Centrum Jaroslava Hájka pro teoretickou a aplikovanou statistiku

Výjezdní zasedání

Telč

25.–27. září 2008

Sborník rozšířených abstraktů

OBSAH

PŘEDMLUVA

Ve dnech 25.9.-27.9.2008 se konalo v Telči ve školícím středisku Masarykovy univerzity Výjezdní zasedání Centra Jaroslava Hájka pro teoretickou a aplikovanou statistiku. Členové řešitelského týmu z Masarykovy univerzity, Univerzity Karlovy a Technické univerzity v Liberci referovali o výsledcích řešení jednotlivých dílčích cílů.

V průběhu zasedání se vytvořila velmi příjemná pracovní atmosféra s plodnými diskusemi. V závěru tohoto zasedání se účastníci shodli na tom, že výjezdní zasedání by se mělo konat i v roce 2009.

10.12.2008

Ivana Horová
řešitel–koordinátor

NAVRHOVANIE A VYHODNOCOVANIE ZLOŽITÝCH MERANÍ, NOVÉ KALIBRAČNÉ POSTUPY, MODELOVANIE

GEJZA WIMMER

Kľúčové slová: kalibrácia, modelovanie.

Dielčí cieľ: V06

V dielčom cieli V06 sa prispelo k riešeniu

- kalibračného problému
- problematiky kľúčových medzilaboratórnych porovnávaní
- matematického modelovania v jazykovede
- problémov spojených s digitalizáciou údajov.

1. Kalibračný problém

Doterajšie prístupy k vyhodnoteniu meraní kalibrovaným meradlom sú v súčasnej dobe v mnohých oblastiach nedostačujúce (napr. v metrológii). Zaoberali sme sa komparatívnou kalibráciou, teda situáciou, keď jeden merací prístroj (meracia metóda) je kalibrovaný oproti druhému prístroju (metóde), pričom merania na obidvoch prístrojoch sú zaťažené chybami. Navrhol sa približný konfidenčný interval pre jedno alebo niekoľko meraní kalibrovaným meradlom v prípade lineárneho vzťahu medzi stupnicami meradiel. Navrhla sa konfidenčná oblasť pre parametre kalibračnej priamky aj v prípade korelovaných meraní. Simulačne sa overovali štatistické vlastnosti navrhovaných oblastí a ukazuje sa, že sú pre praktické účely vyhovujúce (dostatočne "úzke") a empiricky získané pravdepodobnosti pokrytia sú veľmi blízke teoretickým pre širokú oblasť parametrov. Výsledky sa prezentovali na konferencii ROBUST 2006, [1]. Pri riešení tejto úlohy vzniklo aj 8 prác K. Myškovej a I. Molla.

2. Kľúčové medzilaboratórne porovnávania

Existuje niekoľko prístupov k určeniu "Key Comparison Reference Value - KCRV" (metrologický (predpis, norma), frekventistický, bayesovsky, fiduciálny, atď.) Navrhla sa novú metódu určenia KCRV a jej rozšírenej neistoty, založená na tzv. metrologickom prístupe (rešpektujúcom metrologické normy) využívajúcom aj tzv. fiduciálny prístup. Systematické posuny laboratórií sa uvažovali ako realizácie normálne rozdelenej, rovnomerne rozdelenej alebo trojuholníkovo rozdelenej náhodnej veličiny a predpokladala sa heteroskedasticita. Pomocou rozsiahlej simulačnej štúdie sa ukázalo, že novonavrhnutý intervalový odhad vykazuje veľmi dobré štatistické vlastnosti s empirickou pravdepodobnosťou pokrytia blízkou k nominálnej hodnote pre každú uvažovanú distribúciu vychýlenia laboratórií. Výsledky sa prezentovali na 56. zasadaní ISI v Lisabone a sú publikované v [2],[3],[4].

S podporou MŠMT LC06024.

3. Matematické modelovanie v jazykovede

Vo výskume ide o novú cestu hľadania a (matematickú) formuláciu zákonitostí v empirických vedách na príkladoch z jazykovedy. Taktiež sa integruje množstvo jazykovedných zákonitostí do jednotnej teórie. Odhalili sa niektoré lingvistické zákonitostí (morfologická produktivita slovných kmeňov v jazyku, sémantická produktivita jazyka, teória slovných dĺžok) pomocou diskrétnych pravdepodobnostných modelov. Podarilo sa jednotne odvodiť (veľkú) triedu jazykovedných zákonov v dvoch prístupoch - diskrétnom a spojitom. Nové moderné smery matematicko-štatistického modelovania v jazykovede sa prezentovali na pozvanej prednáške na konferencii o výuke a aplikáciach štatistiky STAKAN 2007 v Rusave, 25-27.V.2007 a sú publikované v [5].

4. Problémy spojené s digitalizáciou údajov

Konštrukcia konfidenčného intervalu pre skutočnú meranú hodnotu v prípade digitalizovaných meraní je tiež doteraz nie k úplnej spokojnosti vyriešený problém. V literatúre je riešený čiastočne heuristicky za určitých teoretických predpokladov. Predpokladáme, že chyby meraní sú normálne rozdelené, nezávislé, s nulovou strednou hodnotou a rovnakým (neznámym) rozptylom. Skutočné výsledky merania (realizácie normálnych náhodných veličín) sú v digitálnom tvare. Keď disperzia chyby merania je veľká vzhľadom k digitalizačnému kroku, efekt digitalizácie zaniká. Analyzujeme situácie, keď je rozptyl chyby merania relatívne malý. Sformuloval sa model merania v prípade digitalizovaných údajov a odvodil približný konfidenčný interval pre skutočnú meranú hodnotu založený na ML odhade. Odvodil sa konfidenčný interval pre skutočnú meranú hodnotu tzv. fiduciálnym prístupom. Simulačne sa porovnávali štatistické vlastnosti konfidenčných intervalov (i) získaných ML prístupom, (ii) "štandardných", založených na Studentovom t rozdelení, ktorý ignoruje rozlíšiteľnosť meracieho prístroja, (iii) ktoré sú modifikovanou verziou konfidenčného intervalu navrhnutého Willinkom a založených na upravenom odhade disperzie, (iv) fiduciálnych. Výsledky sa prezentovali na konferencii TIES 2007 v Mikulove (poster) a ROBUST 2008 v Račkovej doline.

Literatúra

[1] Wimmer, G., Niektoré matematicko-štatistické modely kalibrácie, In.: Antoch, J., Dohnal, G., (Eds.) *ROBUST 2006*, Zborník prací, 26-29.

[2] Witkovský, V., Wimmer, G., Key Comparison Reference Value and Its Expanded Uncertainty Under Normally, Uniformly and Triangularly Distributed Laboratory Biases, *Bulletin of the International Statistical Institute 56th Session. Proceedings ISI 2007, Lisboa (CD ROM)*, 2007, 22-29.

[3] Witkovský, V., Wimmer, G., Confidence Interval for the Common Mean in Interlaboratory Comparisons with Systematic Laboratory Biases, *Measurement Science Review*, **6**, 2007, 64-73.

[4] Witkovský, V., Wimmer, G., Estimation of the Common Mean and Determination of the Comparison Reference Value, *Tatra Mt. Math. Publ.*, **39**, 2008, 53-60.

