatelen již byly publikovány v [2], výsledky z oblasti kontaktních míst veřejné správy a datových schránek v [1]. V rámci tohoto příspěvku je prezentováno celkové sjednocení uvedených výzkumů a jejich výsledků a následné rozšíření do nově zkoumané oblasti kvality výkonu státní správy, a to zejména z prostorového hlediska.

2. Data

Geografická data prezentovaná Českým statistickým úřadem v rámci Územně identifikačního registru ÚIR-ZSJ [3] jsou uváděna v metrech v souřadnicovém systému jednotné trigonometrické sítě katastrální (S-JTSK), který je definován v nařízení vlády č. 430/2006 Sb., o stanovení geodetických referenčních systémů a státních mapových děl závazných na území státu a zásadách jejich používání. Výhodou tohoto souřadnicového systému je, že v něm můžeme provádět standardní geodetická měření. Vzdálenosti bodů lze získat s chybou, která může být maximálně v jednotkách metrů. Jelikož aproximujeme obce ČR bodem v místě polohy úřadu příslušného ÚSC, je tato chyba zanedbatelná.

Adresář elektronických podatelen úřadů ÚSC platný k únoru 2008 jsme převzali z Portálu veřejné správy [8], kde jsou data prezentována na základě

nařízení vlády č. 495/2004 Sb., kterým se provádí zákon o elektronickém podpisu. Adresář kontaktních míst veřejné správy podložených zákonem č. 365/2000 Sb., o informačních systémech veřejné správy, ve znění pozdějších předpisů, jsme převzali z portálu Czech POINT [9] v září 2009. Údaje pro dobrovolně aktivované datové schránky v rámci přechodného období definovaného zákonem č. 300/2008 Sb., o elektronických úkonech a autorizované konverzi dokumentů, ve znění pozdějších předpisů, vycházejí z vlastního vyhledání adres datových schránek úřadů ÚSC provedeného pomocí aktivované datové schránky fyzické osoby, tj. odrážejí aktuální stav dostupný přímo v informačním systému datových schránek v dané dny (16. 9. a 6. 10.) přechodného období. Adresář matričních, stavebních a živnostenských úřadů jsme převzali z portálu státní správy [7] v lednu 2010.

3. Metoda

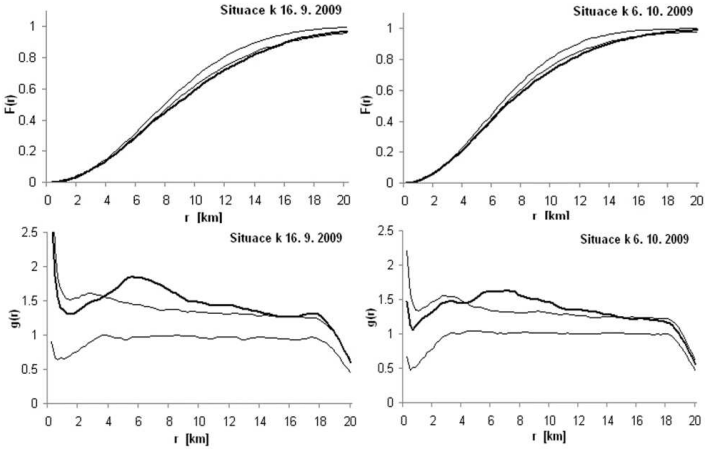
Nechť $X = \{X_1, \ldots, X_n\}$ jsou body, které reprezentují polohy úřadů ÚSC, jimiž aproximujeme obce v ČR. Obce provozující danou službu nechť jsou reprezentovány body $Y = \{Y_1, \ldots, Y_k\}$, přičemž platí, že $Y \subseteq X$. Pozorovacím oknem nechť je území České republiky nebo její část. Budeme testovat hypotézu

H_0: Body Y jsou rovnoměrně náhodně rozdělené na X.
H_1: Body Y nejsou rovnoměrně náhodně rozdělené na X.

Test založíme na sumárních statistikách bodových procesů, kterými lze dobře charakterizovat prostorové rozmístění bodů. Konkrétně jsme použili distribuční funkci nejbližších sousedu (G), sférickou kontaktní distribuční funkci (F) a párovou korelační funkci (g) (viz napr. [5, 6]). Z jejich definic vyplývá, že distribuční funkce $G(r)$ resp. $F(r)$ určují pravděpodobnost, že se do vzdálenosti r od bodu procesu resp. libovolného bodu v prostoru vyskytuje (v případe funkce G jiný) bod procesu. Poznamenejme, že jsme použili Kaplan-Meierovy odhady pro funkce G a F [3] a Ripleyho odhad pro funkci g [4], které jsou implementovány v programu R (balíček SPATSTAT). Všechny zmíněné odhady jsou neparametrické a zahrnují korekci okrajových efektů.

Vlastní Monte Carlo test spočívá získání 95% intervalu spolehlivosti na základě simulací realizací příslušného bodového procesu s k body a odhadnutí příslušné sumární statistiky. Porovnáním tohoto intervalu s odhadnutou sumární statistikou pro Y můžeme rozhodnout o platnosti testované hypotézy na hladině testu 5 %. Pokud výše zmíněnou hypotézu zamítneme, jsme navíc schopni rozhodnout, zda se vyskytují mezi body přitažlivé či odpudivé interakce.

Hlavní výhodou testu je, že výsledek není závislý na konkrétním rozmístění úřadů v rámci území, tj. dává spolehlivé odpovědi, i když úřady nejsou rovnoměrně náhodně rozdělené v rámci území, což je realitou. Zároveň zodpovídá, zda se vyskytují či nevyskytují mezi body Y nějaké interakce.

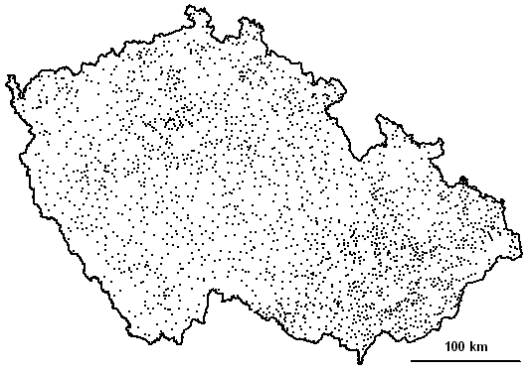


Obrázek 1. Grafy sférické kontaktní distribuční funkce (horní řádek) a párové korelační funkce (dolní řádek) spočtené pro data (tučné křivky). Dále jsou zde zobrazeny obálky (tenké křivky) vymezující 95% interval spolehlivosti pro testování hypotézy o rovnoměrně náhodném rozmístění obcí s datovými schránkami v obcích ČR z 16. 9. 2009 (levý sloupec) a z 9. 10. 2009 (pravý sloupec).

4. **Numerické výsledky**

Navzdory nařízení vlády č. 495/2004 Sb., které dává povinnost všem orgánům veřejné moci provozovat elektronické podatelny, provozovalo v únoru 2008 pouze 15 % úřadů ÚSC funkční elektronickou podatelnu. Protože zákon č. 500/2004 Sb., správní řád, ve znění pozdějších předpisů, zavádí možnost zajistit provozování elektronických podatelen pro menší obce obcemi s rozšířenou působností v rámci spádových oblastí, zajímalo nás, zdali lze skutečně pozorovat shluky elektronických podatelen odpovídající uvedeným spádovým oblastem. Výsledky získané výše popsanou metodou a publikované v [2] ukazují, že v případě zajištění provozu elektronické podatelny mezi sebou úřady ÚSC výrazně nespolupracují, což není ve shodě se záměrem vtěleným do legislativy.

Dalším nástrojem elektronické komunikace po elektronických podatelnách jsou datové schránky podložené zákonem č. 300/2008 Sb., o elektronických úkonech a autorizované konverzi dokumentů, ve znění pozdějších předpisů. Protože v tomto případě jde o zcela nový typ elektronické komunikace na rozdíl od elektronických podatelen, které byly pouze jakýmsi funkčním rozšířením e-mailové pošty, ptáme se, zda v tomto případě mezi sebou úřady ÚSC v rámci přechodného období, definovaného zákonem od 1. července 2009 do 31. října 2009, již spolupracují.

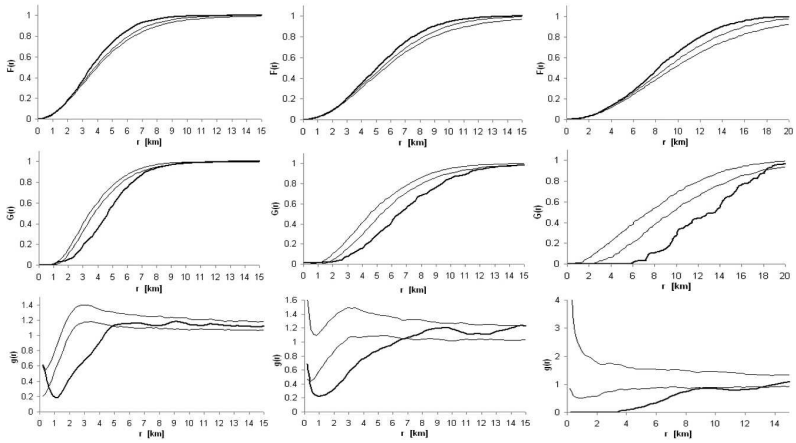


Obrázek 2. Na obrázku jsou vyznačeny polohy obcí České republiky, které provozují Czech POINT, stav v říjnu 2009.

Výsledky získané popsanou metodou (viz Obr. 1) a publikované také v [1] ukazují, že na základě funkce G nelze hypotézu, že úřady s aktivní DS jsou rovnoměrně náhodně rozdělené na množině všech úřadů ÚSC, oproti alternativě, že tomu tak není, zamítnout, neboť křivka odhadu G funkce pro zkoumaná data leží pro všechny vzdálenosti mezi obálkami 95% intervalu spolehlivosti. Zatímco na základě funkce F i funkce g hypotézu zamítáme ve prospěch shlukování, což odpovídá přitažlivým interakcím mezi těmito úřady. Znamená to, že na rozdíl od legislativně předpokládané spolupráce v oblasti provozu elektronických podatelen, kde spolu úřady ÚSC významně nespolupracují, v případě postupného zavádění zcela nového způsobu komunikace spolupráce obcí existuje. To může být velmi důležitou okolností pro realizaci tzv. technologických center [1], která mají zajistit bezpečné zálohování elektronických dokumentů vždy v rámci spádové oblasti obce s rozšířenou působností.

Jinou otázku si budeme klást v případě kontaktních míst veřejné správy, i když použitá stochastická metoda výzkumu je samozřejmě shodná jako v předchozích dvou případech. Kontaktní místa veřejné správy známá jako Czech POINT (Český podací ověřovací informační národní terminál) poskytují například následující služby: získání ověřeného výpisu z informačních systémů veřejné správy, podání podle živnostenského zákona či autorizovanou konverzi dokumentů. Ptáme se tedy, jak je těmito službami pokryto území ČR. Zde by tedy negativní odpověď byla, pokud bychom zamítali hypotézu, že úřady poskytující službu Czech POINT jsou rovnoměrně náhodně rozdělené na množině všech úřadů ÚSC, ve prospěch shlukování, a pozitivní odpověď bude, pokud hypotézu zamítat nebudeme anebo ji zamítneme ve prospěch odpudivých sil, což by znamenalo dobré rozmístění poskytovatelů služby s ohledem na celkové území ČR.

V říjnu 2009 poskytovalo službu Czech POINT 2335 z 6249 úřadů ÚSC (viz Obr. 2).



Obrázek 3. Grafy sférické kontaktní distribuční funkce (první řádek), distribuční funkce nejbližších sousedů (druhý řádek) a párové korelační funkce (třetí řádek) spočtené pro data (tučné křivky) matričních úřadů (levý sloupec), stavebních úřadů (prostřední sloupec) a živnostenských úřadů (pravý sloupec). Dále jsou zde zobrazeny obálky (tenké křivky) vymezující 95% interval spolehlivosti pro testování hypotézy o rovnoměrně náhodném rozmístění obcí s úřady vykonávající danou přenesenou působnost.

Výsledky získané popsanou metodou a publikované také v [1] ukazují, že na základě funkce G, F i g zamítáme hypotézu o náhodném rozmístění úřadů poskytující službu Czech POINT na množině všech úřadů ÚSC a pozorujeme odpudivé interakce ve vzdálenostech 2–3,5 km, ve větších vzdálenostech je rozmístění víceméně rovnoměrně náhodné. To pro Českou republiku znamená poměrně dobré rozložení a pokrytí obcí kontaktními místy veřejné správy.

Zcela nově jsme aplikovali popsanou stochastickou metodu na výzkum kvality výkonu přenesené působnosti z podobného hlediska jako v případě kontaktních míst veřejné správy, tj. dostupnosti a dobrého pokrytí území ČR matričními, stavebními a živnostenskými úřady. Detailní výsledky jsou na Obr. 3.

Přenesenou působnost na úseku matrik vykonává přibližně 19,6 % úřadů ÚSC, na úseku stavebního úřadu asi 10 % úřadů ÚSC a na úseku živnostenského úřadu přibližně 3,3 % úřadů ÚSC. Na základě distribuční funkce nejbližších sousedů (G), sférické kontaktní distribuční funkce (F) i párové korelační funkce (g) hypotézu o náhodném rozmístění zmíněných úřadů na množině všech úřadů ÚSC zamítáme ve prospěch regularit tj. pozorujeme zde odpudivé interakce mezi těmito úřady. Tyto výsledky ukazují, že výkon přenesené působnosti úřady ÚSC na území ČR je nastaven v souladu s principem co nejlepšího pokrytí území ČR.

5. Závěr

Aplikace popsané stochastické metody založené na sumárních statistikách bodových procesů umožňuje zodpovědět důležité otázky z oblasti veřejné správy v ČR. Bylo tak zjištěno, že při implementaci nástrojů e-Governmentu začínají úřady ÚSC více spolupracovat, než tomu bylo v minulosti. Plyne to ze srovnání výsledků pro elektronické podatelny a dobrovolnou aktivaci datových schránek. Zjištěná situace je důležitá pro přípravu dalšího rozvoje e-Governmentu v území; zejména v podobě vazeb na základní registry veřejné správy (v roce 2012) a budoucí Národní digitální archiv (v roce 2013).

V rámci výkonu přenesené působnosti úřady ÚSC jsou příslušné matriční, stavební a živnostenské úřady rozmístěny v území s ohledem na co nejlepší dostupnost těchto služeb pro všechny občany, jak ověřily námi získané výsledky. Podstatné totiž je, že určitá míra dostupnosti veřejných služeb pro daného občana nemá být závislá na faktu, jak velké je sídlo, v němž občan žije, a proto jsou naše výsledky relevantní. Obdobně pozitivní výsledek byl získán také pro rozmístění kontaktních míst veřejné správy.

Literatura

[1] Lechner T., Lechnerová R. (2009) *Vývoj e-Governmentu v České republice – ekonomické a prostorové aspekty.* Sborník Regionálna a miestna verejná správa v znalostnej ekonomike. E. Žárska, V. Vlčková, T. Černěnko (Eds.), Ekonomická univerzita v Bratislavě, Bratislava, Slovensko, XIII-1 – XIII-10.

[2] Lechnerová, R., Lechner T. (2009) *Analýza rozmístění elektronických podatelen obcí v České republice.* In Sborník prací 15. letní školy JČMF Robust 2008. J. Antoch a G. Dohnal (Eds.), JČMF, Praha, 231 – 238.

[3] Moller J., Waagepetersen R.P. (2003) *Statistical Inference and Simulation for Spatial Point Processes.* Chapman & Hall/CRC, New York.

[4] Stoyan D., Stoyan H. (1994) Fractals, rando m shapes and point fields: methods of geometrical statistics. John Wiley and Sons, Chichester.

[5] Český statistický úřad: Územně identifikační registr. Citace 1. 9. 2008. Dostupné na: `http://www.czso.cz/csu/rso.nsf/i/prohlize_uir_zsj`

[6] Datové schránky: `http://www.datoveschranky.info/seznam.php`

[7] Portál státní správy: `shttp://www.statnisprava.cz`

[8] Portál veřejné správy (PVS): Adresář elektronických podatelen orgánů veřejné moci. Citace 1. 2. 2008. Dostupné na: `http://portal.gov.cz/wps/portal/_s.155/696/_s.155/696?kam=epodatelny &paging=10&epodatelnyTable.stk_page=0&epodatelnyTable.stk_npage=1 &epodatelnyTable.stk_pageSize=10`

[9] Projekt Czech POINT. Citace 1. 10. 2009. Dostupné na: `http://www.czechpoint.cz`

Poděkování: Tato práce byla podporována granty GAAV IAA 101120604 a VŠE IG508010.

Adresa: R. Lechnerová, SVSEŠ, s.r.o., Lindnerova 575/1, 180 00 Praha 8-Libeň; T. Lechner VŠE v Praze, Národohospodářská fakulta, katedra práva, nám. W. Churchilla 4, 130 67 Praha 3

E-mail: `radka.lech@seznam.cz, lechner@triada.cz`

TESTY DOBRÉ SHODY PRO MODEL ZRYCHLENÉHO ČASU V ANALÝZE PŘEŽITÍ

Petr Novák

Klíčová slova: Analýza přežití, testy dobré shody, model zrychleného času.

Abstrakt: V příspěvku studujeme regresní modely pro analýzu přežití, věnujeme se především možnostem, jak sestavit testy dobré shody pro model zrychleného času. Porovnáváme je s testy pro Coxův model proporcionálního rizika založenými na teorii čítacích procesů. Na simulovaných datech zkoumáme empirické vlastnosti testů těchto modelů, pozorujeme jejich sílu v závislosti na velikosti sledovaného výběru, typu regresorů a tvaru základního rizika. Hledáme, v jakých situacích je možné dobře rozlišit, podle kterého modelu se data chovají a naopak kdy je rozlišení mezi modely obtížnější.

Abstract: In present work we study regression models in survival analysis, we focus mainly on options how to perform goodness-of-fit tests for the Accelerated Failure Time model. We compare those methods with existing tests for the Cox proportional hazards model which are based on counting process theory. On simulated data, we study empirical properties of these tests. We compare their empirical power for various sample sizes, covariate types and basic hazard. We try to find cases when it is possible to distinguish between the models well and when not.