[5] Wimmer, G., Matematické modelovanie v jazykovede, *Informační Bulletin České statistické společnosti*, **19**, 2008, 1-17.

Gejza Wimmer,
Ústav matematiky a statistiky, Přírodovědecká fakulta MU,
Kotlářská 2, 611 37 Brno

e-mail: wimmer@math.muni.cz

SIGNAL OPTIMALITY – VIEW OF STATISTICS AND INFORMATION THEORY

PETR LÁNSKÝ AND ONDŘEJ POKORA

Klíčová slova: optimality of signal, Fisher information, mutual information
Particular aim: V05

Neuronal responses evoked in sensory neurons by static stimuli of various intensities are characterized by their input-output transfer function, i.e. by plotting the firing frequency (or any other measurable neuron response) versus the corresponding stimulus intensity or by plotting the ratio of activated receptors versus the concentration of ligand (signal). Stimulus intensities can be considered as "optimal" from two different points of view: to transfer as much information as possible and to code the intensity as precisely as possible. Variability is considered as a consequence of noise acting upon the transfer function and it may substantially influence the determination of the stimulus intensities considered as "optimal". For details see [4, 5].

To obtain the range of stimuli which can be identified from the transfer function with greatest precision, we propose to use measures based on Fisher information as known from the theory of statistical inference (as used e.g. by [3]) The Fisher information with respect to the signal s is defined by formula

$$J(s) = \int \left(\frac{\partial \ln g(r;s)}{\partial s} \right)^2 g(r;s)\,\mathrm{d}r \ \ ,$$

where $g(r;s)$ is the probability density of the response R and the signal s plays the role of deterministic parameter. The use of Fisher information as a tool to locate the optimal signal for information transfer is theoretically motivated by Cramer-Rao inequality for the signal as an estimated parameter. An approximation (lower bound) of the Fisher information can also be used, e.g.

$$J_2(s) = \frac{1}{\mathrm{Var}(R(s))} \left(\frac{\partial \mathrm{E}(R(s))}{\partial s} \right)^2 \ \ .$$

Classification of signals by their information content is very common in computational neuroscience (see [6]). For a given distribution $f(s)$ of the signal S, there is not a unique response but a family of responses $g(r|s)$ in dependency on realization of S. Bayes formula $g(r) = \int g(r|s) f(s)\,\mathrm{d}s$ gives the unconditional distribution $g(r)$ of the response. The "distance" between $g(r|s)$ and $g(r)$ can be expressed by using mutual information. One possibility is to use the formula

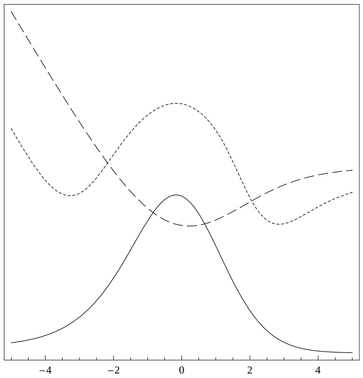
$$I_1(R|s) = \int g(r|s) \ln \frac{g(r|s)}{g(r)}\,\mathrm{d}r \ \ ,$$

Supported by MŠMT LC06024.

which was called "specific surprise" by [2] and mentioned e.g. by [1]. DeWeese and Meister ([2]) proposed a "stimulus-specific information"

$$I_2(R|s) = -\int g(r) \ln g(r)\, \mathrm{d}r + \int g(r|s) \ln g(r|s)\, \mathrm{d}r \ .$$

The optimal signal is located by searching for local maxima of mentioned measures of optimality. Comparing results of both approaches on theoretical and empirical models it can be shown that both the most identifiable signal and the most informative signal are not unique. An example of optimality criteria – Fisher information $J(s)$ (solid line) and information criteria $I_1(R|s)$ (dotted), $I_2(R|s)$ (dashed line) as functions of signal s in logarithmic scale for a binding model – are plotted in the figure. For detailed description see [4, 5].



References

[1] M. Bezzi, Quantifying the information transmitted in a single stimulus, *BioSystems*,**89**, 2007.

[2] M. R. DeWeese, M. Meister, How to measure the information gained from one symbol, *Network*, **10**, 1999.

[3] P. Lansky, and P. E. Greenwood, Optimal signal estimation in neuronal models, *Neural Computation*, **17**, 2005.

[4] P. Lánský, O. Pokora, J.-P. Rospars, Classification of stimuli based on stimulus-response curves and their variability, *Brain Research*, **122**, 2008.

[5] O. Pokora, P. Lánský, Statistical approach in search for optimal signal in simple olfactory neuronal models, *Mathematical Biosciences*, **214**, 2008.

[6] F. Rieke, D. Warland, R. de Ruyter Van Steveninck, W. Bialek, *Spikes: Exploring the Neural Code*, MIT Press, Cambridge, 1999.

[7] M. Stemmler, A single spike suffices: The simplest form of stochastic resonance in model neurons, *Network: Computation in Neural Systems*, **7**, 1996.

Petr Lánský,
Fyziologický ústav, Akademie věd ČR,
Vídeňská 1083, 142 20 Praha 4

Ondřej Pokora,
Ústav matematiky a statistiky, Přírodovědecká fakulta, Masarykova univerzita
Kotlářská 2, 611 37 Brno

e-mail: lansky@biomed.cas.cz, pokora@math.muni.cz

ODHAD NĚKTERÝCH CHARAKTERISTIK NEUROFYZIOLOGICKÝCH DAT

DAVID HAMPEL

Klíčová slova: diferenciální entropie, intervalově cenzorovaná data, reálný neuron, refrakterní perioda
Dílčí cíl: V05

V současnosti se v neurofyziologických vědách vynakládá mimořádné úsilí ve výzkumu v oblasti přenosu informace mezi neurony. Většina použitých přístupů k této otázce je do značné míry založena na přesné identifikaci času generování pulsu. Proto je důležité charakterizovat posloupnosti akčních potenciálů formálním popisem, který by ovšem odpovídal experimentálním poznatkům.

Mezi dobře známé a ověřené vlastnosti neuronů patří tzv. refrakternost. Absolutní refrakterní perioda, kterou se zabýváme, začíná po generování pulsu. V jejím průběhu není možné emitovat další puls za jakkoliv silného stimulu. Z hlediska statistického se jedná o odhad posunu hustoty rozdělení definované na polopřímce. Absolutní refrakterní periodu odhadujeme na základě časových intervalů mezi pulsy.

Srovnali jsme několik metod odhadu refrakterní periody (maximálně věrohodný, s minimálním rizikem, odhady založené na odhadu celého nosiče hustoty, odhad minimální hodnotou) v kombinaci se třemi modely generování mezipulsních časových intervalů. Navrhli jsme a ověřili také odhad posunutí na základě adjustované minimální hodnoty. Zkoumané metody jsme rozdělili na skupinu parametrických a neparametrických metod. V obou případech jsme srovnávali jejich přesnost a asymptotické vlastnosti na simulovaných datech. Z výsledků plyne, že aplikace neparametrických metod, konkrétně Cookova odhadu, dává nejstabilnější výsledky. Na druhou stranu, žádná z metod se nejeví zřetelně lépe než ostatní metody. Můžeme ale konstatovat, že odhad pomocí minimální hodnoty, obvykle používaný pro refrakterní periodu, není nejlepší volbou pro neurofyziologická data. Analýza odhadů pro experimentální data poukázala na omezené možnosti užití jednotlivých metod. Na základě provedené studie můžeme říci, že posun (refrakterní perioda) může být dobře odhadnut pouze pro stimulovaná data.

Dalším objektem našeho zájmu byla přímo komunikace mezi neurony. S využitím teorie informace, konkrétně pomocí konceptu diferenciální entropie, můžeme

- identifikovat změny v chování neuronu a
- srovnat chování dvou či více neuronů

za různých experimentálních situací. Pro tento účel je třeba mít k dispozici kvalitní odhady diferenciální entropie. Tyto lze použít k výpočtu Kullback-Leiblerovy vzdálenosti, pomocí níž lze určit případné změny v chování neuronu či rozdíly mezi neurony. Vezmeme-li v úvahu refrakterní vlastnost, vyvstává otázka zda a popřípadě jak přítomnost refrakterní periody ovlivní odhad diferenciální entropie.

S podporou M©MT LC06024.

Pozornost jsme věnovali především plug-in metodám a tzv. přímým metodám (odhad založený na rozestupu vzorků, Vasickuv odhad) odhadu entropie. V praktických úlohách je potřebné odhadnout diferenciální entropii v relativně krátkém čase a navíc na základě malého rozsahu dat. K tomuto účelu by mohly posloužit námi diskutované rychlé plug-in odhady založené na momentovém odhadu hustoty. Vzhledem k dobrým výsledkům při jejich nasazení jsme na jejich základě odvodili přímý odhad a diskutovali jeho vlastnosti. Podobně jako při analýze odhadů refrakterní periody i zde jsme provedli rozsáhlou studii na simulovaných datech.

Experimentální záznam neuronální aktivity bývá komplikován faktem, že je nemožné pořídit časově neomezený záznam. Konkrétně, pokud zaznamenáváme neuronální aktivitu po nějakém stimulu, je reakce neuronu pouze dočasná. Abychom získali dostatečné množství dat, musíme tento experiment několikrát opakovat. Další podobná situace nastane, pokud sledujeme větší množství neuronů najednou a chceme identifikovat možnou neuronální aktivitu ihned po stimulu.

V obou případech dostaneme množství relativně krátkých intervalově cenzorovaných záznamů neuronální aktivity namísto jednoho dlouhého záznamu. Z těchto vzorků chceme opět odhadnout diferenciální entropii a poté spočítat Kullback-Leiblerovu vzdálenost od rozdělení neuronální aktivity bez stimulu. K odhadu entropie z intervalově cenzorovaných dat můžeme přikročit několika strategiemi.

(1) Můžeme pracovat s intervalově cenzorovanými daty, jako by šlo o kompletní pozorování. Budeme moci využít všechny možné metody odhadu entropie, nicméně se vědomě dopouštíme chyby, když považujeme neúplné vzdáleností mezi pulzy za úplné.
(2) Dále můžeme vyloučit z dat neúplná pozorování, a entropii odhadovat pouze na základě úplných. Podle situace nám zbude relativně málo dat, nutno bude použít metody vhodné pro malé vzorky. Zbavujeme se části informace, naše výpočty budou méně přesné než při další uvažované strategii.
(3) Zohledníme charakter dat, a použijeme adekvátní, byť složitou, metodu. Využitelné budou jen plug-in odhady, do kterých dosadíme vhodný odhad hustoty.

Posledním případem se zabýváme. Analyzujeme výstupy dvou metod odhadu používajících EM algoritmus, které jsme přizpůsobili pro náš typ dat. Odhad entropie na základě těchto metod je vždy lepší než odhady nerespektující podstatu cenzorovaných dat.

Reference

[1] Dempster, A. P., Laird, N. M., Rubin, D. B., Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, **Vol. 39 (1)**, 1977.

[2] Hampel, D., Estimation of differential entropy for positive random variables and its application in computational neuroscience. In: *Mathematical Modeling of Biological Systems. Vol. II.* Birkhauser, Boston, 2007.

[3] Hampel, D., Lánský, P., On the estimation of refractory period, *Journal of Neuroscience Methods*, **Vol. 171**, 2008.

David Hampel,
Ústav matematiky a statistiky PřF MU,
Kotlářská 2, 611 37 Brno

e-mail: david.hampel@ukzuz.cz

CONVOLUTIONS OF KERNELS AND THEIR USING

JIŘÍ ZELINKA

ABSTRACT. Convolutions can be found in some expressions describing properties of all sorts of kernel estimates. The bias of kernel estimate of probability density function is one of these expressions. It can be approximated (Jones et al., 1991) by the integral estimate whose evaluating leads to the convolution of the kernel function with itself. This paper presents the exact construction of convolution of polynomial kernels, properties of this convolution and application of it for the iterative method of bandwidth choice (Horová & Zelinka, 2007).

Keywords: kernel estimate of density, convolution
Particular aim: V04

1. Introduction

Let us denote S_k the class of the polynomial kernels K of order k satisfying the the usual moment conditions (see e.g. Horová & Zelinka (2007)).

Epanechnikov and quartic kernels are typical examples of kernels belonging to S_2.

The standard kernel estimate of the density (e.g Wand & Jones (1995)) is defined as

$$\hat{f}_{h,K}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right), \quad K \in S_k. \tag{1}$$

The smoothing parameter h is called bandwidth and it mostly influences the shape of the estimate.

2. Properties of the estimate

Basic properties of kernel estimate of the density are given by the *Mean Integrated Square Error* ($MISE$). The leading term of $MISE$ is

$$\overline{MISE}\ \left(\hat{f}_{h,K}\right) = \frac{1}{nh}V(K) + h^{2k}\beta_k^2(K)D_k\,. \tag{2}$$

Bandwidth minimizing $\overline{MISE}$ is called *optimal bandwidth* and it is denoted by $h_{opt}(\hat{f}_{h,K})$ or shortly h_{opt}. It is easy to derive the relationship

$$h_{opt}^{2k+1} = \frac{V(K)}{2nkD_k\beta_k^2(K)}\,. \tag{3}$$

3. Convolution of kernels

3.1. Basic properties.

If we denote as

$$K_h(x) = \frac{1}{h}K\left(\frac{x}{h}\right)$$

the expectation of the density estimate can be evaluated as

$$E\hat{f}_{h,K}(x) = \int K_h\left(x-y\right)f(y)dy = (K_h * f)(x) \tag{4}$$

and the bias of the estimate as

$$bias\hat{f}_{h,K}(x) = (K_h * f)(x) - f(x). \tag{5}$$

Müller & Wang (1990) suggested the following estimate of the $bias\hat{f}_{h,K}(x)$ derived from (5):
(6)

$$\widehat{bias}\hat{f}_{h,K}(x) = (K_h * \hat{f}_{h,K})(x) - \hat{f}_{h,K}(x) = \frac{1}{nh}\sum_{i=1}^{n}\left[\mathcal{C}_K\left(\frac{x-X_i}{h}\right) - K\left(\frac{x-X_i}{h}\right)\right],$$

where

$$\mathcal{C}_K(u) = \int K\left(u-y\right)K(y)dy = (K * K)(u)\,.$$

Theorem
The convolution $\mathcal{C}_K$ is continuous and piecewise polynomial function for any polynomial kernel K, support$(\mathcal{C}_K) = [-2,2]$, $\int \mathcal{C}_K(x)dx = 1$. If K is non-negative or even on $\mathbb{R}$ then the same property is valid for $\mathcal{C}_K$.