1. Regrese v analýze spolehlivosti

Studujeme data reprezentující dobu od začátku pozorování do dosažení nějaké předem definované události - poruchy - v závislosti na vysvětlujících proměnných. Počítáme s nezávislým cenzorováním zprava, tj. že u některých jedinců je pozorování ukončeno před dosažením poruchy. Označíme T_i^* skutečné časy událostí a C_i časy cenzorování. Data máme ve tvaru $(T_i, \Delta_i, \boldsymbol{X}_i)_{i=1}^n$, kde $T_i = \min(T_i^*, C_i)$, $\Delta_i = I(T_i \leq C_i)$ a $\boldsymbol{X}_i$ je vektor regresorů.

Dále označme $\alpha_i(t) = \lim_{h\to 0} P(t \leq T_i^* < t+h | T_i^* \geq t)/h$ rizikovou funkci. Data se reprezentují také jako čítací procesy, označme $N_i(t) = I(T_i \leq t, \Delta_i = 1)$, $Y_i(t) = I(t \leq T_i)$, intenzity $\lambda_i(t) = Y_i(t)\alpha_i(t)$ a kumulované intenzity $\Lambda_i(t) = \int_0^t \lambda_i(s)ds$. Bylo dokázáno, že $M_i(t) := N_i(t) - \Lambda_i(t)$ jsou za platnosti daného modelu martingaly vzhledem k filtraci [3]

$$F_{t^-} = \sigma\left\{N_i(s), Y_i(s), X_i, 0 \leq s < t, i = 1, ..., n\right\}$$

Pomocí čítacích procesů se dá přepsat logaritmická věrohodnostní funkce dat a jejím derivováním dle případných parametrů získáváme skórový proces $U(t, \beta)$, pro odhady používáme tento proces až do nějakého času τ, vyššího než je čas poslední události (píšeme $U(\beta) = U(\tau, \beta)$).

2. Nejpoužívanější modely

Srovnáme zde dva ze základních regresních modelů analýzy přežití a možnosti jak provést přílušné testy dobré shody. Nejčastěji používaným je Coxův model proporcionálního rizika [2]:

$$\alpha_i(t) = \exp(\boldsymbol{X}_i^T\boldsymbol{\beta})\alpha_0(t), \qquad i = 1, ..., n, \qquad t = [0, \tau],$$

kde $\alpha_0(t)$ je rizikovou funkcí tzv. základního rozdělení. Dalším obvyklým je model zrychleného času (Accelerated Failure Time - AFT, [1]):

$$\log(T_i^*) = -\boldsymbol{X}_i^T\boldsymbol{\beta} + \epsilon_i, \qquad i = 1, ..., n.$$

kde ϵ_i jsou (iid). Pozor, neznáme skutečné hodnoty T_i^*, ale pouze pozorované T_i. Platí $\alpha_i(t) = \alpha_0(e^{\boldsymbol{X}_i^T\boldsymbol{\beta}}t)e^{\boldsymbol{X}_i^T\boldsymbol{\beta}}$, kde $\alpha_0(t)$ je rizikovou funkcí pro veličiny $\exp(\epsilon_i)$. Pro $\alpha_0(t)$ odpovídající Weibullovu rozdělení se modely shodují pro $\boldsymbol{\beta}_C = \delta\boldsymbol{\beta}_A$, kde δ je parametr tvaru Weibullova rozdělení. Oba modely se od sebe odlišují interpretací parametrů, i tím, jak jsou motivovány. V Coxově modelu působí hodnoty kovariát přímo na rizikovou funkci, v AFT modelu regesory způsobují, že virtuálně běží čas pro daný subjekt rychleji nebo pomaleji. Je proto dobré umět rozlišit, podle kterého modelu se data chovají.

Testy dobré shody pro AFT model

Dosazením odhadů $\hat{\boldsymbol{\beta}}$ do rovnice modelu získáme rezidua

$$r_i := \log(T_i) + \boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}.$$

Ta narozdíl od ϵ_i nejsou ani nezávislá ani stejně rozdělená, protože odhady $\hat{\boldsymbol{\beta}}$ jsou založené na celém datovém souboru. Vzhledem k asymptotické konzistenci odhadů $\hat{\boldsymbol{\beta}}$ [4] mají ale r_i mít přibližně stejnou střední hodnotu. Pokud máme cenzorovaná data, odhadneme rezidua jako

$$\hat{r}_i := \Delta_i r_i + (1 - \Delta_i)E(\epsilon|\epsilon > r_i^C),$$

kde $E(\epsilon|\epsilon > r_i^C)$ odhadneme jako průměrnou hodnotu všech reziduí necenzorovaných pozorování vyšších než $r_i^C = \log T_i + \boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}$. Rozdělíme data do dvou skupin podle hodnot regresorů a testujeme shodu středních hodnot mezi těmito podvýběry. Použijeme t-test a Wilcoxonův test, kvůli nestejnému rozdělení reziduí budou výsledky pouze přibližné. Vyhodnotíme zde proto empirickou sílu testů v závislosti na velikosti výběru, abychom mohli stanovit, jaká je rychlost asymptotické konvergence.

Testy dobré shody pro Coxův model

Za platnosti Coxova modelu je možné pomocí martingalové dekompozice a centrální limitní věty simulovat proces, který je asymptoticky ekvivalentní skórovému procesu $\tilde{U}(t, \hat{\boldsymbol{\beta}}) = \sum_{i=1}^n \boldsymbol{X}_i\hat{M}_i(t)$ (blíže viz [5]). Takto získané

replikace pak porovnáme s hodnotou spočítanou z dat. Pro testování použijeme supremovou statistiku $sup_{t\in[0,\tau]}\|\tilde{U}(\hat{\boldsymbol{\beta}},t)\|$. Pokud její hodnota překročí $(1-\alpha)\%$ hodnot simulovaných statistik, zamítáme hypotézu, že data se chovají podle Coxova modelu. Vždy jsme vyráběli 1000 replikací.

3. Simulační studie

Generovali jsme data z Coxova i z AFT modelu, jako základní rozdělení bylo použito Gamma rozdělení $\Gamma(a = 1/100, p = 5)$ a Lognormální rozdělení $LN(\mu = 5, \sigma^2 = 1)$. Použili jsme data s jedním regresorem, jednak spojitým s hodnotami generovanými z $N(3, 1)$ a jednak faktorovým s hodnotami 0 a 1 z $Alt(1/2)$. Hodnoty parametru jsme uvažovali $\beta = 1$ a 2 abychom porovnali vliv síly závislosti. Vždy byly zkoumány dvě varianty, bez cenzorování a s nezávislým náhodným cenzorováním (okolo jedné čtvrtiny dat). Byly použity vzorky velikosti 20, 50, 100, 200, 500 a 1000. Na data simulovaná podle Coxova modelu jsme zkoušeli testy AFT modelu a naopak. Zvolili jsme hladinu $\alpha = 0.05$, vždy jsme nagenerovali 1000 opakování a počítali, kolikrát test na této hladině hypotézu zamítne. Tak získáme empirickou sílu proti dané alternativě. Výsledky viz tabulky 1 a 2.

Výsledky - testy Coxova modelu na datech z AFT

- Empirická síla roste s velikostí výběru vždy vyjma případu Gamma rozdělení s faktorovým regresorem a $\beta = 2$.

Zákl.rozd.	Gamma				Lognormální			
β	1		2		1		2	
Cenzorování	NC	C	NC	C	NC	C	NC	C
Regresor	Spojitý							
20	0.054	0.053	0.071	0.04	0.133	0.103	0.148	0.047
50	0.107	0.09	0.094	0.079	0.291	0.206	0.288	0.204
100	0.234	0.131	0.146	0.112	0.499	0.425	0.423	0.305
200	0.336	0.257	0.249	0.169	0.785	0.614	0.773	0.555
500	0.63	0.552	0.347	0.24	0.995	0.968	0.97	0.84
1000	0.928	0.769	0.557	0.293	1.000	0.997	1.000	0.987
Regresor	Faktorový							
20	0.194	0.26	*0.921*	*0.931*	0.128	0.132	0.225	0.276
50	0.131	0.115	*0.659*	*0.754*	0.166	0.166	0.362	0.300
100	0.245	0.221	0.433	0.532	0.330	0.272	0.562	0.513
200	0.424	0.395	0.203	0.243	0.570	0.508	0.845	0.796
500	0.784	0.696	0.37	0.250	0.904	0.888	0.996	0.990
1000	0.960	0.92	0.661	0.520	0.992	0.984	1.000	1.000

TABULKA 1. Podíl výběrů kde byl Coxův model zamítnut na hladině 0.05 - data z AFT modelu

- Síla vyšší v případech bez cenzorování.
- U lognormálního základního rozdělení je síla vyšší u $\beta = 2$ než u $\beta = 1$, u Gamma rozdělení naopak.
- Při lognormálním rozdělení síla výrazně vyšší při stejném n než při Gamma.

Výsledky - testy AFT na datech z Coxova modelu

- Empirická síla roste s velikostí výběru ve všech případech.
- Síla vyšší v případech bez cenzorování při spojitém regresoru, při faktorovém naopak vyšší s cenzorováním.
- Síla vyšší u $\beta = 2$ než u $\beta = 1$.

Zákl.rozd.		Gamma				Lognormální			
β		1		2		1		2	
Cenzorování		NC	C	NC	C	NC	C	NC	C
Regresor		Spojitý							
20	T	0.012	0.011	0.003	0.007	0.052	0.008	0.001	0.007
	W	0.010	0.008	0.003	0.009	0.052	0.007	0	0.003
50	T	0.012	0.008	0.011	0.026	0.014	0.024	0.014	0.019
	W	0.004	0.004	0.004	0.023	0.009	0.016	0.008	0.018
100	T	0.022	0.024	0.022	0.008	0.065	0.034	0.012	0.041
	W	0.017	0.020	0.014	0.016	0.063	0.019	0.011	0.026
200	T	0.071	0.047	0.063	0.035	0.181	0.092	0.147	0.039
	W	0.047	0.027	0.049	0.025	0.163	0.031	0.186	0.028
500	T	0.086	0.058	0.131	0.055	0.653	0.364	0.829	0.390
	W	0.065	0.023	0.151	0.042	0.632	0.205	0.941	0.276
1000	T	0.294	0.182	0.353	0.237	0.890	0.668	0.988	0.666
	W	0.238	0.114	0.356	0.208	0.888	0.402	0.997	0.504
Regresor		Faktorový							
20	T	0	0.003	0.001	0.010	0	0.017	0.002	0.016
	W	0	0	0	0.007	0	0	0	0.004
50	T	0.002	0.004	0.007	0.019	0	0.012	0.02	0.040
	W	0	0.003	0	0.012	0	0.003	0	0.012
100	T	0	0.010	0.021	0.061	0.005	0.027	0.119	0.224
	W	0	0.005	0.003	0.057	0.005	0.008	0.092	0.066
200	T	0.012	0.023	0.109	0.218	0.050	0.106	0.640	0.676
	W	0.002	0.002	0.051	0.215	0.071	0.022	0.619	0.342
500	T	0.076	0.177	0.663	0.849	0.516	0.608	1	0.994
	W	0.047	0.089	0.542	0.875	0.572	0.296	1	0.960
1000	T	0.361	0.580	0.981	0.998	0.966	0.980	1	1
	W	0.224	0.462	0.971	1	0.965	0.801	1	1

Tabulka 2. Podíl výběrů kde byl AFT model zamítnut na hladině 0.05 - data z Coxova modelu. T - t-test, W - Wilcoxonův test

- Při lognormálním rozdělení síla výrazně vyšší při stejném n než u Gamma. Použitelný počet zamítnutých výběrů je dosažen u Lognormálního rozdělení pro 200 až 500 pozorování, u Gamma pro 500 až 1000.
- Wilcoxonův a t-test srovnatelné u necenzorovaných dat, u cenzorovaných je lepší t-test.
- Celkově nižší síla než u testů Coxova modelu

4. Shrnutí

Aby bylo možné rozlišit, podle kterého z modelů se data chovají, je potřeba v některých případech velký počet pozorování. Testy Coxova modelu vykazují vyšší empirickou sílu než testy pro model zrychleného času. To můžeme přisoudit tomu, že použité metody jsou zde pouze přibližné. Zlepšení by mohlo přinést vyvynutí testů založených na martingalových reziduálech, podobně jako pro Coxův model. Dalším předmětem zkoumání by mohly být i situace s regresory s proměnlivými hodnotami v čase.

Literatura

[1] Buckley J., James I.R.: *Linear regression with censored data*, Biometrika 66, 429–436, 1979.
[2] Cox D.R.: *Regression models and life tables*, J. Roy. Statist. Soc. Ser. B 34, 187–220, 1972.
[3] Fleming T. R., Harrington D. P.: *Counting Processes and Survival Analysis*, Wiley, New York, 1991.
[4] Lin D.Y., Wei L.J., Ying Z.: *Accelerated failure time models for counting processes*, Biometrika 85, 605–618, 1998.
[5] Nikulin M., Bagdonavičius V.: *Accelerated Life Models*, Chapman&Hall, 2002.

Poděkování: Tato práce byla podporována granty GAAV No. IAA101120604 a SVV 261315/2010.

Adresa: MFF UK, KPMS, Sokolovská 83, 186 75 Praha 8, ÚTIA AV ČR, Pod Vodárenskou věží 4, 182 08 Praha 8

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

ON A PROBLEM CONNECTED WITH MIXTURE PARAMETER ESTIMATION

Bobosharif Shokirov

Keywords: Mixture parameter, estimator, expected value, variance.

Abstract: With a sample $X_1, \ldots, X_n$ drawn from a mixture of two distribution functions $F(x)$ and $G(x)$ the paper deals with estimating the mixture parameter θ. It is proposed a method of estimating the mixture parameter. The explicit form of the estimator is given and some of its properties are discussed.

Abstrakt: Tento článek studuje odhad mixujicího parametru θ pomoci vyběru $X_1, \ldots, X_n$ z směsi dvě distribuční funkce $F(x)$ a $G(x)$. Je navhrnut přístup pro odhad parametru směsi. Je dana explicitní forma odhadu a jsou diskutovany některé jeho vlastnosti.

1. The problem

Let $X_1, \ldots, X_n$ be a sample of size n drawn from a known distribution function (d.f.) $H(x)$ of the form

$$H(x) = \theta F(x) + (1-\theta)G(x, \quad x \in [0,1], \quad (\theta \in (0,1)). \tag{1}$$

In (1) $F(x)$ is a known d.f., while d.f. $G(x)$ and a parameter $\theta \in [0,1]$ are unknown. Our aim is to estimate parameter θ, which we call a mixture parameter. Similar problems were considered in [1], [2]. In such formulation without any additional conditions imposed on d.f. $G(x)$ we cannot estimate θ. First of all representation (1) of d.f. $H(x)$ is not unique. With the same d.f. $F(x)$, an appropriate choice of the parameter θ and d.f. $G(x)$ we can construct a representation of d.f. $H(x)$ different from (1). Indeed, let

$$H(x) = \theta_1 F(x) + (1-\theta_1)G_1(x), \quad x \in [0,1], \quad (\theta_1 \in (0,1)), \tag{2}$$

where $G_1(x)$ is some unknown d.f. different from $G(x)$, be a representation (1) of d.f. $H(x)$. Choose $0 < \theta_2 < \theta_1$ and define the function

$$G_2(x) = \frac{(\theta_1 - \theta_2)F(x) + (1-\theta_1)G_1(x)}{1-\theta_2}.$$

For all $x \in \mathbb{R}$ and $\theta_2 < \theta_1$ function $G_2(x)$ has the following properties:

(i) $0 \le G_2(x) \le 1$;
(ii) $\lim_{x\to-\infty} G_2(x) = 0$;
(iii) $\lim_{x\to+\infty} G_2(x) = 1$;
(iv) is a monotonic nondecreasing:

$$G_2'(x) = \frac{(\theta_1 - \theta_2)F'(x) + (1-\theta_1)G'(x)}{1-\theta_2} \ge 0.$$

Properties (i-iv) of the function $G_2(x)$ show that it is a d.f. Now by using d.f. $G_2(x)$ we might have another representation (1) of d.f. $H(x)$

$$H(x) = \theta_2 F(x) + (1 - \theta_2)G_2(x),$$

which is different from (2).

Without loss of generality we can assume that the support of d.f.'s $F(x)$ is the interval $[0,\ 1]$ ($S_F = \text{supp}F(x) = [0,\ 1]$), otherwise by transformation it could be reduced to the interval $[0,\ 1]$. Also we assume that the support of d.f. $G(x)$ is some proper subset of S_F, that is, $S_G = \text{supp}\, G(x) \subset S_F$. Although under the last assumption we still cannot guarantee the identifiability of the model within the whole S_F, it turns out the model to be well-defined and under certain conditions enables to estimate the mixture parameter. The support S_G of d.f. $G(x)$ could be any proper subset of S_F of the forms $[0,\ 1-\delta]$, $[1-\delta,\ 1]$, $[\delta,\ 1-\delta]$ for some $0 < \delta < 1$. For our further discussions we only assume that $S_G = [0,\ 1-\delta]$, for some $\delta > 0$. Also we need d.f.'s $F(x)$ and $G(x)$ to be continuously differentiable.

2. **Estimator of θ and its properties**

Now we proceed to estimation of the mixture parameter in representation (1). As mentioned above without additional conditions on d.f. $G(x)$ representation (1) is not identifiable and hence the estimator of the mixture parameter θ cannot be defined uniquely. Therefore one needs to tighten the class of the unknown d.f.'s $G(x)$ in (1) to the extent which allows one to estimate θ. To be more specific, we assume the following conditions are satisfied:

$$G(x) > F(x), \quad \forall x \in [0,\ 1] \tag{3}$$

and

$$S_G \subset [0,\ 1-\delta], \quad \text{for some} \quad \delta > 0. \tag{4}$$

Under conditions (3) and (4) the estimator of the mixture parameter θ formally could be expressed in the following form

$$\theta^*(x) = \frac{1 - H_n(x)}{1 - F(x)}, \tag{5}$$

where $H_n(x)$ is the empirical distribution function, constructed by the sample $X_1, \ldots, X_n$. In (5) estimator $\theta^*(x)$ is expressed as a function of x for all $x \in [0, 1]$. Below we show that under certain conditions the expected value of $\theta^*(x)$ is a non-increasing function of x in the intersection of the supports of d.f.'s $F(x)$ and $G(x)$: $x \in S_F \cap S_G$ and is constant in the complement of the supports S_G in S_F: $x \in S_F \backslash S_G$. Regarding random variable x as time-variable $\theta^*(x)$ could be considered as random process. In this setting the problem of our interest consists in finding a value of x^* at which $\theta^*(x^*)$ is the optimal estimator of the mixture parameter θ. By optimal estimator of $\theta^*(x)$ we mean those values of the estimator which have minimal variance and are as close as possible to the right border of S_F.