Let $K \in S_k$ and $K \in C^s(\mathbb{R})$, $s \geq 0$. Then $C_K \in C^{2s+2}(\mathbb{R})$ and the derivatives of $\mathcal{C}_K$ can be expressed as

$$\begin{array}{rcl} \mathcal{C}'_K & = & K' * K \\ \mathcal{C}''_K & = & K' * K' \\ & \vdots & \\ \mathcal{C}^{(2s)}_K & = & K^{(s)} * K^{(s)} \\ \mathcal{C}^{(2s+1)}_K & = & K^{(s+1)} * K^{(s)} \\ \mathcal{C}^{(2s+2)}_K & = & K^{(s+1)} * K^{(s+1)} \end{array}$$

3.2. **Convolutions of higher order.** Let us denote by
$\mathcal{C}_{K,1} = K$,
$\mathcal{C}_{K,2} = K * K = \mathcal{C}_K$,
$\mathcal{C}_{K,3} = K * K * K = \mathcal{C}_K * K$,
$\mathcal{C}_{K,4} = K * K * K * K = \mathcal{C}_K * \mathcal{C}_K$,
Estimate of $\int bias^2\hat{f}_{h,K}(x)dx$ (see (6)) can be evaluated as

$$\begin{array}{rcl} \int \widehat{bias}^2\hat{f}_{h,K}(x)dx & = & \int \Big((K_h * \hat{f}_{h,K})(x) - \hat{f}_{h,K}(x)\Big)^2 dx = \\ = & \frac{1}{n^2h}\sum\limits_{i,j=1}^{n} & \left[\mathcal{C}_{K,4}\left(\frac{X_j-X_i}{h}\right) - 2\mathcal{C}_{K,3}\left(\frac{X_j-X_i}{h}\right) + \mathcal{C}_{K,2}\left(\frac{X_j-X_i}{h}\right)\right] \end{array}$$

for even kernels.

4. Application for bandwidth selection

The iterative method for bandwidth selection was suggested in Horová & Zelinka (2007). The estimate $\hat{h}_{opt}$ of the optimal bandwidth can be obtained as the solution of the equation

$$h = \frac{V(K)}{2nk\int \widehat{bias}^2 \hat{f}_{h,K}(x)dx}\,. \tag{7}$$

Steffensen's method and numerical calculation of the integral were used in the mentioned paper. But the integral can be evaluated exactly using kernel convolutions. Equation (7) can be rewritten as

$$h = \frac{V(K)}{2nk\frac{1}{n^2h}\sum\limits_{i,j=1}^{n}\left[\mathcal{C}_{K,4}\left(\frac{X_j-X_i}{h}\right) - 2\mathcal{C}_{K,3}\left(\frac{X_j-X_i}{h}\right) + \mathcal{C}_{K,2}\left(\frac{X_j-X_i}{h}\right)\right]}$$

or

$$\Phi(h) = \frac{2k}{n}\sum_{i,j=1}^{n}\mathcal{C}\left(\frac{X_j-X_i}{h}\right) - V(K) = 0 \tag{8}$$

for

$$\mathcal{C}(x) = \mathcal{C}_{K,4}(x) - 2\mathcal{C}_{K,3}(x) + \mathcal{C}_{K,2}(x)\,.$$

For solving (8) Newton's method can be used as the derivative of function Φ is easily obtained:

$$\Phi'(h) = -\frac{2k}{nh^2}\sum_{i,j=1}^{n}(X_j-X_i)\mathcal{C}'\left(\frac{X_j-X_i}{h}\right),$$

$$\mathcal{C}'(x) = \mathcal{C}'_{K,4}(x) - 2\mathcal{C}'_{K,3}(x) + \mathcal{C}'_{K,2}(x)$$

taking into account that the convolution are piecewise polynomials. Then the Newton's method gets the form

$$h_{l+1} = h_l - \frac{\Phi(h_l)}{\Phi'(h_l)} = h_l + h_l^2\frac{\sum\limits_{i,j=1}^{n}\mathcal{C}\left(\frac{X_j-X_i}{h}\right) - \frac{n}{2k}V(K)}{\sum\limits_{i,j=1}^{n}(X_j-X_i)\mathcal{C}'\left(\frac{X_j-X_i}{h}\right)}\,.$$

The maximal smoothing principle (see Horová & Zelinka (2007)) can be applied for the initial approximation h_0.

References

Horová, I., Zelinka, J., 2007. Contribution to the Bandwidth Choice for Kernel Density Estimates, Computaional Statistics 22, 31–47.

Jones, M.C., Marron, J.S., Park, B.U., 1991. A simple root-n bandwidth selector. Annals of Statistics 19, 1919–1932.

Müller, H.G., Wang, J.L., 1990. Locally addaptive hazard smoothing. Prob. Th. Rel. Fields 85, 523–538.

Wand, I.P., Jones, I.C., 1995. Kernel smoothing. Chapman & Hall, London.

Department of Mathmatics and Statistics,
Faculty of Science, Masaryk University,
Kotlářská 2, 611 37 Brno, Czech Republic e-mail: zelinka@math.muni.cz

BOUNDARY EFFECTS IN KERNEL CDF ESTIMATION

JAN KOLÁČEK

ABSTRACT. In this presentation we focus on kernel estimates of cumulative distribution functions in case that random variables $X_1, \dots X_n$ are nonnegative. It is well known that kernel distribution estimators are not consistent when estimating a distribution function near the point $x = 0$. This fact is regrettable in many applications, for example in kernel ROC curve estimation (Koláček and Karunamuni (2007)). In order to avoid this problem we propose a bias reducing technique which is a kind of generalized reflection method. Our method is based on ideas of Karunamuni and Alberts (2005) and Zhang et al. (1999) developed for boundary correction in kernel density estimation. Finally, the proposed estimator is compared with the traditional kernel estimator and with the estimator based on "classical" reflection method using simulation studies.

Klíčová slova: kernel estimation, reflection, distribution estimation.
Particular aim: V04

1. INTRODUCTION

The most commonly used nonparametric estimate of a cumulative distribution function F is an empirical distribution function F_n. But F_n is a step function even in case that F is continuous. Another type of nonparametric estimators for F is derived from kernel smoothing methods. Kernel smoothing is most widely used because it is easy to derive and has good properties. Kernel smoothing has received a lot of attention in density estimation. Good references in this area are Gasser at al. (1985), Silverman (1986) and Wand and Jones (1995). However, results in kernel distribution function estimation are relatively few. Theoretical properties of kernel distribution function estimator have been investigated by Nadaraya (1964), Reiss (1981) and Azzalini (1981). Although there is a vast literature on boundary correction in density estimation context, boundary effects problem in distribution function context has been less studied.

In this presentation, we develop a new kernel type estimator of the cumulative distribution function that removes boundary effects near the end points of the support. Our estimator is based on a new boundary corrected kernel estimator of distribution functions and it is based on ideas of Karunamuni and Alberts (2005) and Zhang et al. (1999) developed for boundary correction in kernel density estimation. The basic technique of construction of the proposed estimator is kind of a generalized reflection method involving reflecting a transformation of the observed data. In fact, the proposed method generates a class of boundary corrected estimators. We derive expressions for the bias and variance of the proposed estimator. Furthermore, the proposed estimator is compared with the traditional estimator and with the estimator based on "classical" reflection method using simulation

Supported by MSMT LC06024.

studies. We observe that the proposed estimator successfully remove boundary effects and performs considerably better than the others two.

Kernel smoothing in distribution function estimation and boundary effects are discussed in the first part of the presentation. The proposed estimator is introduced in the next part. Finally, some simulation results are given at the end of the presentation.

2. Conclusion

In this presentation we proposed a new kernel-type distribution estimator to avoid the difficulties near the boundary. The technique implemented is a kind of generalized reflection method involving reflecting a transformation of the data. The proposed method generates a class of boundary corrected estimators and it is based on ideas of boundary corrections for kernel density estimators presented in Karunamuni and Alberts (2005). We showed some good properties of our proposed method (e.g., local adaptivity). Furthermore, it is shown that bias of the proposed estimator is better than that of the "classical" case.

Acknowledgements: The research was supported by The Jaroslav Hájek center for theoretical and applied statistics (MSMT LC 06024).

References

[1] Azzalini, A., A note on the estimation of a distribution function and quantiles by a kernel method, *Biometrika*, **Vol. 68**, No 1, pp. 326–328, 1981.

[2] Bowman, A., Hall, P., Prvan, T., Bandwidth selection for the smoothing of distribution functions, *Biometrika*, **Vol. 85**, No 4, pp. 799–808, 1998.

[3] Gasser, T., Müller H.G. and Mammitzsch V., Kernels for nonparametric curve estimation, *Journal of the Royal Statistical Society*, Series B, **Vol. 47**, No. 2, 238–252, 1985.

[4] Karunamuni, R.J., Alberts T., On boundary correction in kernel density estimation, *Statistical Methodology*, **Vol. 2**, pp. 191–212, 2005.

[5] Nadaraya, E.A., Some new estimates for distribution functions. *Theory Prob. Appl.*, **Vol. 15**, 497–500, 1964.

[6] Reiss, R.D., Nonparametric estimation of smooth distribution functions, *Scandinavian Journal of Statistics*, **Vol. 8**, 116–119, 1981.

[7] Silverman, B.W.: *Density estimation for statistics and Data Analysis*, Chapman and Hall, New York, 1986.

[8] Wand, I.P. and Jones, I.C.: *Kernel smoothing*, Chapman & Hall, London, 1995.

[9] Zhang, S., Karunamuni, R.J., Jones, M.C., An improved estimator of the density function at the boundary, *Journal of the Amer. Stat. Assoc.*, **Vol. 448**, pp. 1231–1241, 1999.

Jan Koláček,
Department of Mathematics and Statistics,
Kotlářská 2, Brno, 611 37

e-mail: kolacek@math.muni.cz

KERNEL ESTIMATES OF ROC CURVES

JIŘÍ ZELINKA, IVANA HOROVÁ

ABSTRACT. The Receiver Operating Characteristic (ROC) curve describes the performance of a diagnostic test which classifies subjects into either group without condition $\mathcal{G}_0$ or group with condition $\mathcal{G}_1$ by means of a continuous discriminant score X. The present paper aims to estimate the ROC curve as e special case of the cumulative distribution function (c.d.f.) by means of kernel methods. The beta distribution is used for construction of this estimate.

Keywords: ROC curve, beta distribution, kernel estimate
Particular aim: V04

1. INTRODUCTION

Let $\mathcal{G}_1$ be group of n_1 subjects with a condition and $\mathcal{G}_0$ group of n_0 subjects without a condition, $D = 0, 1$ random variable denotes absence or presence of the condition. Let the falling into particular groups is tested by a test that gives as result the random variable T: $T = 1$ positive test result, $T = 0$ negative test result. Results of the test can be presented by the confusion matrix:

	Positive test, $T = 1$	**Negative test,** $T = 0$	**Total**
$\mathcal{G}_1$ $(D = 1)$	True positive (TP)	False negative (FN)	$TP + FN$
$\mathcal{G}_0$ $(D = 0)$	False positive (FP)	True negative (TN)	$FP + TN$
Total	$TP + FP$	$FN + TN$	$n = n_0 + n_1$

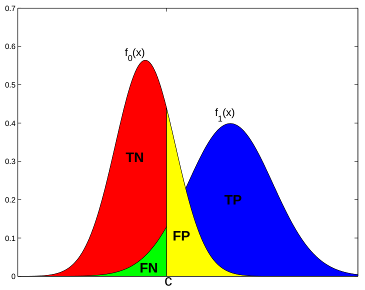
Let X be the diagnostic test variable (one-dimensional absolutely continuous random variable) and c – given cutoff point, $c \in \mathbb{R}$. The subject is classified as $\mathcal{G}_1$ if $X \geq c$ and $\mathcal{G}_0$ otherwise for given cutoff point c.

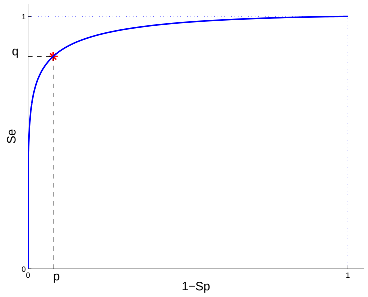
$$F_0(c) = P(X \leq c|\mathcal{G}_0) = \int\limits_{-\infty}^{c} f_0(x)dx, \quad F_1(c) = P(X \leq c|\mathcal{G}_1) = \int\limits_{-\infty}^{c} f_1(x)dx$$

F_0 or F_1 are c.d.f.s of group $\mathcal{G}_0$ or $\mathcal{G}_1$ and f_0 and f_1 are corresponding probability density functions (p.d.f.).

F_0: the specificity (Sp) of the test
$1 - F_1$: the sensitivity (Se) of the test

ROC curve is displayed by plotting $\underbrace{1 - F_1(c)}_{Se}$ against $\underbrace{1 - F_0(c)}_{1-Sp}$ for a range of cutoff points $c \in \mathbb{R}$.





TP – True positive FP – False positive
FN – False negative TN – True negative

$$\begin{aligned} p &= 1 - F_0(c) \\ q &= 1 - F_1(c) \end{aligned}$$

2. Nonparametric estimates of cumulative distribution function

Let $Z_1, \ldots, Z_n$ be random sample from random variable Z with c.d.f. F.

Empirical distribution function: $\hat{F}_{n,Z}(x) = \frac{1}{n}\sum\limits_{i=1}^{n} I(Z_i \le x)$.