For the expected value of $\theta^*(x)$ the following statement is true.

Theorem 1. *Assume condition (3) holds. Let d.f.'s $F(x)$ and $G(x)$ are continuously differentiable and satisfy the relation*

$$\frac{F'(x)}{1-F(x)} \leq \frac{G'(x)}{1-G(x)}. \tag{6}$$

Then the expected value of the estimator $\theta^*(x)$ is a monotonic non-increasing on the interval $[0,1]$ function and $\theta \leq \mathbb{E}[\theta^*(x)] \leq 1\ \forall x \in [0,1]$.

Důkaz. Taking expectation from (5) yields

$$\mathbb{E}[\theta^*(x)] = \theta + (1-\theta)\frac{1-G(x)}{1-F(x)}. \tag{7}$$

Then the statement follows immediately from

$$\frac{d}{dx}\left(\frac{1-G(x)}{1-F(x)}\right) = \frac{1-G(x)}{1-F(x)}\left(\frac{F'(x)}{1-F(x)} - \frac{G'(x)}{1-G(x)}\right) \leq 0,$$

if only (6) holds. By virtue of (3)

$$0 \leq \frac{1-G(x)}{1-F(x)} \leq 1.$$

Therefore from (7) we get $\theta \leq \mathbb{E}[\theta^*(x)] \leq 1,\ 0 \leq x \leq 1$.

Corollary 1. *If condition (4) holds, then for $x \in (1-\delta, 1]$ $\theta^*(x)$ is an unbiased estimator of θ: $\mathbb{E}[\theta^*(x)] = \theta$.*

Unbiasedness of the estimator $\theta^*(x)$ still is not enough to judge it as the most suitable for our purposes. One needs to know how it deviates with x within the sets $S_F \cap S_G$ and $S_F \backslash S_G$ Therefore, we must clarify the behavior of its variance defined as

$$\sigma^2_{\theta^*(x)} = \mathbb{E}[\theta^*(x)]^2 - [\mathbb{E}\theta^*(x)]^2. \tag{8}$$

As mentioned above, we would like the estimator to be as close as possible to 1 with minimal possible variance. For the variance of $\theta^*(x)$ of the mixture parameter estimator the following statement holds.

Theorem 2. *If conditions (3) and (4) hold, then the variance $\sigma^2_{\theta^*(x)}$ of the estimator $\theta^*(x)$ has the form*

$$\sigma^2_{\theta^*(x)} = \frac{H(x)(1-H(x))}{n(1-F(x))^2} \tag{9}$$

or

$$\sigma^2_{\theta^*(x)} = \frac{A(x;\theta)}{n}\left(\frac{1}{1-F(x)} - A(x;\theta)\right), \tag{10}$$

where

$$A(x;\theta) = \theta\left(1 - \frac{1-G(x)}{1-F(x)}\right) + \frac{1-G(x)}{1-F(x)}$$

Důkaz. Evaluate $\mathbb{E}[\theta^*(x)]^2$. From (5) we have

$$\mathbb{E}[\theta^*(x)]^2 = \frac{1}{(1-F(x))^2}\mathbb{E}[1-2H_n(x)+H_n^2(x)]. \tag{11}$$

$H_n(x) = \frac{1}{n}\sum_{i=1}^n I_{\{X_i<x\}}$. Therefore

$$H_n^2(x) = \frac{1}{n^2}\sum_{i=1}^n I_{\{X_i<x\}} + \frac{1}{n^2}\sum_{i\neq j}^n I_{\{\min(X_i,X_j)<x\}}. \tag{12}$$

Due to $\mathbb{E}[H_n(x)] = H(x)$ from equation (12) we obtain

$$\mathbb{E}[H_n^2(x)] = H^2(x) + \frac{1}{n}H(x)(1-H(x)). \tag{13}$$

By virtue of (13) from (7) we have

$$\mathbb{E}[\theta^*(x)]^2 = \frac{(1-H(x))^2 + \frac{1}{n}H(x)(1-H(x))}{(1-F(x))^2}$$

By using the last relation from (8) and (5) we obtain

$$\sigma^2_{\theta^*(x)} = \frac{H(x)(1-H(x))}{n(1-F(x))^2}.$$

Corollary 2. (a) *If condition (4) holds, then*

$$\sigma^2_{\theta^*(x)} = \frac{\theta}{n}\left(\frac{1}{1-F(x)} - \theta\right), \quad \textit{for} \quad 1-\delta < x \leq 1. \tag{14}$$

(b) *If condition (3) holds, then*

$$\begin{aligned}\sigma^2_{\theta^*(x)} &= \frac{1}{n(1-F(x))}\left[\theta\left(1-\frac{1-G(x)}{1-F(x)}\right)+\frac{1-G(x)}{1-F(x)}\right] - \\ &- \frac{1}{n}\left[\theta\left(1-\frac{1-G(x)}{1-F(x)}\right)+\frac{1-G(x)}{1-F(x)}\right]^2, \quad \textit{for} \quad 0 \leq x \leq 1-\delta.\end{aligned}$$

Theorem 3. *Let conditions (3) and (4) be satisfied. Then if (6) also holds, then the variance $\sigma^2_{\theta^*(x)}$, defined in (10) is a monotonic nondecreasing function of x for all $x \in [0,1]$.*

Důkaz. We first show that if (6) holds then $A(x;\theta)$ is non-increasing with x for $0 \leq x \leq 1$. Calculate the first derivative of $A(x;\theta)$ with respect to x. We obtain

$$\frac{d}{dx}(A(x;\theta)) = \theta\frac{1-G(x)}{1-F(x)}\left(\frac{F'(x)}{1-F(x)} - \frac{G'(x)}{1-G(x)}\right) \leq 0,$$

if only

$$\frac{F'(x)}{1-F(x)} \leq \frac{G'(x)}{1-G(x)}.$$

Now take two arbitrary points $x_1,\ x_2 \in [0,1]$ such that $x_1 < x_2$. Show that $\sigma^2_{\theta^*(x_2)}/\sigma^2_{\theta^*(x_1)} \geq 1$. We have

$$\frac{\sigma^2_{\theta^*(x_2)}}{\sigma^2_{\theta^*(x_1)}} = \frac{H(x_2)(1-H(x_2))}{H(x_1)(1-H(x_1))}\left[\frac{1-F(x_1)}{1-F(x_2)}\right]^2. \tag{15}$$

For $x_1 < x_2,\ H(x_1) \leq H(x_2)$ and $1-F(x_1) \geq 1-F(x_2)$. Therefore from (15) we obtain

$$\frac{\sigma^2_{\theta^*(x_2)}}{\sigma^2_{\theta^*(x_1)}} \geq \frac{1-H(x_2)}{1-H(x_1)}\frac{1-F(x_1)}{1-F(x_2)} = \frac{1-H(x_2)}{1-F(x_2)}\frac{1-F(x_1)}{1-H(x_1)}. \tag{16}$$

Since

$$\frac{1-H(x)}{1-F(x)} = \frac{1}{1-F(x)} - A(x;\theta),$$

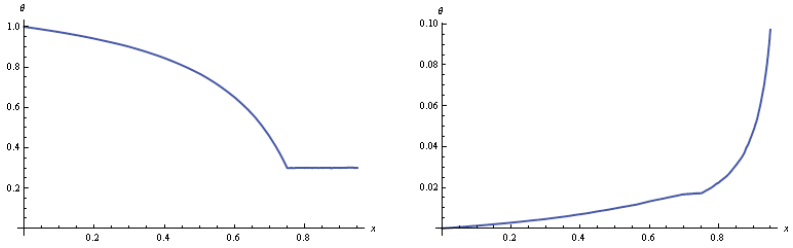
then from the right hand side of (16) we get

$$\frac{1-H(x_2)}{1-F(x_2)}\frac{1-F(x_1)}{1-H(x_1)} = \frac{1-A(x_2;\theta)(1-F(x_2))(1-F(x_1))}{1-A(x_1;\theta)(1-F(x_1))(1-F(x_2))}.$$

Function $A(x;\theta)$ is non-increasing with x: $A(x_1;\theta) \geq A(x_2;\theta)$, therefore $1-A(x_2;\theta)C(x_1;x_2) \geq 1-A(x_1;\theta)C(x_1;x_2)$, where $C(x_1;x_2) = (1-F(x_1))(1-F(x_2)) \geq 0$ and hence $\frac{\sigma^2_{\theta^*(x_2)}}{\sigma^2_{\theta^*(x_1)}} \geq 1$.

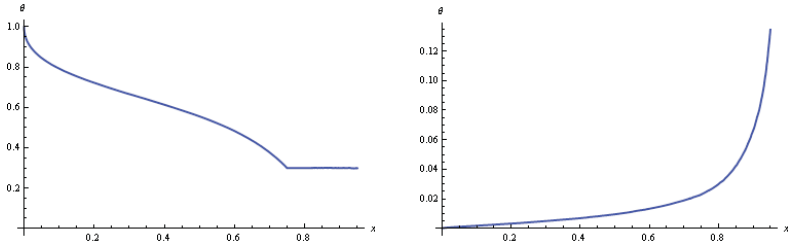
3. **Simulation study**

Simulated data from different distributions show that in condition (6) holds, then the expected value of the estimator θ decreases with x, while its variance increases with x for all $x \in S_F$. Although monotonicity keeps direction within the whole set of S_F we observe a change in the behavior of both the expected value and the variance of the $\theta^*(x)$ once the random variable x runs beyond the $S_F \cap S_G$. Uniform distribution is the most suitable case to our theoretical explanation (Theorems 1 and 3). When one of the distributions in the mixture is different from the uniform we observe some deviation from theoretical explanation but, in general, estimator of $\theta^*(x)$ has a behavior very similar to the uniform case: having decreased with x in the interval $[0,\ 1-\delta]$, the support of the alternative distribution, we observe some stabilization of the expected value of $\theta^*(x)$ once random variable x crosses the right border of the support. Here by stabilization we mean that oscillations of the expected value are not that high and could be negligible. But for $x > 1-\delta$ the variance is strictly increasing. If in the uniform case the expected value of $\theta^*(x)$ remains constant for $x > 1-\delta$, in other cases it behaves as a function of bounded variation or slow change. Thus, the right border of the support of $G(x)$ can serve as a lower bound for the estimator of θ. Since we would like $\theta^*(x)$ to be as close as possible to 1, we can choose $\theta^*(x)$ greater than $1-\delta$ with the minimal standard deviation. Some simulated data, which illustrate the behavior of the expected value and the variance of the estimator $\theta^*(x)$ are shown in Figures 1



(a) Expected value (b) Variance

FIGURE 1. The expected value and the variance of the mixture parameter estimator $\theta^*(x)$, calculated by the sample generated from the mixture of d.f.'s $F(x) = U[0,\ 1]$ and $G(x) = U[0,\ 0.75]$.



(a) Expected Value (b) Variance

FIGURE 2. The expected value and the variance of the mixture parameter estimator $\theta^*(x)$, calculated by the sample generated from the mixture of d.f.'s $F(x) = U[0, 1]$ and $G(x) = B[0.5, 1]$.

and 2. In both cases the true value of the mixture parameter is 0.3 and the number of simulations is 10000.

Figure 1 presents the behavior of the expected value and the variance of $\theta^*(x)$, derived from the mixture of two uniform distributions on the intervals [0, 1] and [0, 0.75]. Here we generated a sample of size 100 from the mixture of these two distributions and with the mixture parameter θ and calculated the expected value and variance of the estimator $\theta^*(x)$.

In Figure 2 are shown graphs of the expected value and the variance of $\theta^*(x)$, derived from the mixture of the uniform on the interval [0, 1] distribution $U[0,\ 1]$ and beta distribution $B[0.5,\ 1]$ on the interval [0, 0.75]. Here calculation of the expected value and the variance of the estimator $\theta^*(x)$ are based on the sample of size 100, generated from a mixture of these two distributions.

4. Summary

We presented a method of estimating the mixture parameter from the mixture of two distribution functions, where one of the d.f.'s is unknown. Having imposed some restrictions on the components of the mixture we derived an explicit form of the estimator as a random process. We studied the behavior of the expected value and the variance of the estimator; we showed that this is an unbiased estimator in S_G and its expected value is monotonic non-increasing, while its variance is monotonic nondecreasing with random variable x. We illustrated our results by simulating data from different d.f.'s. Simulations confirms that the mixture of two uniform distributions (with different supports) is the most suitable to our theoretical explanation and the mixture from other distributions do not have large deviation from uniform cases.

References

[1] Meinhausen N., Rice J.P. (2006) *Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses.* The Annals of Statistics, **34**, 373 – 393.

[2] Wu W.B. (2008) *On false discovery control under dependence.* The Annals of Statistics, **36**, 364 – 380.

Acknowledgement: The author is sincerely grateful to professor Klebanov L.B. for his generous help. This work was supported by the grant SVV 261315/2010.

Address: MFF UK, KPMS, Sokolovská 83, 186 75 Praha 8 – Karlín

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

DIFÚZE V UZAVŘENÉ OBLASTI

Jakub Staněk, Josef Štěpán

Klíčová slova: Difúze s odrážející hranicí, difúze s pohlcující hranicí, difúze v omezené oblasti.

Abstrakt: Předpokládejme funkci $f \in \mathcal{C}^2(\mathbb{R}^n)$, konstantu c a definujme oblast $K = [f \leq c]$. V článku jsou prezentovány podmínky zaručující, že difúze určená sochastickou diferenciální rovnicí $dX_t = b(X_t)dt + \sigma(X_t)dB_t$ neopustí oblast K. Dále je zkoumáno, za jakých podmínek je hranice oblasti $S = \partial K$ odrážející a kdy pohlcující.

Abstract: Consider a map $f \in \mathcal{C}^2(\mathbb{R}^n)$, a constant c and define the region $K = [f \leq c]$. Further, a diffusion given by stochastic differential equation $dX_t = b(X_t)dt + \sigma(X_t)dB_t$ is considered. The paper presents a conditions for having the diffusion inside K and it is also studied when the boundary $S = \partial K$ is absorbing and reflecting, respectively.

1. Úvod

Předpokládejme difúzi danou rovnicí

$$dX_t = b(X_t)dt + \sigma(X_t)dB_t, \tag{1}$$

kde B_t je n-dimensionální Wienerův proces a $b(x) = (b_1(x), \ldots, b_n(x))^T$ a $\sigma(x) = (\sigma_{ij}(x)_{1\leq i,j\leq n})$ jsou borelovské funkce. Dále uvažujme oblast K určenou

$$K := \{x : f(x) \leq c\}, \tag{2}$$

kde $f \in \mathcal{C}^2(\mathbb{R}^n)$ a $c \in \mathbb{R}$.

Příspěvek se zabývá podmínkami, které zaručují, že difúze určená rovnicí (1) neopustí oblast K. Dále budou prezentovány podmínky zaručující, že hranice $S := \partial K = \{x : f(x) = c\}$ je odrážející, respektive pohlcující.

Připomeňme, že spojitý $\mathcal{F}_t$-adaptovaný proces $X = (X_1, \ldots, X_n)$ řeší rovnici (1), pokud

$$X_t = X_0 + \int_0^t b(X_s)ds + \int_0^t \sigma(X_s)dB_s \quad \text{platí skoro jistě} \quad \forall t \geq 0,$$

kde $\mathcal{F}_t$ je zúplněná kanonická filtrace Wienerova procesu B_t. Aby byla pravá strana předchozího výrazu dobře definována, předpokládáme, že platí

$$\int_0^t |\sigma(X_s)|^2 + |b(X_s)|ds < \infty \quad \text{s.j. pro libovolné} \quad t \geq 0.$$

Dále budeme používat následující značení:

- X^x je řešením rovnice (1) s počáteční podmínkou $X_0 = x$, $x \in \mathbb{R}^n$,
- $K^e = \mathbb{R}^n \setminus \text{int}(K) = (\mathbb{R}^n \setminus K) \cup S$.

2. Difúze v oblasti K

Na úvod této části uvedeme již známé výsledky prezentovány v [1] a [2]. Nejprve však zadefinujeme používané pojmy.

Uvažujme nyní pevné řešení X rovnice (1). Řekneme, že hranice S je **nedosažitelná** pro řešení X, pokud

$$P[X_t^x \in S \quad \text{pro nějaké} \quad t \geq 0] = 0, \quad \forall x \notin S,$$

hranice S je **pohlcující** hranice pro X, pokud vně P-nulové množiny platí

$$X_t \in S \Rightarrow X_{t+s} \in S, \quad s \geq 0, t \geq 0$$

a hranice S bude **odrážející** hranicí pro X, pokud vně P-nulové množiny X neopustí oblast K a zároveň neexistuje dvojice časů $0 \leq u < v < \infty$ taková, že $X_s \in S$ pro všechna $s \in (u, v)$.

Uvažujme nyní omezenou, uzavřenou oblast K s $\mathcal{C}^3$-spojitou hranicí $S = \partial K$ a předpokládejme, že koeficienty b a σ splňují následující podmínky:

- Existuje konstanta C taková, že pro všechna $x \in \mathbb{R}^n$

$$\sum_{i=1}^{n} |b_i(x)| + \sum_{i,j=1}^{n} |\sigma_{i,j}(x)| \leq C(1 + |x|), \tag{3}$$

- pro každé $R > 0$ existuje konstanta C_R taková, že

$$\sum_{i=1}^{n} |b_i(x) - b_i(y)| + \sum_{i,j=1}^{n} |\sigma_{i,j}(x) - \sigma_{i,j}(y)| \leq C_R|x - y| \tag{4}$$

  platí pro všechna $|x| < R$ a $|y| < R$.

Dále označme $v = (v_1, ..., v_n)$ vnější normálový vektor k hranici S a funkci $\rho(x) = d(x, K)$ ($d(x, K)$ značí vzdálenost bodu x od množiny K), kterou uvažujeme pouze na množině $K^e \cup S$. Pak můžeme vyslovit následující větu, která stanovuje podmínky, za kterých je hranice S nedosažitelná a pohlcující.