The ROC curve $ROC(p) = R(p)$ is the c.d.f. of random variable $Y = 1 - F_0(X_1)$:

$$\begin{aligned} F_Y(p) &= P(Y \le p) = P(1 - F_0(X_1) \le p) = \\ &= P(X_1 \ge F_0^{-1}(1-p)) = 1 - F_1(F_0^{-1}(1-p)) = \\ &= R(p) \end{aligned}$$

Properties of Y:

- $Y \in [0, 1]$
- $E(Y) = 1 - \int\limits_0^1 R(p)dp = 1 - AUC(R)$
- $D(Y) = var(Y) = E(Y^2) - (E(Y))^2 = 1 - 2\int\limits_0^1 p\,R(p)dp - \big(1 - AUC(R)\big)^2 =$ $= 2\int\limits_0^1 (1-p)\,R(p)dp - AUC^2(R)$

Therofore we focus to kernel estimates of c.d.f.:

$$\hat{F}_h(x) = \frac{1}{n}\sum_{i=1}^{n} W\left(\frac{x - Z_i}{h}\right), \qquad W(x) = \int_{-1}^{x} K(t)dt$$

K is a non-negative symmetric function called kernel, supported on $[-1, 1]$, integrated to unity, h is a smoothing parameter called *bandwidth*

Bandwidth selection

Mean Integrated Square Error $MISE$ is

$$MISE(\hat{F}_h) = \int E(\hat{F}_h(x) - F(x))^2 dx,$$

The leading term of MISE:

$$\overline{MISE}(\hat{F}_h) = \underbrace{\frac{1}{n}\int F(x)(1-F(x))dx - c_1\tfrac{h}{n}}_{\overline{Ivar}(\widehat{F}_{h,K})} + \underbrace{\frac{\beta_2^2}{4}\psi_2 h^4}_{\overline{Ibias}^2(\widehat{F}_{h,K})} \ ,$$

$$c_1 = \int_{-1}^{1} W(x)(1-W(x))dx > 0, \quad \beta_2 = \int_{-1}^{1} x^2 K(x)dx, \quad \psi_2 = \int (F''(x))^2 dx$$

Optimal bandwidth minimizing $\overline{MISE}(\hat{F}_h)$ provided that $F \in C^2$:

$$h_{opt} = n^{-1/3}\left(\frac{c_1}{\beta_2^2\psi_2}\right)^{1/3}$$

3. Direct kernel estimate of ROC curve

Let us construct the random sample of the estimate of Y:

$$\hat{Y}_E(x) = 1 - \hat{F}_{n_0,X_0}(x) = \frac{1}{n_0}\sum_{j=1}^{n_0} I(X_{0,j} \le x)$$

for $x = X_{1,1}, \dots, X_{1,n_1}$. The estimate of ROC curve is

$$\hat{R}_E(p) = \hat{F}_{\hat{Y}_E,h_1}(p) = \frac{1}{n_1}\sum_{i=1}^{n_1} W\left(\frac{p - \hat{Y}_{E_i}}{\tilde{h}_1}\right)$$

For bandwidth selection (unknown value ψ_2) we will use the reference c.d.f. F_r or p.d.f. f_r giving the same expectation and variance as $\hat{Y}_E$.

As the refernce p.d.f. let us use the beta distribution:

$$f_r(x) = f(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}$$

$$E(\hat{Y}) = \frac{\alpha}{\alpha+\beta}, \quad \sigma^2_{\hat{Y}} = D(\hat{Y}) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

$$\alpha = E(\hat{Y})\left(\frac{E(\hat{Y})(1-E(\hat{Y}))}{\sigma^2_{\hat{Y}}} - 1\right), \quad \beta = (1-E(\hat{Y}))\left(\frac{E(\hat{Y})(1-E(\hat{Y}))}{\sigma^2_{\hat{Y}}} - 1\right)$$

We use estimates $\hat{E}(\hat{Y})$ and $\hat{\sigma}^2_{\hat{Y}}$ and rounded value of α and β in real calculations for easy evaluation of $\widehat{\psi}_2$.

References

[1] Altman, N., Léger, Ch.: *Bandwidth selection for kernel distribution function estimation*, Journal of Stat. Planning and Inference, **46**, pp. 195–214, 1995.

[2] Bowman, A., Hall, P., Prvan, T.: *Bandwidth selection for the smoothing of distribution functions*. Biometrika, **85**, No. 4, pp. 799–808, 1998.

[3] Hall, P.H., Hyndman R.J.: *Improved methods for bandwidth selection when estimating ROC curves*. Statistics & Probability Letters, **64**, pp 181–189, 2003.

[4] Horová I., Zelinka J.: *Contribution to the bandwidth choice for kernel density estimates*, Computational Statistics, **22**, No. 1, pp. 31–47, 2007.

[5] Horová I. Koláček J., Zelinka, J., El-Shaarawi A.: *Smooth Estimates of Distribution Functions with Application in Environmental Studies*, Accepted to MABE'08, 2007.

[6] Lloyd, C.J., Zhou Yong: *Kernel estimators of the ROC curve are better than empirical.* Statistics and Prob. Letters 44, pp. 221–228, 1999.

[7] Pepe, M. S.:

[8] Wand, I.P., Jones, I.C.: Kernel smoothing. Chapman & Hall, London, 1995.

[10] Zhou X.–H., Harezlak J.: *Comparison of bandwidth selection methods for kernel smoothing of ROC curves*, Statistics in Medicine, **21**, 2045–2055, 2002.

Department of Mathmatics and Statistics,
Faculty of Science, Masaryk University,
Kotlářská 2, 611 37 Brno, Czech Republic e-mail: zelinka@math.muni.cz

JÁDROVÉ ODHADY VÍCEROZMĚRNÝCH HUSTOT

KAMILA VOPATOVÁ

Klíčová slova: jádrové vyhlazování, vícerozměrné hustoty, AMISE.
Dílčí cíl: V04

Pro d-rozměrný náhodný výběr $\mathbf{X}_1, \ldots, \mathbf{X}_n$ definujeme jádrový odhad hustoty rozdělení, z něhož výběr pochází, jako

$$\hat{f}(\mathbf{x}; H) = \frac{1}{n} \sum_{i=1}^{n} |H|^{-1/2} K\big(H^{-1/2}(\mathbf{x} - \mathbf{X}_i)\big),$$

kde K je d-rozměrná jádrová funkce a H je vyhlazovací matice.

Nejzávažnějším úkolem jádrového vyhlazování je nalezení matice H. Kritériem pro volbu optimální matice H je střední integrální kvadratická chyba MISE, resp. AMISE – asymptotický tvar MISE. Optimální vyhlazovací matice H minimalizuje tuto chybu, $H = \arg\min_H AMISE(H)$.

Pokud se omezíme na diagonální matice H řádu 2, tj. $H = \left(\begin{smallmatrix} h_1^2 & 0 \\ 0 & h_2^2 \end{smallmatrix}\right)$, a zavedeme odhad asymptotické střední integrální kvadratické chyby

$$\widehat{AMISE}(H) = \iint \widehat{var} \hat{f}(x_1, x_2; H)\, \mathrm{d}x_1 \mathrm{d}x_2 + \iint \Big(\widehat{bias} \hat{f}(x_1, x_2; H)\Big)^2\, \mathrm{d}x_1 \mathrm{d}x_2,$$

kde

$$\iint \widehat{var} \hat{f}(x_1, x_2; H)\, \mathrm{d}x_1 \mathrm{d}x_2 = \frac{1}{n} |H|^{-1/2} V(K),$$
$$\iint \Big(\widehat{bias} \hat{f}(x_1, x_2; H)\Big)^2\, \mathrm{d}x_1 \mathrm{d}x_2 = \frac{1}{n^2} |H|^{-1/2} g(h_1, h_2),$$

pak užitím vztahu mezi rozptylem a vychýlením obdržíme rovnici

$$\frac{n}{2} V(K) = g(h_1, h_2).$$

Druhým vztahem pro neznámé h_1, h_2 je jednoduchá rovnice $h_2 = c\, h_1$. Konstantu c odhadneme užitím Scottova pravidla $\hat{h}_i = \hat{\sigma}_i n^{-1/6}$, tedy

$$h_2 = \hat{c}\, h_1, \qquad \hat{c} = \frac{\hat{\sigma}_2}{\hat{\sigma}_1}.$$

Prvky h_1, h_2 optimální vyhlazovací matice H jsou pak řešením soustavy rovnic

$$\frac{n}{2} V(K) = g(h_1, h_2), \qquad h_2 = \frac{\hat{\sigma}_2}{\hat{\sigma}_1} h_1.$$

S podporou MŠMT LC06024.

Reference

[1] Horová, I., Zelinka, J., Contribution to the Bandwidth Choice for Kernel Density Estimates, *Comput. Statist.*, **22**, 2007.

[2] Scott, D. W., *Multivariate Density Estimation: Theory, Practise, and Visualization*, John Wiley and Sons, New York, 1992.

[3] Wand, M. P., Jones, M. C., *Kernel Smoothing*, Chapman and Hall, London, 1995.

Kamila Vopatová,
Ústav matematiky a statistiky PřF, Masarykova univerzita,
Kotlářská 2, 611 37 Brno

e-mail: vopatova@mail.muni.cz

TESTS BASED ON REGRESSION RANK SCORES

MARTIN SCHINDLER

Keywords: Regression rank scores, regression quantiles, nonlinear regression.
Particular aim: V02

We focuse on inference based on regression rank scores, particularly on the construction of tests that are functions of regression rank scores. Regression rank scores could hardly be discovered without regression quantiles which are useful for the interpretation of the regression rank scores (and necessary for their definition in the nonlinear model). Thus, before defining the scores, we always mention the regression quantiles.

This work is divided into two parts called *Linear model* and *Nonlinear model*.

1. Linear model

The first part concentrates on the tests in the linear regression model and in the second part we try to extend some of these tests for a nonlinear regression model as well. Both parts consist of three chapters.

The tests presented in the first part are asymptotically distribution free. Firstly we give a motivation and an introduction to the theory of the tests based on regression rank scores, explain that these tests naturally generalize ordinary rank tests and more or less we summarize the known facts about the regression quantiles, regression rank scores and tests based on them in the linear model.

Next we concentrate on a few specific situations of one-way and two-way ANOVA models for which we construct tests described in the foregoing. We will also find out that such tests as Kruskal-Wallis or Friedman tests are in fact based on regression rank scores too.

Finally we deal with a different class of tests based on rank scores. We introduce the Kolmogorov-Smirnov type test when a nuisance linear regression is present and show that the two-sample variant of it is an extension of the classical Kolmogorov-Smirnov test. Moreover, its asymptotic distributions under the hypothesis and the local alternatives coincide with those of the classical test. Similarly, a two-sample Cramér-von Mises type test under a nuisance regression is derived.

2. Nonlinear model

In the second part we try to use a similar concept to build some tests also in the nonlinear regression model. The situation in the nonlinear model is, however, much more complicated, so we pay more attention to finding out the properties of nonlinear regression quantiles and nonlinear regression rank scores.

Next we introduce a class of tests based on nonlinear regression rank scores when a nuisance nonlinear regression is present. The asymptotic distribution as well as its verification by a simulation is included.

S podporou MŠMT LC06024.

Finally, a test for independence with a nuisance nonlinear regression is constructed.

Most of the tests introduced here are followed by at least one example on which we demonstrate their utilization. Computational problems in the nonlinear models are mentioned and a method how to work them out is proposed and applied to real data.

References

[1] Gutenbrunner C., Jurečková J., Koenker R. and Portnoy S., Tests of linear hypotheses based on regression rank scores, *J. Nonparametr. Statist.*, **2**, 1993.

[2] Jurečková J., Regression rank scores in nonlinear model, *In: Beyond Parametrics in Interdisciplinary Research: Festschrift in honor of Professor Pranab K. Sen (N. Balakrishnan, E. A. Peña and M. J. Silvapulle, eds.). Institute of Mathematical Statistics Collections*, **1**, 2008.

[3] Hájek J., Šidák Z. , *Theory of Rank Tests*,Academia, Praha, 1967.

Martin Schindler,
Technical University of Liberec,
Hálkova 6, 461 17 Liberec 1

e-mail: martin.schindler@tul.cz

CONSTRUCTION OF STABLE WAVELET BASES

DANA ČERNÁ, VÁCLAV FINĚK

Keywords: wavelet, stabilization
Particular aim: V03

Wavelets are by now a widely accepted tool in signal and image processing as well as in numerical simulation, statistics and engineering applications. In the field of statistics they are used especially for denoising, regression and density estimation. Originally, wavelet methods were applied to problems defined on the whole Euclidean space or on the torus. However in many applications biorthogonal wavelet bases defined on a bounded domain are needed. The starting point of the construction of wavelet bases on general domain is the construction of wavelet bases on the interval. It consists in retaining basis functions from $L^2(\mathbb{R})$ whose support is contained in the interval and suitable adaptation near the boundary. Such bases were constructed in [4]. These bases are derived from B-splines and their additional advantage in contrast to orthonormal wavelets is their smoothness and their explicit form. The disadvantage of popular bases from [4] is their bad condition which cause problems in practical applications. Some modifications which lead to better conditioned bases were proposed in [6, 1, 5]. In this contribution, we further improve the condition of spline-wavelet bases on the interval.

The primal scaling bases will be the same as bases designed in [2], because they are known to be well-conditioned. Let N be the desired order of polynomial exactness of the primal scaling basis and let $\mathbf{t}^j = (t_k^j)_{k=-N+1}^{2^j+N-1}$ be the Schoenberg sequence of knots. The corresponding B-splines of order N are defined by

$$B_{k,N}^j(x) := \left(t_{k+N}^j - t_k^j\right)\left[t_k^j, \ldots, t_{k+N}^j\right]_t (t-x)_+^{N-1}, \quad x \in [0,1], \tag{1}$$

where $(x)_+ := \max\{0, x\}$ and $[t_1, \ldots t_N]_t f$ is the N-th divided difference of f. The set Φ_j of primal scaling functions is then simply defined as

$$\phi_{j,k} = 2^{j/2} B_{k,N}^j, \quad \text{for} \quad k = -N+1, \ldots, 2^j - 1, \quad j \geq 0. \tag{2}$$

The desired property of the dual scaling basis $\tilde{\Phi}$ is biorthogonality to Φ and polynomial exactness of order $\tilde{N}$. Let $\tilde{\phi}$ be dual scaling function which was designed in [3] and which is shifted so that its support is $\left[-\tilde{N}+1, N+\tilde{N}-1\right]$. In this case $\tilde{N} \geq N$ and $\tilde{N}+N$ must be an even number. We define inner scaling functions as translations and dilations of $\tilde{\phi}$:

$$\theta_{j,k} = 2^{j/2}\tilde{\phi}\left(2^j \cdot - k\right), \quad k = \tilde{N}-1, \ldots 2^j - N - \tilde{N} + 1. \tag{3}$$