Věta 1 *Nechť pro všechny $x \in S$ platí*

$$\sum_{i,j=1}^{n} a_{ij} v_i v_j = 0 \tag{5}$$

a

$$\sum_{1}^{n} b_i v_i + \frac{1}{2} \sum_{i,j=1}^{n} a_{ij} \frac{\partial^2 \rho}{\partial x_i \partial x_j} = 0, \tag{6}$$

kde $a_{ij} = \sum_{k=1}^{n} \sigma_{ik}\sigma_{jk}$.
Pak S je nedosažitelná a pohlcující hranice pro řešení rovnice (1).

Důkaz a další podrobnosti lze nalézt v kapitole 12 v [2].

Poznámka Podmínka lipschitzovskosti (4) je však velmi omezující a pro některé aplikace nevhodná, navíc nám nedovolí zkonstruovat rovnici (1) tak, aby její řešení startovalo z vnitřku oblasti K a dorazilo až na její hranici S,

což může být pro nějaké modely užitečné. Proto v následující části ukážeme jemnější podmínky, za kterých řešení X rovnice (1) neopustí oblast K, a které otvírají cestu ke konstrukci rovnic, jejichž řešení dorazí až k hranici oblasti.

Vraťme se nyní k oblasti K zadané vztahem (2), tedy

$$K := \{x : f(x) \leq c\}.$$

Pak

$$K^e := \{x : f(x) \geq c\} \quad \text{a} \quad S = \partial K = \partial K^e = \{x : f(x) = c\}.$$

Uvědomme si, že takto definovaná oblast K nemusí být omezená, což je v mnoha aplikacích užitečné.

Označíme-li $Z_t = f(X_t)$ pak dostáváme

$$X_t \in K \Leftrightarrow Z_t \leq c,$$

čímž jsme naši úlohu převedli na jednodimenzionální problém.

Použijeme-li Itôovu formuli (viz například Theorem 32.8, str. 60 v [3]) na proces Z, dostáváme

$$dZ_t = df(X_t) = Lf(X_t)dt + dM_t,$$

kde

$$\begin{aligned} Lf(x) &= \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} b_i(x) + \frac{1}{2} \sum_{i,j=1}^{n} \frac{\partial^2 f}{\partial x_i \partial x_j}(x) a_{ij}(x) \\ &= \mathrm{grad} f(x)^T \cdot b(x) + \frac{1}{2} \mathrm{tr}\left(f''(x) \cdot a(x)\right), \\ dM_t &= \mathrm{grad} f(X_t)^T \cdot \sigma(X_t) dB_t, \\ a(x) = \sigma(x)\sigma(x)^T, &\quad \mathrm{grad} f(x) = \left(\frac{\partial f}{\partial x_1}(x), ..., \frac{\partial f}{\partial x_n}(x)\right) \end{aligned}$$

a

$$f''(x) = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}(x)_{1, \leq i,j \leq n}\right).$$

Chceme-li, aby proces Z_t nepřekročil hodnotu c, pak je v okolí hranice S třeba utlumit difúzní koeficient $\mathrm{grad} f(x)^T \cdot \sigma(x)$ a zařídit, aby $Lf(x) \leq 0$. Tato úvaha je formulována v následujícím lemmatu.

Lemma 1 *Nechť existuje otevřené okolí G hranice S takové, že pro všechna $x \in G \cap K^e$ platí*

$$\mathrm{grad} f(x)^T \cdot a(x) \cdot \mathrm{grad} f(x) = 0 \tag{7}$$

a

$$Lf(x) \leq 0. \tag{8}$$

Pak $X \in K$ skoro jistě pro libovolné řešení X rovnice (1) s počáteční podmínkou $X_0 = x_0$, kde $x_0 \in K$.

Důkaz: Nechť X je řešení rovnice (1) s počáteční podmínkou $X_0 = x_0 \in K$, pak vně P-nulové množiny N platí

$$f(X_v) - f(X_u) = \int_u^v Lf(X_s)ds + \int_u^v \operatorname{grad}^T f(X_s) \cdot \sigma(X_s)dB_s \tag{9}$$

pro všechny $0 < u < v < \infty$.

Nejprve označme $N_r = [f(X_r) > c]$ pro $r \geq 0$ a ukažme, že $P(N_r) = 0$ pro všechna $r \in \mathbb{Q}^+$.

Předpokládejme $\omega \in N_r$ takové, že $\omega \notin N$ a označme

$$u = u(\omega) = \sup\{s \leq r : X_s(\omega) \in K\}.$$

Pak existuje čas v, takový, že $u < v = v(\omega) < r$ a

$$(X_u, X_v) = \{X_s, s \in (u, v)\} \subset G \cap K^e, \quad f(X_u) = c, \quad f(X_v) > c.$$

Jelikož $\omega \notin N$, pak z (9) dostáváme $f(X_v) - f(X_u) = f(X_v) - c \leq 0$, čímž jsme došli ke sporu. Tedy $N_r \subset N$, a proto $P(N_r) = 0$.

Zbytek důkazu plyne ze spojitosti trajektorií procesu X a spočetnosti množiny $r \in \mathbb{Q}^+$. □

Lemma 1 nás motivovalo k definici **hraniční rovnice**, to jest rovnice, jejíž řešení se pohybuje pouze po hranici S.

Řekneme, že rovnice (1) je **hraniční rovnice** pro hranici S, pokud existuje otevřené okolí $G \supset S$ takové, že

$$Lf(x) = 0 \tag{10}$$

a

$$\operatorname{grad} f(x)^T \cdot \sigma(x) = 0 \tag{11}$$

platí pro všechna $x \in G$.

Poznámka Aplikací Lemmatu 1 na dvojice (f, c) a $(-f, -c)$ lze ukázat, že libovolné řešení hraniční rovnice X s počáteční podmínkou $X_0 = x_0 \in S$ zůstává skoro jistě na hranici S.

Nyní se zabývejme otázkou, za jakých podmínek bude hranice S odrážející, respektive pohlcující hranicí pro řešení X.

Lemma 2 *Nechť X je řešením rovnice (1) a nechť existuje otevřené okolí G hranice S, takové, že podmínky (7) a (8) platí pro všechna $x \in G$. Dále nechť $Lf(x) < 0$ pro všechna $x \in S$. Pak S je odrážející hranicí pro proces X.*

Důkaz: Z Lemmatu 1 vyplývá, že řešení X neopustí oblast K, zbývá tedy dokázat, že v hranici S nesetrvá.

Postupujme stejně jako v důkazu Lemmatu 1. Nechť N je P-nulová množina taková, že (9) platí vně N. Předpokládejme, že $\omega \notin N$ a existuje dvojice časů $u < v$ takových, že $X_s(\omega) \in S$ pro všechna $s \in (u, v)$. Pak

$$f(X_v) - f(X_u) = 0 = \int_u^v Lf(X_s)ds,$$

tedy jsme došli ke sporu, čímž je důkaz hotov. □

Lemma 3 *Předpokládejme rovnici (1) a hraniční rovnici*

$$dX_t = b^*(X_t)dt + \sigma^*(X_t)dB_t. \tag{12}$$

Nechť rovnice (1) má slabé, jednoznačné řešení $(X^x, x \in \mathbb{R}^n)$ *a lokálně omezené koeficienty* b *a* σ*. Dále předpokládejme, že hraniční rovnice (12) má slabé řešení, pro libovolnou počáteční podmínku* $x \in S$ *a platí následující rovnosti:*

$$b(x) = b^*(x) \quad \sigma(x) = \sigma^*(x) \quad \forall x \in S.$$

Pak

$$P[X^x \in K] = 1 \quad \forall x \in K$$

a hranice S *je pohlcující hranicí.*

Důkaz: Uvažujme pevné $x \in S$ a slabé řešení Y rovnice (12) s počáteční podmínkou $Y_0 = x$. Pak dle Poznámky 1 dostáváme $P[Y \in S] = 1$, a tedy Y je rovněž slabým řešením rovnice (1) s počáteční podmínkou $Y_0 = X$. Ze slabé jednoznačnosti rovnice (1) dostaváme $P[X^x \in S] = 1$ pro libovolné řešení X^x rovnice (1).

Z lokální omezenosti koeficientů rovnice (1) dostáváme silnou markovskou vlastnost řešení X této rovnice. Položme

$$X_0 = \tilde{x} \in K \quad \text{a} \quad \lambda := \inf\{t \geq 0 : X_t \in S\},$$

pak z markovské vlastnosti procesu X dostáváme

$$P^{\tilde{x}}[X_{\lambda+t} \in S \quad \forall t \geq 0, \lambda < \infty] = P^{\tilde{x}}[\lambda < \infty].$$

Tedy hranice S je pohlcující hranicí pro proces X. □

Poznámka Při podrobnějším zkoumání lze ukázat, že podmínky (5) a (6) použité ve větě 1 odpovídají podmínkám (11) a (10), které byly využity v předchozím lematu.

3. Závěr

Představili jsme podmínky zaručující, že řešení X rovnice (1) neopustí oblast K a podmínky zaručující pohlcující, respektive odrážející hranici S tak, aby nebylo nutné předpokládat lipschitzovskost koeficientů b a σ. Zároveň byla oblast K zvolena tak, aby prezentované výsledky bylo možno aplikovat například na oblast $K = \{x = (x_1, ..., x_n) \in \mathbb{R}^n : x_1 \geq 0\}$.

Nezodpovězenou otázkou zůstává, jak obecně zaručit existenci a slabou jednoznačnost řešení rovnice (1), která je předpokládána v Lemmatu 3, tak,

aby nebyla splněna podmínka lipschitzovskosti (4). Další otázkou zůstává, jak obecně za předpokladu splnění podmínek (7) a (8) volit koeficienty b a σ tak, aby řešení X rovnice (1) startující z vnitřku oblasti K dorazilo v konečném čase na hranici S s kladnou pravděpodobností.

Literatura

[1] Friedman A.(1976) *Stochastic Differential Equations and Applications-volume 1*, Academic press, INC., New York.
[2] Friedman A.(1976) *Stochastic Differential Equations and Applications-volume 2*, Academic press, INC., New York.
[3] Rogers L. C. G., Williams D. (2000) *Diffusions, Markov Processes and Martingales-vilume 2 Itô Calculus*, Cambridge university press, Cambridge.

Poděkování: Tato práce byla podporována projektem MŠMT 1M06047 Centrum pro jakost a spolehlivost výroby a výzkumným záměrem MSM 0021620839.

Adresa: J. Staněk, Ústav technické matematiky, Fakulta strojní, ČVUT v Praze, Karlovo náměstí 13, 121 35 Praha 2; J. Štěpán, MFF UK, KPMS, Sokolovská 83, 186 75 Praha 8 – Karlín

E-mail: jakub.stanek@fs.cvut.cz, stepan@karlin.mff.cuni.cz

DVOUSTUPŇOVÉ NÁHODNÉ VÝBĚRY VE VÝBĚROVÝCH ŠETŘENÍCH

Michaela Šedová, Michal Kulich

Klíčová slova: Výběrová šetření, dvoustupňový náhodný výběr.

Abstrakt: V klasické teorii výběrových šetření jsou předmětem studia parametry charakterizující konečnou populaci, jako např. úhrn nebo průměr N pevných hodnot. Někdy je však vhodnější považovat pozorování za náhodné veličiny a zároveň brát v úvahu, že není k dispozici prostý náhodný výběr. Uvažujeme dvoustupňové výběrové schéma, kde jsou nejprve vybrány domácnosti a poté je z každé domácnosti náhodně určen jeden člen a zařazen do studie. Popisujeme odhad střední hodnoty, který toto výběrové schéma zohledňuje, jeho vlastnosti a porovnáváme ho s odhadem získaným z bernoulliovského výběru.

Abstract: In classical sampling theory the targets of inference are finite population parameters, e.g. total or mean of N fixed values. However, in some situations it is more appropriate to consider observations as realizations of random variables and in the same time to take into consideration that simple random sample is not available. We deal with two-stage sampling where households are selected at random and then one eligible member is sampled from each household and included in the study. We describe an estimator of expectation, which takes into account the sampling scheme, present its properties and compare this estimator with the estimator based on Bernoulli sample.

1. Výběr z domácností

V kontextu výběrových šetření se zpravidla zabýváme parametry, které charakterizují konečnou populaci (např. úhrn nebo průměr N pevných hodnot). Někdy však může nastat situace, kdy bychom rádi výsledky zobecnili na jiné populace, nebo i tutéž populaci v jiném čase. V takovém případě je vhodné chápat naše pozorování jako realizace náhodných veličin (viz např. [1] a [2]). Takto přistupují k datům "klasické"statistické metody. Ty však předpokládají, že je k dispozici prostý náhodný výběr, což v kontextu výběrových šetření často není možné.

Uvažujme dvoustupňové výběrové schéma, kde jsou nejprve (se stejnou pravděpodobností) vybrány domácnosti a poté je ze všech členů dané domácnosti náhodně určen jeden a zařazen do studie. V teorii výběrových šetření tento postup patří mezi schémata nazývaná „Two-stage element sampling" [4]. V daném kontextu však není možné stanovit rozptyl odhadů zkoumaných paramerů, neboť je-li z každé domácnosti („clusteru") vybrán pouze jeden zástupce, nelze určit rozptyl v jednotlivých domácnostech. Ani „klasické" metody nemůžeme aplikovat beze změny (např. odhadnout střední

hodotu průměrem pozorování), neboť máme k dispozici výběr, který nadhodnocuje počet členů malých domácností a naopak podhodnocuje zastoupení domácností velkých.

Proto je potřebné zvolit analýzu dat, která kombinuje oba tyto přístupy. Ukážeme odhad střední hodnoty, který zohledňuje popsané výběrové schéma, a uvedeme jeho vlastnosti. Formulujeme podobnou úlohu s „bernoulliovským“ výběrovým schématem a odhady získané z obou výběrů porovnáme.

2. Odhad střední hodnoty

Předpokládejme, že máme n domácností (prostý náhodný výběr z nekonečné populace, resp. z rozdělení). Nechť náhodná veličina M_i představuje počet členů v i-té domácnosti. Její hustotu značme $f(m)$ a střední hodnotu μ. Celkový počet jedinců označme $N = \sum_{i=1}^{n} M_i$.

Nechť je dále Y_{ir} sledovaná náhodná veličina pro r-tého člena i-té domácnosti a ξ_{ir} náhodná veličina, pro kterou platí:

$$\xi_{ir} = \begin{cases} 1 & \text{je-li } r\text{-tý jedinec z } i\text{-té domácnosti zahrnut do výběru} \\ 0 & \text{jinak,} \end{cases}$$

a tedy $\pi_{ir} = \mathrm{E}\,\mathrm{E}\,(\xi_{ir}|M_i) = \frac{1}{M_i}$ je pravděpodobnost zahrnutí r-tého člena z i-té domácnosti do výběru, je-li dáno M_i. Veličina Y_{ir} je pozorovaná pouze pro jedince z výběru, tj. pro $\xi_{ir} = 1$.

Y_{ir} v i-té domácnosti jsou nezávislé stejně rozdělené (iid) náhodné veličiny, jejichž rozdělení závisí na velikosti domácnosti M_i a náhodném parametru $\boldsymbol{b_i}$. Je dáno hustotou

$$f(y|m, \boldsymbol{b}).$$

Střední hodnotu Y_{ir} v i-té domácnosti značíme

$$\theta_i = \int y f(y|m, \boldsymbol{b})\, dy.$$

Hustota veličiny Y_{ir} v jakékoliv domácnosti o velikosti m je

$$f(y|m) = \int f(y|m, \boldsymbol{b}) f(\boldsymbol{b}|m) d\boldsymbol{b},$$

kde $f(\boldsymbol{b}|m)$ je hustota parametru $\boldsymbol{b_i}$, je-li dáno M_i. Hustota veličiny Y_{ir} je tudíž

$$\begin{aligned} f(y) &= \frac{\int m f(y|m) f(m) d\mu(m)}{\int m f(m)\, d\mu(m)} = \frac{1}{\mu} \int m \Big[\int f(y|m, \boldsymbol{b}) f(\boldsymbol{b}|m)\, d\boldsymbol{b} \Big] f(m)\, d\mu(m) \\ &= \frac{1}{\mu} \iint m f(y|m, \boldsymbol{b}) f(m, \boldsymbol{b})\, d\boldsymbol{b} d\mu(m), \end{aligned}$$

kde $f(m, \boldsymbol{b})$ je sdružená hustota M_i a $\boldsymbol{b}_i$. Pro střední hodnotu veličiny Y_{ir} tedy platí

$$\theta = \mathrm{E}\, Y_{ir} = \frac{1}{\mu} \iint m \Big[\int y f(y|m, \boldsymbol{b}) dy \Big] f(m, \boldsymbol{b})\, d\boldsymbol{b} d\mu(m) = \frac{1}{\mu} \mathrm{E}\, M_i \theta_i. \tag{1}$$

Odhad parametru θ definujeme:

$$\hat{\theta} = \frac{\sum_{i=1}^{n} \sum_{r=1}^{M_i} \frac{\xi_{ir}}{\pi_{ir}} Y_{ir}}{\sum_{i=1}^{n} M_i}.$$

Tvrzení 1. *Nechť $M_i, i = 1 \dots n$, jsou iid náhodné veličiny. Nechť také $(Y_{ir}, \xi_{ir}), i = 1 \dots n$ a $r = 1 \dots M_i$, jsou stejně rozdělené náhodné veličiny. Nechť dále $(Y_{i1}, Y_{i2} \dots Y_{iM_i})$ a rovněž $(\xi_{i1}, \xi_{i2} \dots \xi_{iM_i})$ jsou nezávislé náhodné vektory, $\sum_{r=1}^{M_i} \xi_{ir} = 1$ a ξ_{ir} je nezávislé s Y_{ir}, je-li dáno M_i. Předpokládejme, že* $\operatorname{var} Y_{ir} < \infty$. *Potom*

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \Sigma_{\hat{\theta}}),$$

kde

$$\Sigma_{\hat{\theta}} = \frac{1}{\mu} \mathrm{E}\, M_i (Y_{ir} - \theta)^2.$$

Náznak důkazu. Máme-li $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} M_i$, Taylorovým rozvojem $\frac{1}{\hat{\mu}}$ kolem $\frac{1}{\mu}$ dostaneme

$$\sqrt{n} \left(\frac{1}{\hat{\mu}} - \frac{1}{\mu} \right) = -\frac{1}{\sqrt{n}\mu} \sum_{i=1}^{n} \left(\frac{M_i}{\mu} - 1 \right) + o_p(1).$$

Pak

$$\begin{aligned}
\sqrt{n}(\hat{\theta} - \theta) &= \frac{1}{\sqrt{n}\hat{\mu}} \sum_{i=1}^{n} \sum_{r=1}^{M_i} \frac{\xi_{ir}}{\pi_{ir}} Y_{ir} - \sqrt{n}\theta \\
&= \frac{1}{\sqrt{n}\mu} \sum_{i=1}^{n} M_i \sum_{r=1}^{M_i} \xi_{ir} Y_{ir} - \frac{1}{\sqrt{n}} \theta \sum_{i=1}^{n} \left(\frac{M_i}{\mu} - 1 \right) - \sqrt{n}\theta + o_p(1) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Q_i + o_p(1),
\end{aligned}$$

kde $Q_i = \frac{1}{\mu} M_i \big(\sum_{r=1}^{M_i} \xi_{ir} Y_{ir} - \theta \big)$ jsou iid náhodné veličiny. Podle centrální limitní věty $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(\mathrm{E}\, Q_i, \operatorname{var} Q_i)$. Platí

$$\mathrm{E}\, Q_i = \mathrm{E}\, (\mathrm{E}\, (Q_i | i)) = \frac{1}{\mu} \mathrm{E}\, M_i \theta_i - \theta = 0.$$

Podle (1),

$$\operatorname{var} Q_i = \frac{1}{\mu^2} \mathrm{E} \left[M_i \Big(\sum_{r=1}^{M_i} \xi_{ir} Y_{ir} - \theta \Big) \right]^2 = \frac{1}{\mu} \mathrm{E}\, M_i (Y_{ir} - \theta)^2.$$

□

3. Porovnání s bernoulliovským výběrem

Zajímá nás, zda je rozptyl odhadu parametru θ z výběru domácností srovnatelný s rozptylem odhadu získaného na základě bernoulliovského výběru. Představme si tedy situaci, že bychom měli *pevné* N jedinců, z nichž by každý byl nezávisle na ostatních vybrán s pravděpodobností $1/m$, kde m je počet členů domácnosti, do které patří. Nyní tedy velikost výběru bude náhodná, se střední hodnotou n.

Je nutné závést jiné značení. Nechť Y_j je sledovaná náhodná veličina pro j-tého jedince a ξ_j náhodná veličina, pro kterou platí

$$\xi_j = \begin{cases} 1 & \text{je-li } j\text{-tý jedinec zahrnut do výběru} \\ 0 & \text{jinak,} \end{cases}$$

a tedy $\pi_j = \mathrm{E}\,(\xi_j | M_j) = \frac{1}{M_j}$ je pravděpodobnost zahrnutí j-tého jedince do výběru, kde M_j je velikost domácnosti, ze které pochází. Všimněme si, že π_j je náhodná veličina. Veličina Y_j je pozorovaná pouze pro jedince z výběru, tj. pro $\xi_j = 1$.

Odhad parametru θ definujeme:

$$\tilde{\theta} = \frac{\sum_{j=1}^{N} \frac{\xi_j}{\pi_j} Y_j}{\sum_{j=1}^{N} \frac{\xi_j}{\pi_j}}.$$

Tvrzení 2. *Nechť* $(Y_j, \xi_j, M_j), j = 1 \dots N$, *jsou iid náhodné veličiny a* ξ_j *je nezávislé s* Y_j, *je-li dáno* M_j. *Předpokládejme, že* $\operatorname{var} Y_j < \infty$. *Potom*

$$\sqrt{N}(\tilde{\theta} - \theta) \xrightarrow{d} N(0, \Sigma_{\tilde{\theta}}),$$

kde

$$\Sigma_{\tilde{\theta}} = \mathrm{E}\, M_i (Y_i - \theta)^2.$$

Náznak důkazu. Máme-li $\hat{N} = \sum_{i=1}^{N} \frac{\xi_i}{\pi_i}$, Taylorovým rozvojem $\frac{1}{\hat{N}}$ kolem $\frac{1}{N}$ dostaneme

$$\sqrt{N}\left(\frac{1}{\hat{N}} - \frac{1}{N}\right) = -N^{-\frac{3}{2}}\left(\hat{N} - N\right) + o_p(1) = -N^{-\frac{3}{2}} \sum_{i=1}^{N} \left(\frac{\xi_i}{\pi_i} - 1\right) + o_p(1).$$

Pak

$$\begin{aligned} \sqrt{N}(\tilde{\theta} - \theta) &= \sqrt{N}\Big(\frac{1}{N}\sum_{i=1}^{N} \frac{\xi_i}{\pi_i} Y_i + \Big(\frac{1}{\hat{N}} - \frac{1}{N}\Big) \sum_{i=1}^{N} \frac{\xi_i}{\pi_i} Y_i - \theta\Big) \\ &= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{\xi_i}{\pi_i} Y_i + \theta \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \big(\frac{\xi_i}{\pi_i} - 1\big) - \sqrt{N}\theta + o_p(1) \\ &= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} Q_i + o_p(1), \end{aligned}$$

kde $Q_i = \frac{\xi_i}{\pi_i}(Y_i - \theta)$ jsou iid náhodné veličiny. Podle centrální limitní věty $\sqrt{n}(\tilde{\theta} - \theta) \xrightarrow{d} N(\mathrm{E}\, Q_i, \mathrm{var}\, Q_i)$.

Zřejmě

$$\mathrm{E}\, Q_i = 0,$$

a tedy

$$\mathrm{var}\, Q_i = \mathrm{E}\, Q_i^2 = \mathrm{E}\, \frac{1}{\pi_i}(Y_i - \theta)^2.$$

□

Všimněme si, že zatímco v případě výběru z domácností je asymptotika založena na rostoucím počtu domácností, tedy $n \to \infty$, pro bernoulliovský výběr je rozhodující rostoucí počet jedinců, $N \to \infty$. Pro srovnatelný rozsah výběru mají tedy oba odhady stejný rozptyl (využijeme-li asymptotický rozptyl k aproximaci rozptylu odhadů):

$$\begin{aligned}\mathrm{var}\, \tilde{\theta} &= \frac{1}{N}\mathrm{E}\, M_i(Y_i - \theta)^2 = \frac{1}{n\frac{1}{n}\sum_{j=1}^{n} M_j}\mathrm{E}\, M_i(Y_i - \theta)^2 \\ &\xrightarrow{P} \frac{1}{n\mu}\mathrm{E}\, M_i(Y_i - \theta)^2 = \mathrm{var}\, \hat{\theta}.\end{aligned}$$

Poznámka. U bernoulliovského výběru (a tedy i v Tvrzení 2) předpokládáme, že počet jedinců N, ze kterých vybíráme, je pevný, zatímco v prvním případě (a tedy v Tvrzení 1) je pevný počet domácností n a celkový počet jedinců $\sum_{i=1}^{n} M_i$ je náhodná veličina. Kdybychom hledali přesnější analogii výběru z domácností, museli bychom i v případě bernoulliovského výběru považovat N za náhodné, což by samozřejmě vedlo k většímu rozptylu odhadu. Další nepřesností je, že u bernoulliovského výběru považujeme velikosti domácností M_j za nezávislé, což opět v přísné analogii výběru z domácností neplatí.

Přesto mají uvedená tvrzení důležitý důsledek pro praxi. Podle nich je totiž možné pro analýzu dat na základě výběru z domácností použít dostupný software, např. balík `survey` [3] ve statistickém softwaru `R`, kde jsou implementovány základní statistické metody pouze pro výběr bernoulliovský.

4. Ilustrace

Výsledky ilustrujeme na malé simulační studii. Předpokládejme, že domácnosti mohou mít se stejnou pravděpodoností velikost od jednoho do pěti členů. Představme si, že jsme u jejich členů měřili míru daných sociálních dovedností, které byly ohodnoceny určitým skóre. Střední hodnota skóre pro člena z domácnosti o velikosti m je

$$\theta_m = 75 + 25m,$$

Skutečná hodnota θ	$166,667$
Průměrný odhad θ	$166,620$
Asymptotický rozptyl $\hat{\theta}$	$3,395$
Empirický rozptyl $\hat{\theta}$	$3,264$
Průměrný odhad rozptylu $\hat{\theta}$	$3,395$

TABULKA 1. Průměrné výsledky simulace (1 000 opakování)

střední hodnota skóre jakéhokoliv jedince je tudíž

$$\theta = \frac{1}{3}\sum_{m=1}^{5} m(75+25m)\frac{1}{5} = 166,667.$$

Rozptyl měřeného skóre, je-li dána velikost rodiny, je 2000. Podle Tvrzení 1 je tedy $\Sigma_{\hat{\theta}} = 3395,062$.

Nejprve byla vygenerována populace 1000 rodin podle právě popsaného modelu a potom byl z každé rodiny vybrán jeden člen. Na základě získaného výběru jsme odhadli střední hodnotu ($\hat{\theta}$). Tento proces byl zopakován $1000\times$. Výsledky jsou uvedené v Tabulce 1.

5. Diskuse a závěr

V tomto příspěvku jsme se pro jednoduchost zabývali pouze situací, kdy máme k dispozici prostý náhodný výběr n domácností a z nich náhodně vybereme jednoho zástupce. Uvedli jsme odhad střední hodnoty odpovídající zvolenému výběrovému schématu a ukázali jsme, že v takovém případě má tento odhad stejný rozptyl jako odhad střední hodnoty při bernoulliovském výběru. Uvedený postup je snadno zobecnitelný na případ, kdy je výběr domácností složitější, např. stratifikovaný, nebo vybíráme více členů jedné domácnosti.

Literatura

[1] Graubard B. I., Korn E. L. (2002) *Inference for Superpopulation Parameters Using Sample Surveys*. Statistical Science **17**, 73–96.
[2] Korn E. L.,Graubard B. I. (1998) *Variance estimation for superpopulation parameters*. Statistica Sinica **8**, 1131–1151.
[3] Lumley T. (2004) *Analysis of complex survey samples*. J Stat Softw **9**, 1–19.
[4] Särndal C. E., Swensson B. and Wretman J. (1991) Model Assisted Survey Sampling. Springer-Verlag, New York.

Adresa: MFF UK, KPMS, Sokolovská 83, 186 75 Praha 8 – Karlín
E-mail: sedova@karlin.mff.cuni.cz

BERNSTEIN – VON MISES THEOREM AND ITS APPLICATION IN SURVIVAL ANALYSIS

Jana Timková

Keywords: Cox model, Bayesian asymptotics, Hadamard differential, functional delta method, survival function, median residual life.

Abstract: In this paper we deal with asymptotic properties of functionals of parameters of Cox model from frequentist and Bayesian point of view.

Abstrakt: Článek se zabývá frekvenčnými a Bayesovskými asymptotickými vlastnostemi funkcionálů parametrů Coxova modelu.

1. Introduction

When we deal with regression models in survival analysis, we estimate various parameters as is cumulative hazard functions and regression parameter. The large sample properties of the estimators are usually known. However, sometimes we need to transfer these asymptotic features from estimators to functionals of estimators. Then, the infinite-dimensional (functional) delta method hand in hand with Hadamard differentiability may serve a tool.

However, sometimes the classical asymptotics is tedious or impossible to conduct. Then, the Bernstein-von Mises theorem (BvM) as a bridge between Bayesian and frequentist asymptotics represents a way since the asymptotic properties can be always estimated from posterior sample. Basically, the theorem states that under mild conditions the posterior distribution of the model parameter centered at the maximum likelihood estimator (MLE) is asymptotically equivalent to the sampling distribution of the MLE. In turn, we can use the Bayesian asymptotics as an alternative to deriving the frequentialone.

In following we will summarize the frequentist and Bayesian asymptotic properties of parameters of Cox model and show the way of establishing the same for their functionals.

2. Cox's regression model

Let us have a multivariate counting process $\mathbf{N}(t) = (N_1(t), N_2(t), .., N_n(t))^\top$ observed in time interval $[0, \tau]$. We assume *the multiplicative intensity model*, so that the intensity takes form $I_i(t) = Y_i(t)\lambda_i(t)$, where $\lambda_i(t)$ is a deterministic bounded nonnegative continuous hazard rate function and $Y_i(t)$ is a predictable $\{0, 1\}$-valued process indicating whether the i-th individual is at risk of event whenever $Y_i(t) = 1$. The processes $Y_1, .., Y_n$ are assumed to be observed alongside with $N_1, .., N_n$. Further, for each i, let $\mathbb{Z}_i$ be a p-variate column vector of time-independent covariates associated with the i-th object. We adopt the well-known *Cox model* of Cox [3], so the hazard rate λ_i is of following form

$$\lambda_i(t) = \exp\{\beta^\top \mathbb{Z}_i\}\lambda_0(t),$$

where β is a column vector of p unknown regression coefficients and λ_0 is an unknown and unspecified baseline hazard rate common for all individuals (the hazard rate function for individual with $\mathbb{Z} = 0$).

The traditional approach to the regression parameter estimation is via the partial maximum likelihood theory. The estimator $\hat{\beta}$ of β is defined as a solution of $U(\beta, \tau) = 0$, where $U(\beta, t), t \in [0, \tau]$, is the score process equal to

$$U(\beta, t) = \sum_{i=1}^{n} \int_0^t \left(\mathbb{Z}_i - \frac{\sum_{j=1}^n Y_j(s)\mathbb{Z}_j \exp\{\beta^\top \mathbb{Z}_j\}}{\sum_{j=1}^n Y_j(s) \exp\{\beta^\top \mathbb{Z}_j\}} \right) dN_i(s).$$

The cumulative baseline hazard function $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$ is usually estimated using the Breslow estimator

$$\hat{\Lambda}_0(t) = \int_0^t \left[\sum_{i=1}^{n} Y_i(s) \exp\{\hat{\beta}^\top \mathbb{Z}_i\} \right]^{-1} d \sum_{i=1}^{n} N_i(s).$$

Notation: Let β_{tr} and λ_{tr} (as well as $\Lambda_{tr}(t) = \int_0^t \lambda_{tr}(s)ds$) represent the true values of parameters. Before stating following theorem we introduce necessary notation:

$$\begin{aligned}
q_j(\beta, s) &= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} Y_i(s)\mathbb{Z}_i^{\otimes j} \exp\{\beta^\top \mathbb{Z}_i\}, \quad j \in \{0, 1, 2\}, \\
\Sigma(\beta, t) &= \int_0^t \left[\frac{q_2(\beta, s)}{q_0(\beta, s)} - \frac{q_1(\beta, s)^{\otimes 2}}{q_0(\beta, s)^2} \right] q_0(\beta, s)\lambda_{tr}(s)ds \\
V(t) &= \int_0^t \frac{1}{q_0(\beta_{tr}, s)} \lambda_{tr}(s)ds \\
E(t) &= \int_0^t \frac{q_1(\beta_{tr}, s)}{q_0(\beta_{tr}, s)} \lambda_{tr}(s)ds,
\end{aligned}$$

where $t \in [0, \tau]$ and $\beta \in \mathbb{R}^p$. Here we use the operator $^{\otimes j}$ for $j = 0, 1, 2$, that is $\phi(s)^{\otimes 0} = 1$, $\phi(s)^{\otimes 1} = \phi(s)$ and $\phi(s)^{\otimes 2} = \phi(s)\phi(s)^\top$.

Theorem 1 (Asymptotics for β and Λ_0, [1])**.** *Under Conditions A-D of Andersen and Gill* [1] *the following is true:*

1.
$$\sqrt{n}(\hat{\beta} - \beta_{tr}) \xrightarrow{\mathscr{D}} \mathcal{N}(0, \Sigma(\beta_{tr}, \tau)^{-1})$$

2.
$$\mathscr{L}(\sqrt{n}(\hat{\Lambda}_0(\cdot) - \Lambda_{tr}(\cdot))|\sqrt{n}(\hat{\beta} - \beta_{tr}) = x) \xrightarrow{\mathscr{D}} W(V(\cdot) - xE(\cdot))$$

on the space of functions continuous to the right and with limits to the left, $D[0, \tau]$. W denotes the standard Brownian motion.

3. **Bayesian modelling**

In semiparametric Bayes method, the nonparametric part is assumed to be a realization of a stochastic process. In Cox model, among the most popular choices of a prior process for cumulative hazard function fall the Gamma and Beta process, or alternatively the Dirichlet process when modelling the distribution function. All of these processes belong to a wider family of priors conjugate to the right-censored survival data introduced by Kim and Lee in [6] and [5]. Following their notation, it is said that a prior process on the c.d.f. F_0 is *a process neutral to the right* if corresponding $\Lambda_0 = \int dF_0(s)/(1 - F_0(s_-))$ is a positive nondecreasing independent increment process (a nonstationary subordinator in the language of Lévy processes, further NII) such that $\Lambda_0(0) = 0, 0 \leq \Delta\Lambda_0(t) \leq 1$, for all t, w.p. 1, and either $\Delta\Lambda_0(t) = 1$ for some $t > 0$ or $\lim_{t\to\infty} \Lambda_0(t) = \infty$ w.p. 1.

The Lévy measure ν of an NII process is defined

$$\nu([0,t] \times B) = \mathrm{E}\left(\sum_{s\in[0,t]} I(\Delta\Lambda_0(s)) \in B \setminus \{0\}\right)$$

where $t \geq 0$, B is a Borel subset of $[0, 1]$. Let us assume that the baseline c.d.f. F_0 is, a priori, a process neutral to the right and the corresponding Λ_0 is an NII process with the Lévy measure

$$\nu(dt, dx) = \frac{1}{x} g_t(x)\zeta(t)\ dx\ dt, \quad t \geq 0, x \in [0, 1],$$

where $\int_0^1 g_t(x)dx = 1,\ \forall t$, and ζ is bounded and positive on $[0, \tau]$. And let $\pi(\beta)$ be prior distribution for β which is continuous at β_{tr} with $\pi(\beta_{tr}) > 0$.

Theorem 2 (Bernstein - von Mises theorem for β and Λ_0, [5])**.** *Under conditions (A1)-(A5), (C1) and (C2) in* [5] *the following holds:*

1. $$\lim_{n\to\infty} \int_{\mathbb{R}^p} |f_n(x) - \phi(x)|dx = 0$$
   with probability 1, where f_n is the marginal posterior density of $x = \sqrt{n}(\beta - \hat{\beta})$ and ϕ is the normal density with mean 0 and variance $\Sigma(\beta_{tr}, \tau)^{-1}$.
2. $$\mathscr{L}(\sqrt{n}(\Lambda_0(\cdot) - \hat{\Lambda}_0(\cdot)|\sqrt{n}(\beta - \hat{\beta}) = x, \sigma\{N_i, \mathbb{Z}_i, Y_i; i = 1, .., n\}) \xrightarrow{\mathscr{D}} W(V(\cdot) - xE(\cdot))$$
   on the space of functions continuous to the right and with limits to the left, $D[0, \tau]$, with probability 1, as $n \to \infty$. W denotes the standard Brownian motion.

In first proposition of Theorem 2 we actually have convergence in L_1 norm which is stronger than the usual Bernstein-von Mises statement and also the frequentists' result in Theorem 1.

4. Asymptotics for functionals of parameters

Joint posterior distribution of β and Λ_0 and Hadamard differentiability with the functional delta method (see II.8 in [2] or [4]) gives a way to establish analogical result to Theorem 2 for any smooth functional of β and Λ_0.

Let us take a sneak peek into the world of functionals and their differentiability. Firstly, let us endow the space of *cadlag* functions $D[0,\tau]$ with supremum norm instead of usual Skorohod metric and let $\mathscr{B}$ be σ-algebra generated by the supremum-norm open balls. We also need to switch to broader definition of weak converegence: a sequence X_n of random elements of $(D[0,\tau],\mathscr{B})$ *converges weakly* to X, $X_n \xrightarrow{\mathscr{D}} X$, if $\mathrm{E}\, f(X_n) = \mathrm{E}\, f(X)$ for every bounded continuous real-valued measurable function f on $D[0,\tau]$.

The next step is the definition of differentiability of elements of normed vector spaces like $D[0,\tau]$ or $D[0,\tau]\times\mathbb{R}^p$. As it turns out the Hadamard differentiability (or differentiability on compact sets) is well attunned for the weak convergence theory.

Definition 1. *Let us have two normed vector spaces B_1, B_2 , let $\eta : B_1 \to B_2$ be some function and let $\mathscr{S}$ be set of all compact subsets of B_1. Then the function η is called Hadamard (compactly) differentiable at point $x \in B_1$ with derivative $d\eta_x$ (where $d\eta_x(h)$ is linear and continuous as a function of h) if for all $S \in \mathscr{S}$*

$$\frac{\eta(x+th)-\eta(x)-d\eta_x(th)}{t} \longrightarrow 0 \quad \textit{uniformly in } h \in S.$$

Now we can introduce the functional delta method.

Theorem 3 (The delta method, [4])**.** *Let B_1 and B_2 be normed vector spaces with σ-algebras $\mathscr{B}_1$ and $\mathscr{B}_2$ nested between open-balls and open-sets σ-algebras. Suppose $\eta : B_1 \to B_2$ is Hadamard differentiable at a point $\mu \in B_1$ with derivative $d\eta_\mu$ and both η and $d\eta_\mu$ are measurable w.r.t. $\mathscr{B}_1$ and $\mathscr{B}_2$. Let X_n be a sequence in B_1 such that $Z_n = n^{1/2}(X_n-\mu) \xrightarrow{\mathscr{D}} Z$ in B_1, where the distribution of Z is concentrated on a separable subset of B_1. Then*

$$n^{1/2}(\eta(X_n)-\eta(\mu)) - d\eta_\mu(n^{1/2}(X_n-\mu)) \xrightarrow{P} 0$$

and

$$n^{1/2}(\eta(X_n)-\eta(\mu)) \xrightarrow{\mathscr{D}} d\eta_\mu(Z).$$

In application a functional might often be a composition of several functionals. Then *the chain rule* comes in handy, since it states that, for some normed vector spaces B_1, B_2 and B_3, if $\eta : B_1 \to B_2$ and $\varsigma : B_2 \to B_3$ are Hadamard differentiable at $x \in B_1$ and $\eta(x) \in B_2$ respectively, then $\eta\circ\varsigma : B_1 \to B_3$ is Hadamard differentiable at x with derivative $d\varsigma_{\eta(x)} \circ d\eta_x$.

Combining the results of Theorem 1 and 2 we get the large sample results for an arbitrary functional of model parameters as long as it is Hadamard differentiable.

Corollary 1 (Frequential asymptotics for smooth functionals of β and Λ_0)**.** *Assume that the conditions of Theorem 1 are fulfilled and that B is a normed vector space with a σ-algebra $\mathscr{B}$ nested between open-balls and open-sets σ-algebras. If a functional η of the parameters β and Λ_0, $\eta : \mathbb{R} \times D[0,\tau] \to B$, is Hadamard differentiable at the point $(\beta_{tr}, \Lambda_{tr})$ with derivative $d\eta_{(\beta_{tr},\Lambda_{tr})}$ then the following is true:*

$$\sqrt{n}(\eta(\hat{\beta}, \hat{\Lambda}_0) - \eta(\beta_{tr}, \Lambda_{tr})) \xrightarrow{\mathscr{D}} d\eta_{(\beta_{tr},\Lambda_{tr})}(X, W(V + E^\top \Sigma^{-1}(\beta_{tr}, \tau)E)).$$

Corollary 2 (Bernstein-von Mises for smooth functionals of β and Λ_0)**.** *Let the assumptions of Theorem 2 be fulfilled. Assume that B is a normed vector space with a σ-algebra $\mathscr{B}$ nested between open-balls and open-sets σ-algebras. If a functional η of the parameters β and Λ_0, $\eta : \mathbb{R} \times D[0,\tau] \to B$, is Hadamard differentiable at the point $(\beta_{tr}, \Lambda_{tr})$ with derivative $d\eta_{(\beta_{tr},\Lambda_{tr})}$ then, with probability 1,*

$$\begin{aligned}\mathscr{L}(\sqrt{n}(\eta(\beta, \Lambda_0) - \eta(\hat{\beta}, \hat{\Lambda}_0)|\sigma\{N_i, \mathbb{Z}_i, Y_i; i = 1, .., n\}) \\ \xrightarrow{\mathscr{D}} \quad d\eta_{(\beta_{tr},\Lambda_{tr})}(X, W(V + E^\top \Sigma^{-1}(\beta_{tr}, \tau)E)).\end{aligned}$$

In next we will deal with most common functionals present in Cox regression model.

Baseline survival function. The baseline survival function $S(t) = 1 - F(t)$ can be expressed as

$$S_0(t) = \prod_{[0,t]} [1 - d\Lambda_0]$$

where with $\prod_{[a,b]}$ we denote the product integral over the interval $[a, b]$. It can be seen that the mapping $\eta : D[0,\tau] \to D[0,\tau]$ such that $\eta : \Lambda_0 \mapsto S_0(\cdot)$ is Hadamard differentiable (see Prop. II.8.7 in [2]). The derivative at the point $\Lambda_0 \in D[0,\tau]$ is equal to

$$\begin{aligned}(d\eta_{\Lambda_0}(H))(t) &= -\int_{s\in[0,t]} \prod_{[0,s)} [1 - d\Lambda_0]\, H(ds) \prod_{(s,t]} [1 - d\Lambda_0] \\ &= - \quad S_0(t_-)H(t), \quad t \in (0,\tau].\end{aligned}$$

The MLE estimator of S_0 is $\hat{S}_0(t) = \prod_{[0,t]}[1 - \hat{\Lambda}_0]$ and in case of no covariates coincides with Kaplan-Meier estimator. Let us denote the true survival function by S_{tr} . Using this result, Corollary 1 and supposing that the distribution is absolutely continuous, we have the convergence in every $t \in [0,\tau]$

$$\sqrt{n}(\hat{S}_0(t) - S_{tr}(t)) \xrightarrow{\mathscr{D}} - S_{tr}(t)W(V(t) + E(t)^\top \Sigma^{-1}(\beta_{tr}, \tau)E(t)).$$

The asymptotic variance $S_{tr}(t)^2[V(t) + E(t)^\top \Sigma^{-1}(\beta_{tr}, \tau)E(t)]$ can be estimated by plugging-in the estimators $\hat{\beta}$, $d\hat{\Lambda}_0$ and $\hat{S}_0$ instead of β_{tr}, $\lambda_{tr}ds$ and S_{tr} in $V(t)$, Σ and $E(t)$. This result may be used to calculate the pointwise confidence limits for $S_0(t)$ or alternatively we can specify the limiting distribution

as the supremum of transformed Brownian motion since using the continuous mapping theorem gives

$$\sup_{t\in[0,\tau]}\left\{\frac{n}{\hat{V}(\tau)+\hat{E}(\tau)\hat{\Sigma}^{-1}(\hat{\beta},\tau)\hat{E}(\tau)}\right\}^{1/2}\frac{|\hat{S}_0(t)-S_{tr}(t)|}{\hat{S}_0(t)}\xrightarrow{\mathscr{D}}\sup_{x\in[0,1]}|W(x)|$$

Using Corollary 2 we get the Bayesian asymptotic properties. The posterior distribution of the process S_0 centered around ML estimator converges weakly w. p. 1 to the same limiting process

$$\mathscr{L}(\sqrt{n}(S_0(\cdot)-\hat{S}_0(\cdot)|\sigma\{N_i,\mathbb{Z}_i,Y_i;i=1,..,n\}) \xrightarrow{\mathscr{D}} -S_{tr}(\cdot)W(V+E^\top\Sigma^{-1}(\beta_{tr},\tau)E).$$

This knowledge can be used when we want to avoid the deriving of the asymptotic variance or using its plug-in estimator and we can create pointwise credibility bands from a posterior sample instead. Bayesian version of the distribution of a supremum of asymptotic distribution can be obtain from the sample of supremum values for each of posterior realisations of $S_0^{(k)}=\eta(\beta^{(k)},\Lambda_0^{(k)})$, $k=1,\dots,K$. Then, for example, we can find $\alpha>0$ such that

$$\mathscr{P}(\sup\sqrt{n}|S_0(\cdot)-\hat{S}_0(\cdot)|>\alpha|\sigma\{N_i,\mathbb{Z}_i,Y_i;i=1,..,n\})=0.95$$

by taking the 95% sample quantile of the supremum values of all posterior realisations.

Survival function for $\mathbb{Z}=\mathbb{Z}^\star$. The survival function for an individual with certain value of covariate is defined as

$$S(t;\mathbb{Z}^\star)=\prod_{[0,t]}\left[1-\exp\{\beta^\top\mathbb{Z}^\star\}d\Lambda_0\right]$$

The mapping $\eta:\mathbb{R}\times D[0,\tau]\to D[0,\tau]$ which assigns a point $(\beta,\Lambda_0)\in\mathbb{R}\times D[0,\tau]$ the value $S(\cdot;\mathbb{Z}^\star)$ is again Hadamard differentiable. Here we, however, need to use the chain rule feature for the composition of two mappings $\eta=\eta_2\circ\eta_1$ where $\eta_1(\beta,\Lambda_0)=\exp\{\beta^\top\mathbb{Z}^\star\}\Lambda_0$ and $\eta_2(x)=\prod_{[0,\cdot]}[1-dx]$.

The derivative at the point $(\beta,\Lambda_0)\in\mathbb{R}\times D[0,\tau]$ is equal

$$(d\eta_{(\beta,\Lambda_0)}(h,H))(t)=-\int_{s\in[0,t]}\prod_{[0,s)}\left[1-e^{\beta^\top\mathbb{Z}^\star}d\Lambda_0\right]\Big(e^{\beta^\top\mathbb{Z}^\star}h^\top\mathbb{Z}^\star\Lambda_0(ds) + e^{\beta^\top\mathbb{Z}^\star}H(ds)\Big)\prod_{(s,t]}\left[1-e^{\beta^\top\mathbb{Z}^\star}d\Lambda_0\right],\quad t\in[0,\tau].$$

So, the limiting process in both frequential and Bayesian asymptotics is

$$-S_{tr}(t;\mathbb{Z}^\star)e^{\beta_{tr}^\top\mathbb{Z}^\star}[X^\top\mathbb{Z}^\star\Lambda_{tr}(t)+W(V(t)+E(t)^\top\Sigma^{-1}(\beta_{tr},\tau)E(t))],\ t\in[0,\tau].$$

where X is normally distributed zero-mean variable with variance $\Sigma^{-1}(\beta_{tr},\tau)$. The asymptotic variance equals

$$\{e^{\beta_{tr}^\top\mathbb{Z}^\star}S_{tr}(t;\mathbb{Z}^\star)\}^2\left[(E-\mathbb{Z}^\star\Lambda_{tr})^\top\Sigma^{-1}(\beta_{tr},\tau)(E-\mathbb{Z}^\star\Lambda_{tr})+V\right].$$

and its estimator can be found by plugging-in the estimated parameters $\hat{\beta}$ and $d\hat{\Lambda}_0$ instead of β_{tr} and $\lambda_{tr}ds$. Similarly as when dealing with baseline survival function, the pointwise bands or supremum can be obtained via plugged-in estimator variance or by using the posterior sample of $S(\beta, \mathbb{Z}^\star)$.

Median residual life. The median residual life for individual with the covariate $\mathbb{Z} = \mathbb{Z}^\star$ is $\gamma_{t_0}(\mathbb{Z}^\star)$ such that

$$\frac{S(\gamma_{t_0}(\mathbb{Z}^\star); \mathbb{Z}^\star)}{S(t_0; \mathbb{Z}^\star)} = 0.5, \quad \text{for} \quad t_0 \in (0, \tau).$$

It is not difficult to see that for Cox model the median residual life equals

$$\gamma_{t_0}(\mathbb{Z}^\star) = \Lambda_0^{-1}(\Lambda_0(t_0) + \log 2 \exp\{-\beta^\top \mathbb{Z}^\star\}).$$

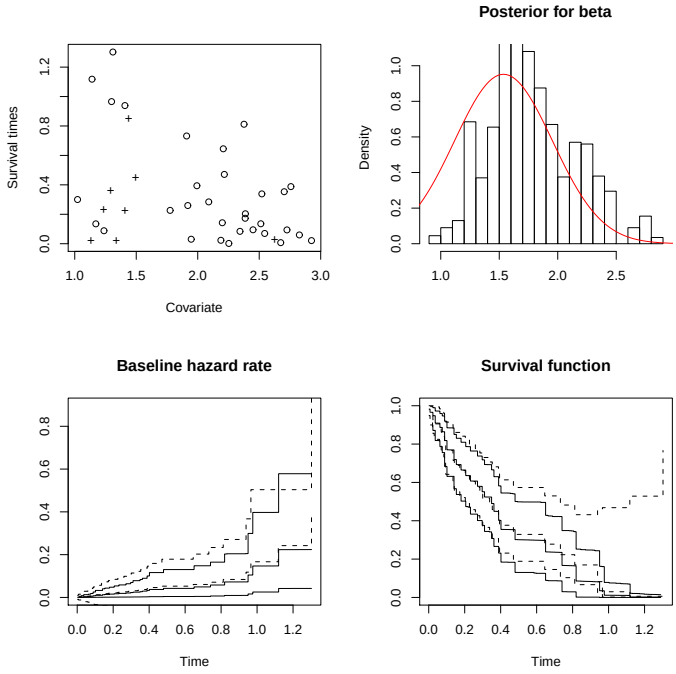


FIGURE 1. Upper: Left: Data, "∘"for failure, "+"for censored. Right: Histogram of posterior sample of β with theoretical limiting distribution in red. Lower: Left: Estimated cumulative BHR with 95% pointwise CI, solid - Bayesian, dashed - frequentist. Right: Estimated survival function with $\mathbb{Z}^\star = \overline{z}$ with 95% pointwise CI, solid - Bayesian, dashed - frequentist.

To be able to obtain asymptotic distribution of η_{t_0} we have to investigate the differentiability of the function $\eta : (\Lambda_0, \beta) \mapsto \gamma_{t_0}$ which could be again expressed as a composition of functions $\eta_1(\Lambda_0, \beta) = \Lambda_0(t_0) + \log 2 \exp\{-\beta^\top \mathbb{Z}^\star\}$ and $\eta_2(\Lambda_0, z) = \Lambda_0^{-1}(z)$. Both η_1 and η_2 are Hadamard differentiable. For derivative of η_2 see Prop. II.8.4 in [2] and application can be seen in e.g. [3].

5. Illustration

We illustrate the model on $n = 40$ simulated survival times from a hazard rate of form $\lambda(t; z) = 0.1t\ e^{1.5z}$ where z was randomly generated from $N(2, 1)$. For the prior of cumulative hazard rate we chose Beta process prior with parameters $\Lambda(t) = 0.05t$ and $c(t) = 10e^{-0.05t}$. The Beta process on the interval $[0, \tau]$ with mean $H \in D[0, \tau]$ and scale parameter $c(t) > 0$ is defined as a nonstationary subordinator with Lévy measure

$$\nu(dt, dx) = c(t)x^{-1}(1 - x)^{c(t)-1}dx\ dH(t).$$

It can be shown that this process satisfies the conditions of Theorem 2. For simulation of Beta process see [7]. We ran 5000 repetitions of MCMC and used last 2000 for analysis of posterior. Posterior summaries on regression parameter: β is $mean(\beta) = 1.78$ and $sd(\beta) = 0.37$. The frequentist' estimator is 1.53 with $sd = 0.41$. The results can be seen in Figure 1. We may see that Bayesian and frequentist estimators of limiting distributions are quite similar.

References

[1] Andersen P.K., Gill R.D. (1982) *Cox's regression model for counting processes: A large sample study.* Ann. Statist. 10, 1100 – 1120.

[2] Andersen P.K., Borgan A., Gill R.D., Kieding N. (1993) *Statistical models based on counting processes.* Springer, New York.

[3] De Blasi P., Hjort N.L. (2007) *Bayesian survival analysis in proportional hazard models with logistic relative risk.* Scand. J. Statist. 34, 229 – 257.

[4] Gill R.D., Wellner J.A., Prestgaard J. (1989) *Non- and semi- parametric maximum likelihood estimators and the Von Mises method (Part 1).* Scand. J. Statist. **16**, 2, 97 – 128.

[5] Kim Y. (2006) *The Bernstein-von Mises theorem for the proportional hazard model.* Ann. Statist. **34**, 4, 1678 – 1700.

[6] Kim Y., Lee J. (2004) *A Bernstein-von Mises theorem in the nonparametric right-censoring model.* Ann. Statist. **32**, 4, 1492 – 1512.

[7] Lee J., Kim Y. (2004) *A new algorithm to generate beta processes.* Comput. Statist. Data Anal. **47**, 441 – 453.

Acknowledgement: This work was supported by grant GA CR 201/05/H007 and by GA AV IAA101120604.

Address: MFF UK, KPMS, Sokolovská 83, 186 75 Praha 8 – Karlín

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

SHLUKOVÁNÍ V SOUBORECH S ODLEHLÝMI OBJEKTY POMOCÍ METOD *k*-PRŮMĚRŮ

Marta Žambochová

Klíčová slova: Shlukování, velké soubory dat, varianty algoritmu k-průměrů, odlehlé objekty.

Abstrakt: Velká citlivost shlukování na odlehlá pozorování je skutečnost, která může záporně ovlivnit kvalitu výsledného rozdělení do shluků. Ve většině případů jsme odkázáni na vhodné předzpracování dat a případné vyloučení odlehlých objektů z dalšího zpracování.V odborné literatuře se však objevují shlukovací metody přímo zaměřené na data obsahující odlehlé objekty. Jedním z takovýchto postupů je například dvoufázový algoritmus k-průměrů. V příspěvku je navržena varianta metody k-průměrů pracující s mrkd-stromy, která je postavena na jiném principu. Identifikace odlehlých objektů probíhá v rámci fáze předzpracování, kterou je nutno provádět i v případě, že nás odlehlé objekty nezajímají. Je to fáze organizující data do stromové struktury, která činí následující fázi shlukování velmi efektivní. Dále článek předkládá třetí možnost detekce odlehlých objektů pomocí modifikace algoritmu k-průměrů$^{++}$. Příspěvek pojednává o srovnání uvedených tří metod.

Abstract: Great sensitivity of clustering to outliers may negatively affect the quality of the resulting division into clusters. In most cases we must rely on an appropriate preprocessing and a possible exclusion of outliers. However, there are clustering methods aimed at the data containing outliers, in professional statistics literature. One such example is the two-step k-means algorithm. The paper proposes an alternative to the k-means method working with mrkd-trees, which is based on another principle. The identification of outliers is in the phase of preprocessing, which must be done even if we are not interested in outliers. It's a phase, which organizes the data into a tree structure, which makes the next phase of clustering very effective. The article also presents a third option involving the detection of outliers by modifying the algorithm k-means^{++}. The paper outlines a comparison between the three methods.

1. Úvod

Citlivost shlukování na odlehlá pozorování je fakt, který může záporně ovlivnit kvalitu výsledného rozdělení do shluků. V mnoha případech, zvlášť pokud data zpracováváme pomocí standardních statistických programových systémů, jsme odkázáni na vhodné předzpracování dat a případné vyloučení odlehlých objektů z dalšího zpracování. Touto problematikou se zabývá například článek [4] či [12]. Jinou možností je, jak navrhuje autor například v [5], spuštění několika málo iterací shlukovacího algoritmu, po kterých se může vytvořit shluk, respektive shluky, obsahující jen zanedbatelné množství objektů. Objekty v těchto shlucích můžeme považovat za odlehlé. Uvedený

způsob není dle mého názoru ideální z důvodu nejasnosti potřebného počtu iterací, po kterých se oddělí malé shluky. Tento počet je různý pro různá počáteční rozdělení do shluků. Pro obzvlášť velké datové soubory není výše uvedený způsob využitelný vůbec z důvodu velké časové náročnosti zpracování jednotlivých iterací.

V odborné literatuře se objevují shlukovací metody přímo zaměřené na data obsahující odlehlé objekty. Článek se zabývá vybranými algoritmy pracujícími na principu metody k-průměrů. Metoda k-průměrů nepracuje s maticí vzdáleností pro všechny dvojice objektů, a proto je velmi vhodná pro zpracování souborů s velkým počtem objektů. Každý ze shluků je reprezentován svým středem, tj. d-rozměrným vektorem skládajícím se z průměrných hodnot jednotlivých proměnných (tzv. centroidem), metoda je tedy použitelná pouze pro zpracování kvantitativních dat. Jedná se o velmi oblíbenou a hojně používanou iterativní shlukovací metodu, jejíž základní myšlenkou je hledání rozkladu objektů do předem daného počtu shluků, pro který je součet vzdáleností jednotlivých objektů od centra jejich shluku minimální, tj. hledání minima účelové funkce

$$Q = \sum_{x} ||x - c(x)||^2,$$

kde x je libovolný objekt, $c(x)$ je centroid nejbližší objektu x.

Vybranými postupy pracujícími na principu metody k-průměrů a umožňujícími odhalování odlehlých objektů jsou například modifikace využívající algoritmus k-průměrů$^{++}$, dvoufázový algoritmus k-průměrů či algoritmus MFA. Tyto algoritmy umí detekovat skupinky objektů s malým počtem objektů (tj. menším než daná konstanta), které jsou od zbylých objektů velmi vzdálené.

2. Dvoufázový algoritmus k-průměrů

Jednou z variant algoritmu k-průměrů umožňujících odhalení odlehlých objektů je dvoufázový algoritmus k-průměrů. Metodu popsali autoři v [6]. Postup v první fázi využívá modifikovaný algoritmus k-průměrů ovlivněný algoritmem ISODATA. Je zde využita heuristika: „pokud je vkládaný objekt velmi vzdálen od všech dosavadních center shluků, je zařazen do nově vzniklého shluku“. Na rozdíl od klasického algoritmu k-průměrů nevytváří předem daný počet shluků, ale konečný počet shluků se pohybuje v předem daném rozmezí. Výsledkem první fáze je rozdělení do k' shluků, kde $k \leq k' \leq n$, kde k je požadovaný počet shluků a n je počet objektů, přičemž objekty v jednom shluku jsou buď všechny odlehlé, nebo není odlehlý ani jeden.

V druhé fázi algoritmus za pomoci minimální kostry odhalí odlehlé objekty a vytvoří cílové rozdělení do požadovaných k shluků. Všech k' center vzniklých v první fázi je považováno za objekty nového shlukování. Na nalezení odlehlých objektů jsou vhodné shlukovací metody, které na základě velké

vzdálenosti oddělí některé objekty od ostatních. Mezi takovéto metody například patří hierarchické metody shlukování, nebo metoda založená na principu minimální kostry. Z důvodu použitelnosti metody pro velmi velké soubory dat autoři v článku [6] zavrhli shlukování pomocí hierarchických metod, jejichž časová náročnost roste s třetí mocninou počtu objektů, a zvolili efektivnější metodu založenou na minimální kostře. Centra shluků vzniklá v první fázi se stanou vrcholy úplného grafu, jehož každá hrana je ohodnocena vzdáleností daných dvou center. Pomocí libovolného, k tomu určeného (viz [3], [11]), algoritmu nalezneme minimální kostru vytvořeného grafu. Vyjmutím hrany s maximálním ohodnocením obdržíme dvě komponenty (souvislé části grafu), které reprezentují dva shluky. Dalším postupným vyjímáním hran s maximálním ohodnocením dostáváme potřebné množství shluků. Objekty málo početných shluků označíme za odlehlé.

3. Modifikovaný algoritmus k-průměrů$^{++}$

Algoritmus k-průměrů$^{++}$ popsali poprvé jeho autoři ve článku [2]. Tato alternativa vytváří speciální inicializační rozdělení do shluků pomocí množiny center na jejímž základě se provede rozdělení do shluků. Prvním centrem je náhodně vybraný objekt ze všech datových objektů. Další centra jsou jedno po druhém vybíráno ze zbývajících datových objektů. Vždy je vybrán objekt s nejvyšší pravděpodobností, která je vypočítána podle vztahu

$$P = \sum_x \frac{D(y)^2}{D(x)^2},$$

kde y je zkoumaný objekt, x je libovolný objekt, $D(x)$ (resp. $D(y)$) je nejkratší vzdálenost objektu x (resp. y) od nejbližšího centra ze všech doposud vybraných.Tímto postupem vybereme zbývajících $k-1$ center. Další postup algoritmu k-průměrů$^{++}$ je shodný s klasickým algoritmem k-průměrů.

Speciálního postupu při výběru objektů do množiny inicializačních center jsem využila k detekci odlehlých objektů. Nové centrum je vybíráno tak, aby od všech doposud vybraných center bylo co nejvíce vzdálené. Z principu výběru jednotlivých center je zřejmé, že odlehlé objekty jsou vždy prvky množiny „dostatečného množství“ inicializačních center. Toto „dostatečné množství“ je tím menší, čím více je objekt odlehlý. To znamená, že velmi odlehlé objekty, které hodně narušují kvalitu výsledného shlukování se odhalí relativně rychle. K vlastní detekci odlehlých objektů stačí po vytvoření množiny inicializačních center provést dvě či tři iterace algoritmu k-průměrů. Již v průběhu těchto iterací se odlehlé objekty znatelně oddělí.

4. Algoritmus MFA

Další variantou metody k-průměrů umožňující detekci odlehlých objektů je algoritmus MFA (Modifikovaný Filtrovací Algoritmus), viz [12] či [13]. Výhodou tohoto způsobu je fakt, že identifikace odlehlých objektů probíhá v rámci

fáze předzpracování, kterou je nutno provádět i v případě, že nás odlehlé objekty nezajímají. Je to fáze organizující data do stromové struktury, která činí následující fázi shlukování velmi efektivní.

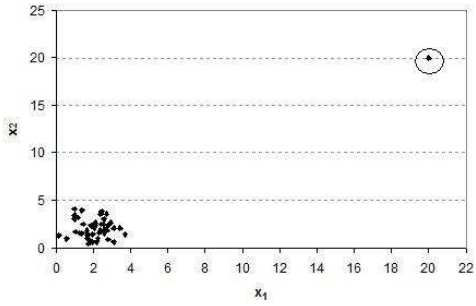
Filtrovací algoritmus, z něhož jsem při návrhu nového algoritmu vycházela, je jednou z implementací Lloydova shlukovacího algoritmu využívající speciální stromovou strukturu, tzv. mrkd-stromy. Tento algoritmus je podrobněji popsán v [8], principy, na kterých je algoritmus postaven, v [7] a [9]. Algoritmus je velkým zefektivněním klasického přístupu algoritmu k-průměrů.

Mrkd-strom je binární forma datové stromové struktury, která reprezentuje rekurzivní dělení konečné množiny bodů z d-dimenzionálního prostoru na k částí (d-dimenzionálních hyperkvádrů), pomocí $d-1$ dimenzionálních ortogonálních nadrovin. Mrkd-strom je zkonstruován pouze jedenkrát pro daný soubor objektů a celá struktura nemusí být přepočítávána v každém iteračním kroku algoritmu k-průměrů. Autoři filtrovacího algoritmu provádí rozdělení ortogonálně k nejdelší straně hyperkvádru na úrovni mediánu ze všech bodů hyperkvádru. Výsledkem algoritmu s touto myšlenkou dělení jsou velmi vyvážené stromy. Hloubka stromu se v jednotlivých větvích liší maximálně o jednu úroveň. Toto je způsobeno faktem, že dělení na úrovni mediánu zaručuje, že vrcholy v jedné úrovni stromu mají počet objektů odlišný maximálně o jeden objekt.

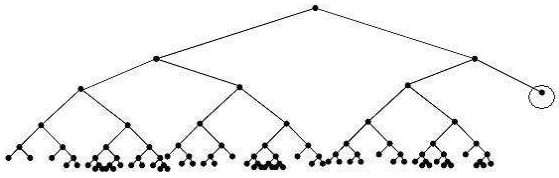
Hlavní myšlenka modifikace základního algoritmu pro tvorbu mrkd-stromu je změna způsobu dělení prostoru objektů. Pokud provedeme dělení hyperkvádru na úrovni průměru místo na úrovni mediánu obdržíme stromy, které nejsou tak dokonale vyvážené. To znamená, že se délka cest od kořene k jednotlivým listům znatelně liší. Tato nevyváženost je zapříčiněna známou vlastností aritmetického průměru, který je silně ovlivňován odlehlými hodnotami. Proto v části oddělené průměrem, která obsahuje odlehlé hodnoty, je umístěn zpravidla mnohem menší počet objektů než v části druhé. Tato zdánlivá nevýhoda může být však docela podstatnou výhodou. Relativně dobře se daří odhalit odlehlé objekty, které mohou znehodnotit celkový výsledek konečného shlukování. Čím jsou objekty odlehlejší, tím dříve je algoritmus detekuje. Cesta stromu končící listem, jež obsahuje odlehlý objekt, je tím kratší, čím je objekt odlehlejší. Po oddělení odlehlého objektu se stávají data stejnorodější a hodnota aritmetického průměru se přibližuje hodnotě mediánu, dělení hyperkvádru je symetričtější. Vznikající podstrom je již vyváženější. Mrkd-strom vzniklý variantou dělení na úrovni aritmetického průměru se znatelně člení na několik vyvážených větších podstromů a případně několik krátkých osamocených větví.

Příkladem je mrkd-strom na obrázku 2, který je vytvořen nad dvourozměrnými daty s jedním odlehlým objektem, jejichž struktura je zřejmá z obrázku 1. Minimální hloubka stromu je dvě. Tuto délku má cesta od kořene stromu končící listem reprezentujícím odlehlý objekt. Maximální hloubka v takto vytvořeném stromu je sedm. Na obrázku 3 je zobrazen mrkd-strom vytvořený nad stejnými daty pomocí původního algoritmu. Takto vytvořený strom má

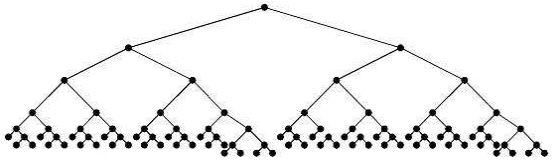
minimální hloubku pět a maximální hloubku šest. Odlehlý objekt není ve struktuře nijak viditelně odlišen.



Obrázek 1. Data s odlehlým objektem.



Obrázek 2. Mrkd-strom s dělením na pozici aritmetického průměru nad daty s odlehlým objektem.



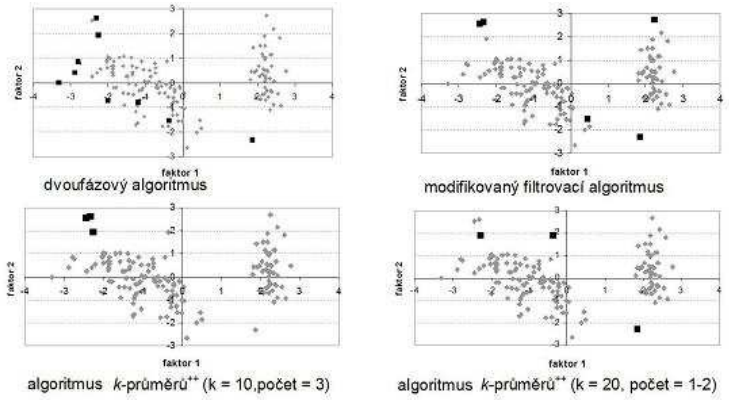
Obrázek 3. Mrkd-strom s dělením na pozici mediánu nad daty s odlehlým objektem.

5. **Provedené experimenty**

Všechny algoritmy byly naprogramovány v prostředí MATLAB. Experimenty byly prováděny na dvou souborech obsahujících reálná data a jednom souboru se speciálně vygenerovanými daty. Oba reálné soubory jsou k dispozici na internetové stránce [14].

Soubor IRIS byl vybrán z důvodu malého počtu objektů a tím i možnosti podrobného porovnání výsledků jednotlivých algoritmů. Viz obrázek 4. Soubor však neobsahuje výrazně odlehlé objekty. Výrazně odlehlý objekt obsahuje VOWEL, druhý z použitých souborů. Soubor GENER, s miliónem speciálně vygenerovaných dvourozměrných dat, byl použit k porovnání výpočetní náročnosti jednotlivých algoritmů.

Odlehlý objekt souboru VOWEL odhalily všechny sledované algoritmy dobře. Na obrázku 4. jsou názorně zobrazeny výsledné množiny detekovaných odlehlých objektů pomocí vybraných algoritmů. Obrázek 5. znázorňuje průběh času zpracování pomocí jednotlivých algoritmů v závislosti na počtu shlukovaných objektů.

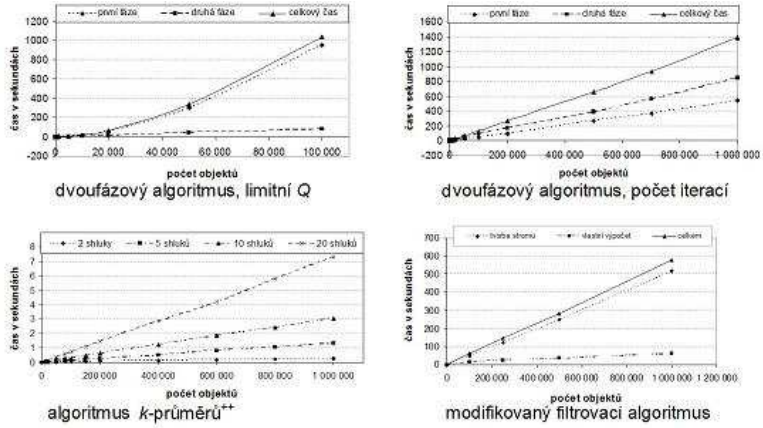


Obrázek 4. Odlehlé objekty souboru IRIS detekované pomocí různých algoritmů.

6. **Shrnutí**

Experimenty ukázaly, že výsledky dvoufázového algoritmu k-průměrů a algoritmu k-průměrů$^{++}$ jsou závislé na nastavení vstupních parametrů programu. Výsledky algoritmu MFA nejsou ovlivněny žádným uživatelským nastavením.

Pokusy zaměřené na zkoumání chování algoritmu k-průměrů$^{++}$ při detekci odlehlých objektů v souboru IRIS ukázaly, že pro hodnoty parametru k nižší než 10 algoritmus neodhalí žádné odlehlé objekty. Pro hodnotu $k = 10$ se vytvoří shluk obsahující tři objekty, který má výrazně menší počet objektů než ostatní shluky. Objekty v tomto shluku lze již označit za odlehlé. Pro



Obrázek 5. Výpočetní náročnost jednotlivých algoritmů.

hodnotu parametru $k = 15$ se vytvořil shluk obsahující samostatný objekt a dále se vytvořily shluky po dvou objektech. Při nastavení hodnoty parametru $k = 20$ se vytvoří tři shluky obsahující jediný objekt a dále tři shluky obsahující dva objekty. Pro volbu vyšší hodnoty parametru k již vzniká velké množství malých shluků, a tím i nepřiměřeně velké množství objektů, jež jsou detekovány jako odlehlé.

Dále jsem sledovala množinu detekovaných souborů vzniklou při spuštění programu dvoufázového algoritmu k-průměrů s různě volenými parametry $k_{\max}$ pro určení maximálního počtu center v první fázi a k pro požadovaný cílový počet shluků. Výsledné množiny se lišily. Se vzrůstajícími hodnotami parametru k vzrůstal počet shluků s jedním objektem, a tím i počet detekovaných odlehlých objektů. Například pro hodnotu $k = 30$ algoritmus detekoval devatenáct odlehlých objektů detekovaných pomocí shluků s jediným objektem a osm odlehlých objektů detekovaných pomocí shluků s dvěma objekty. Naopak pro zvyšující se hodnotu parametru $k_{\max}$ se počet detekovaných odlehlých objektů mírně snižuje, od určité hranice se ustálí. Některé objekty se objevily mezi detekovanými pro různá nastavení vstupních parametrů. Devět s nejčetnějším výskytem je znázorněno na obrázku 4.

Většinu algoritmů jednotlivých variant se mi podařilo naprogramovat tak, že následné experimenty naznačily lineární růst času v závislosti na počtu objektů ve zpracovávaném datovém souboru. Toto je zřejmé z obrázku 5. Průměrné hodnoty jsem vždy vypočítala z deseti naměřených hodnot. V případech, kdy se běh zpracování skládal z několika samostatných částí, jsem zaznamenala časy jednotlivých částí samostatně. I přes relativizaci výsledků (ovlivněno vlastní implementací) je zřejmé, že velmi výjimečné postavení

z hlediska časové náročnosti má algoritmus MFA postavený na mrkd-stromech.

Existuje ještě řada dalších variant, které by bylo dobré prozkoumat. Další výzkum bude zaměřen na podrobnější experimentální ověření jednotlivých algoritmů na speciálně vygenerovaných velkých souborech dat s cílem dokonalejšího porovnání jednotlivých variant. Dále by bylo dobré zaměřit se především na další vylepšení stávajících variant, zvláště z hlediska výpočetní náročnosti zpracování. Nadějně se jeví kombinace minimalizace počtu průchodů datovým souborem a využití různých forem stromových struktur. Pokud se týká vytváření nových modifikací, jedná se především o rozšíření použitelnosti metody k-průměrů v případě přípustnosti překrývání se výsledných shluků (tzv. fuzzy shlukování) či použitelnosti metody pro shlukování i jiných než kvantitativních dat.

Literatura

[1] Anděl, J., J. Zichová (2002): A method for estimating parameter in nonnegative $MA(1)$ models. *Communications in Satistics - Theory and Methods, 31, 2101 - 2111.*

[2] Arthur D., Vassilvitskii S. (2007) *k-means++ The Advantages of Careful Seeding.* Symposium on Discrete Algorithms (SODA), New Orleans, Louisiana, 1027 – 1035.

[3] Demel J. (2002) *Grafy a jejich aplikace.* Academia, Praha, 257 s.

[4] Duan L., Xu L., Liu Y., Lee J. (2009) *Cluster-based outlier detection.* Annals of Operations Research, **168** (1), 151 – 168.

[5] Goswami A., Ruoming J., Agrawal G. (2004) *Fast and exact out-of-core k-means clustering.* Data Mining, ICDM apos;04. Fourth IEEE International Conference on Volume, Issue, 83 – 90.

[6] Jiang M.F., Tseng S.S., Su C.M. (2001) *Two-phase clustering process for outliers detection.* SPattern Recognition Letters, **22**, 691 – 700.

[7] Kanungo T., Mount D.M., Netanzahu N.S., Piatko CH.D. Silverman R., Wu A.Y. (2000) *The analysis of a simple k-means clustering algorithm.* Proceedings of the Sixteenth Annual Symposium on Computational Geometry, Hong Kong 100 – 109.

[8] Kanungo T., Mount D.M., Netanzahu N.S., Piatko CH.D. Silverman R., Wu A.Y. (2002) *An Efficient k-means clustering algorithm: analysis and implementation.* Proc ACM SIGKDD Int'l Conf. IEEE Transactions on Pattern Analysis and Machina Inteligence, **24** (7).

[9] Moore A. (1999) *Very fast EM-based mixture model clustering using multiresolution kd-trees.* Advances in Neural Information Processing Systems, 543 – 549.

[10] Zichová, J. (1996): On a method of estimating parameters in non-negative ARMA models. *Kybernetika 32, 409 - 424.*

[11] Žambochová, M. (2008) *Teorie grafů v příkladech.* Skripta FSE UJEP, Ústí nad Labem, 102 s.

[12] Žambochová, M. (2009) *Odlehlé objekty a shlukovací algoritmy.* Mezinárodní statisticko-ekonomické dny na VŠE [CD-ROM]. Praha, 1 – 6

[13] Žambochová, M. (2010) *Shluková analýza rozsáhlých souborů dat: nové postupy založené na metodě k-průměrů.* Disertační práce (před obhajobou), Praha.

[14] `http://archive.ics.uci.edu/ml/datasets/`.

Adresa: FSE UJEP, KMS, Moskevská 54, CZ- 400 96, Ústí nad Labem

E-mail: `marta.zambochova@ujep.cz`

ON NONPARAMETRIC ESTIMATORS OF LOCATION OF MAXIMUM

Zdeněk Hlávka

Keywords: Kernel regression, location of maximum, optimal design.

Abstract: An estimator of the maximum of a regression function and its location is often of greater interest than an estimator of the regression curve itself. We review properties of nonparametric estimators of the location of maximum and investigate the influence of the density of design points on the asymptotic distribution of the estimator. Classical calculus of variations is used to find the optimal distribution of the design points for the nonparametric kernel estimator of the location of maximum.

Abstrakt: Odhad maxima funkce a jeho polohy bývá často zajímavější a důležitější, než odhad celé neznámé regresní funkce. Příspěvek pojednává o neparametrických odhadech polohy maxima a některých problémech, se kterými se můžeme setkat při jejich použití. Budeme se zabývat zejména vlivem volby hodnot nezávisle proměnné na asymptotický rozptyl neparametrického jádrového odhadu polohy maxima. Pomocí variačního počtu odvodíme optimální návrh experimentu pro neparametrický jádrový odhad polohy maxima.

RIDGE LEAST WEIGHTED SQUARES

Tomáš Jurczyk

Keywords: Multicollinearity, robust ridge regression, least weighted squares.

Abstract: Multicollinarity and outlier presence are classical problems of the data in linear regression framework. We are going to present a proposal of a new method which can be potential candidate for robust ridge regression as well as robust detector of multicollinearity. This proposal arises as a logical combination of principles used by ridge regression and least weighted squares estimate. We will also show the properties of new method.

Abstrakt: Jedním z problémů dat v regresní analýze může být přítomnost multikolinearity nebo například výskyt odlehlých pozorování. Tento příspěvek představuje návrh nové metody pro odhad parametrů lineárního regresního modelu, která může být kandidátem na robustní verzi hřebenové regrese, stejně jako na robustní detektor multikolinearity. Tento návrh je logickou kombinací postupů metod známých pod názvem hřebenová regrese a nejmenší vážené čtverce. V příspěvku ukážeme také základní vlastnosti nového odhadu.

MAXIMIZATION OF THE INFORMATION DIVERGENCE FROM MULTINOMIAL DISTRIBUTIONS

Jozef Juríček

Keywords: Information divergence, relative entropy, exponential family, information projection, hierarchical models, multi-information, multinomial distribution.

Abstract: The explicit solution of the problem of maximization of information divergence from the family of multinomial distributions is presented, using result of N. Ay and A. Knauf for the problem of maximization of multi-information [2], which is the special case of maximization of information divergence from hierarchical models [4].

The problem studied in this paper is a generalization of the binomial case, which was solved in [3].

The problem of maximization of information divergence from an exponential family has emerged in probabilistic models for evolution and learning in neural networks that are based on infomax principles [1].

The maximizers admit interpretation as stochastic systems with high complexity w.r.t. exponential family [2].

Abstrakt: Explicitní řešení problému maximalizace informační divergence od rodiny multinomických rozdělení bude prezentováno, s použitím výsledku N. Aye a A. Knaufa pro problém maximalizace multi-informace [2]. Jde o speciální podúlohu maximalizace informační divergence od hierarchických modelů [4].

Problém řešený v článku zobecňuje případ rodiny binomických rozdělení, který byl vyřešen v [3].

Úloha maximalizace informační divergence se objevila v pravděpodobnostních modelech pro evoluci a učení Bayesovských sítí, založených na principu infomaxu [1].

Maximalizátory jsou interpretovatelné jako stochastické systémy s vysokou mírou komplexity vzhledem k dané exponenciální rodine [2].

Literatura

[1] Ay, N. (2002) *An information-geometric approach to a theory of pragmatic structuring.* The Annals of Probability **30** (1), 416 – 436.

[2] Ay, N., Knauf, A. (2006) *Maximizing multi-information.* Kybernetika **45**, 517 – 538.

[3] Matúš, F. (2004) *Maximization of information divergences from binary i.i.d. sequences.* Proceedings of IPMU 2004, Perugia, **2**, 1303 – 1306.

[4] Matúš, F. (2009) *Divergence from factorizable distributions and matroid representations by partitions.* IEEE Transactions on Information Theory **55** (12), 5375 – 5381.

DIRECTIONAL QUANTILES

Lukáš Kotík

Keywords: Multivariate analysis, multivariate quantiles, data depth, nonparametric analysis, robust statistic, confidence sets.

Abstract: An univariate quantile plays an important role in the statistics and the data visualization. The presented paper proposes its possible generalization to the multivariate case. The proposed method is based on finding univariate quantiles along rays (directions) starting in some central point. We show basic properties of the proposed quantiles and its estimators.

Abstrakt: Kvantily patří mezi základní nástroje matematické statistiky a vizualizace dat. Bohužel neexistuje obecně uznávané rozšíření kvantilu pro vícerozměrná data. Článek ukazuje jednu z možností rozšíření pojmu kvantil do prostoru vyšších dimenzí. Postup je založen na určení jednorozměrných kvantilů na polopřímkách začínajících v jednom bodě, tzv. centru. Ukážeme si základní vlastnosti navrhovaného rozšíření jednorozměrných kvantilů a také možnosti jejich odhadu.

BOOTSTRAPPING OF M-SMOOTHERS

Matúš Maciak

Keywords: Nonparametric regression, local polynomial M-smoothers, change-point, smooth residual bootstrap, Mallow's metric.

Abstract: Asymptotic distribution of local polynomial M-smoothers depends on some unknown quantities. However, a knowledge of this distribution is crucial for a hypotheses testing problem in a change-point model. Instead of using some plug-in techniques, which provide a poor approximation, a bootstrap algorithm is proposed to approximate the unknown distribution and a proper justification of this algorithm is given. Finally, some results are illustrated through a proposed simulation study.

Abstrakt: Asymptotické rozdelenie lokálne polynomiálných M-vyhladzovačov závisí na niektorých neznámych kvantitách. Znalosť tohto rozdelenia je ale nutná k testovaniu hypotézy o prítomnosti bodu zmeny. Namiesto plug-in techník, ktoré poskytujú často len slabú aproximáciu a pomalú konvergenciu, použitie bootstrapových algoritmov býva často výhodnejším a správnejším rozhodnutím, to však musí byť dostatočne korektne preukázané. V prípade nášho modelu sme navrhli reziduálne založený hladký bootstrap a dôkaz fungovania tohto algoritmu je popísaný v článku. Na záver je algoritmus názorne aplikovaný na simulované data.

RATIO TYPE STATISTICS FOR DETECTION OF CHANGES IN MEAN AND THE BOOTSTRAP METHOD

Barbora Madurkayová

Keywords: Ratio type test statistics, block bootstrap, α-mixing.

Abstract: The paper presents procedures for detection of changes in mean. In particular test procedures based on ratio type test statistics that are functionals of partial sums of residuals are studied. We assume to have data obtained in ordered time points and study the null hypothesis of no change against the alternative of a change occurring at some unknown time point. We explore the possibility of applying the bootstrap method for obtaining critical values of the proposed test statistics and derive the limit behavior of the block bootstrap statistic for the L_2 procedure.

Abstrakt: V článku sú prezentované procedúry pre detekciu zmeny v strednej hodnote. Konkrétne ide o metódy založené na štatistikách podielového typu, ktoré sú funkcionálmi čiastočných súčtov reziduí. Predpokladáme, že máme dáta získané v časovo po sebe nasledujúcich okamihoch a testujeme nulovú hypoptézu o tom, že žiadna zmena nenastala, proti alternatíve, že zmena nastala v neznámom okamihu. Skúmame možnosť aplikácie metódy blokový bootstrap pre získanie kritických hodnôt navrhnutých testovacích štatistík a odvodíme limitné rozdelenie pre bootstrapovú štatistiku pre L_2 procedúru.

ESTIMATION OF INTERARRIVAL TIME DISTRIBUTION FROM SHORT TIME WINDOWS

Zbyněk Pawlas

Keywords: Distribution function estimation, interarrival time distribution, mixed Poisson process, point process, renewal process.

Abstract: We propose several estimators of interarrival time distribution based on observations of independent identically distributed stationary point processes in time windows with length of the same order as the mean interarrival time. This task is motivated by the situation in which a high number of neurons communicates with a target neuron. The comparison of the finite sample performance of the estimators is carried out by a simulation study for three selected models of point processes, namely Poisson point process, renewal process and mixed Poisson process.

Abstrakt: Navrhujeme několik odhadů rozdělení dob mezi událostmi na základě pozorování nezávislých, stejně rozdělených, stacionárních bodových procesů v časových oknech délky stejného řádu jako střední doba mezi událostmi. Tato úloha je motivována situací, ve které velký počet neuronů komunikuje s cílovým neuronem. Na základě simulační studie je provedeno porovnání kvality jednotlivých odhadů v případě konečného rozsahu výběru pro tři vybrané modely bodových procesů, a sice Poissonův bodový proces, proces obnovy a smíšený Poissonův bodový proces.

STRONGLY CONSISTENT ESTIMATION IN DEPENDENT ERRORS-IN-VARIABLES

Michal Pešta

Keywords: Errors-in-variables, dependent errors, strong consistency.

Abstract: *Errors-in-variables* (EIV) model with *dependent* errors is considered. A strong consistency of the *total least squares* (TLS) estimate for *weakly dependent* (α- and ϕ-mixing) measurements—encumbered with errors which are not necessarily stationary and identically distributed—is proved.

Abstrakt: Uvažujeme model *chyby-v-premenných* (EIV) so *závislými* chybami. Odvodíme silnú konzistenciu odhadu získaného metódou *úplne najmenších štvorcov* (TLS) pre *slabo závislé* merania (α- a ϕ-mixing) zaťažené nie nutne stacionárnymi a rovnako rozdelenými chybami.

WEAK $\sqrt{N}$-CONSISTENCY OF THE LEAST WEIGHTED SQUARES UNDER HETEROSCEDASTICITY

Jan Ámos Víšek

Keywords: Robustness, implicit weighting, weak $\sqrt{n}$-consistency of estimate by the least weighted squares, heteroscedasticity.

Abstract: Weak $\sqrt{n}$-consistency of the *Least Weighted Squares* estimator of the coefficients of regression model is proved generally under the heteroscedasticity of error terms. The assumptions required for the weak $\sqrt{n}$-consistency are briefly discussed. The roots of the heteroscedasticity are also critically considered.

Abstrakt: Článek dokazuje slabou $\sqrt{n}$-konsistenci odhadu (získaného metodou nejmenších vážených čtverců) koeficientů lineárního regresního modelu, a to obecně při přítomnosti heteroskedastcity. Předpoklady pro konsistenci jsou krátce diskutovány. Úvahy o zdrojích heteroskedasticity jsou rovněž uvedeny.

SOME APPLICATIONS OF TIME SERIES MODELS TO FINANCIAL DATA

Jitka Zichová

Keywords: Non-negative time series, autoregressive model, quality of forecasting, financial data, exchange rates

Abstract: Some special procedures for parameter estimation in non-negative autoregressive models were proposed in the literature and their small sample behavior investigated in simulation studies. These studies confirmed satisfactory convergence properties. The aim of this article is to study the forecasting quality on real data sets and to compare selected univariate and multivariate models estimated using the mentioned approach with models analyzed by means of standard methods. Some series of exchange rates from finance were used for this purpose.

Abstrakt: V literatuře byly během let navrženy postupy pro odhadování parametrů v nezáporných časových řadách a zkoumáno jejich chování. Vybrané vlastnosti byly též ověřovány pomocí simulačních studiií, jež mimo jiné prokázaly uspokojující konvergenční vlastnosti těchto metod. Cílem tohoto příspěvku je studovat kvalitu předpovědi vybraných finančních časových řad popisujících směnné kurzy pomocí modelů jedno a vícerozměrných nezáporných časových řad.

Obsah

Vybrané příspěvky z konference ROBUST 2010

Informační Bulletin České statistické společnosti vychází čtyřikrát do roka v českém vydání. Příležitostně i mimořádné české a anglické číslo. Časopis je zařazen na Seznamu Rady, více viz `http://www.vyzkum.cz/`.

Předseda společnosti: doc. RNDr. Gejza Dohnal, CSc.
ÚTM FS ČVUT v Praze, Karlovo náměstí 13, Praha 2, CZ-121 35
E-mail: `gejza.dohnal@fs.cvut.cz`

Redakční rada: prof. Ing. Václav Čermák, DrSc. (předseda), prof. RNDr. Jaromír Antoch, CSc., doc. Ing. Josef Tvrdík, CSc., RNDr. Marek Malý, CSc., doc. RNDr. Jiří Michálek, CSc., doc. RNDr. Zdeněk Karpíšek, CSc., prof. Ing. Jiří Militký, CSc.

Technický redaktor: ing. Pavel Stříž, Ph.D., `striz@fame.utb.cz`

Informace pro autory jsou na stránkách `http://www.statspol.cz/`

ISSN 1210–8022