There will be additional two types of basis functions at each boundary. Basis functions of the first type are defined to preserve polynomial exactness in the same

With support of MŠMT LC06024.

way as in [4]:

$$\theta_{j,k} = 2^{j/2} \sum_{l=-N-\tilde{N}+2}^{\tilde{N}-2} \left\langle p^{\tilde{N}-1}_{k+N-1}, \phi\left(\cdot - l\right)\right\rangle \tilde{\phi}\left(2^j \cdot - l\right)|_{[0,1]}, \quad k = 1-N, \ldots, \tilde{N}-N.$$

Here $p^{\tilde{N}-1}_0, \ldots, p^{\tilde{N}-1}_{\tilde{N}-1}$ is a basis of $\mathbb{P}_{\tilde{N}-1}\left([0,1]\right)$. As in [4], $p^{\tilde{N}-1}_k$ are Bernstein polynomials defined by

$$p^{\tilde{N}-1}_k\left(x\right) := b^{-\tilde{N}+1}\binom{\tilde{N}-1}{k} x^k \left(b-x\right)^{\tilde{N}-1-k}, \quad k = 0, \ldots, \tilde{N}-1, \tag{4}$$

because they are known to be well-conditioned on $[0, b]$ relative to the supremum norm. In our numerical experiments the choice $b = 10$ seems to be optimal.

The basis functions of the second type are defined as

$$\theta_{j,k} = 2^{\frac{j+1}{2}} \sum_{l=\tilde{N}-1-2k}^{N+\tilde{N}-1} \tilde{h}_l \tilde{\phi}\left(2^{j+1} \cdot -2k - l\right)|_{[0,1]}, \quad k = \tilde{N}-N+1, \ldots, \tilde{N}-2, \tag{5}$$

where $\tilde{h}_l$ are scaling coefficients corresponding to $\tilde{\phi}$.

The boundary functions at the right boundary are defined to be symmetrical with the left boundary functions:

$$\theta_{j,k} = \theta_{j,2^j-N+1-k}\left(1-\cdot\right), \quad k = 2^j - N - \tilde{N} + 2, \ldots, 2^j - 1. \tag{6}$$

Since the set $\Theta_j := \left\{\theta_{j,k} : k = -N+1, \ldots, 2^j - 1\right\}$ is not biorthogonal to Φ_j, we derive a new set $\tilde{\Phi}_j$ from Θ_j by biorthogonalization. Finally the corresponding wavelets are determined by a method called stable completion.

References

[1] T. Barsch, Adaptive Multiskalenverfahren für Elliptische Partielle Differentialgleichungen - Realisierung, Umsetzung und Numerische Ergebnisse, *PhD. thesis, RWTH Aachen, Shaker Verlag*, 2001.

[2] C. K. Chui and E. Quak, Wavelets on a Bounded Interval, *Numerical Methods of Approximation Theory, edited by D. Braess and L. L. Schumaker, Birkhäuser*, 1992.

[3] A. Cohen and I. Daubechies and J. C. Feauveau, Biorthogonal Bases of Compactly Supported Wavelets, *Comm. Pure and Appl. Math.*, **45**, 1992.

[4] W. Dahmen and A. Kunoth and K. Urban, Biorthogonal Spline Wavelets on the Interval - Stability and Moment Conditions, *Appl. Comp. Harm. Anal.*, **6**, 1999.

[5] M. Primbs, New Stable Biorthogonal Spline-Wavelets on the Interval, *preprint Universität Duisburg-Essen*, 2007.

[6] S. G. Talocia and A. Tabacco, Wavelets on the Interval with Optimal Localization, *Math. Models Meth. Appl. Sci.*, **10**, 2000.

Dana Černá,
Technical university in Liberec,
Studentská 2, 461 17 Liberec

e-mail: dana.cerna@tul.cz

Václav Finěk,
Technical university in Liberec,
Studentská 2, 461 17 Liberec

e-mail: vaclav.finek@tul.cz

ESTIMATING EXTREMES IN CLIMATE CHANGE SIMULATIONS USING THE POT METHOD

JAN PICEK, JAN KYSELÝ

Keywords: peaks-over-threshold method, Poisson process, extreme temperatures
Particular aim: V03

The paper presents a methodology for estimating high quantiles of distributions of daily temperature in a non-stationary context, based on a peaks-over-threshold analysis with a time-dependent threshold expressed in terms of regression quantiles. The extreme value models are applied to estimate 20-yr return values of temperature over Europe in transient GCM simulations, assuming changes in greenhouse gas concentrations according to SRES emission scenarios over the 21st century. A comparison of scenarios of changes in the 20-yr return temperatures based on the non-stationary peaks-over-threshold models with conventional stationary models is performed. It is demonstrated that the application of the stationary extreme value models in temperature data from GCM scenarios yields results that may be to a large extent biased, while the non-stationary models lead to spatial patterns that are robust and enable one to detect those areas in which the projected warming in the tail of the distribution of daily temperatures is largest. The method also allows splitting the projected warming of the high quantiles in two parts that reflect a change in the location and scale of the distribution of extremes, respectively. Their spatial patterns differ in the examined climate change projections over Europe.

The regression quantiles (Koenker and Bassett 1978) are the most natural and intuitive solution to the problem of setting a (time-dependent) threshold in the POT analysis, corresponding to a high quantile of the distribution of the examined variable. The proposed non-stationary POT models with time-dependent thresholds and a homogeneous Poisson process are computationally straightforward and do not violate assumptions of the extreme value analysis, unlike models with an invariable threshold and a non-homogeneous Poisson process used in some previous climate change studies. Quadratic regression 95% quantiles were found superior to linear ones in the present application as they may accommodate to a wider range of changes in the threshold; on the other hand, differences between results obtained with the quadratic and linear quantiles were relatively small. Results are little dependent on the particular choice of the threshold (regression quantile); if the 96% or 97% quantiles are used instead of the 95% quantile, the main findings remain unchanged.

The arguments why an extreme value model with a time-dependent threshold is more appropriate compared to the widely used non-homogeneous Poisson process models with a fixed threshold in the context of climate change studies are straightforward. In fact, it is the sample size that is dealt with when modeling daily meteorological variables what makes the application of the non-homogeneous

Supported by MŠMT LC06024.

process questionable from the statistical point of view. When a significant trend is present in the data, a constant threshold in the POT models with a variable intensity of the Poisson process cannot be not suitable over longer periods of time: there are either too few (or no) exceedances above the threshold in an earlier part of record (which enhances the variance of the estimated model), or too many exceedances towards the end of the examined period (which violates asymptotic properties of the model and leads to bias), or both the deficiencies are present in the examined sample of 'extremes'. The issue of too many exceedances above the threshold (which cannot be considered extremes in fact) towards the end of the examined period would become less severe if the effective sample size from which extremes are drawn was larger than in 'real world' and/or climate change simulations of daily temperatures.

The choice of the particular statistical model for extremes as to the dependence of the model parameters on a time index is based on the likelihood ratio tests (e.g. Coles, 2001) for pairs of models with increasing complexity (i.e. the number of parameters describing the dependence on the covariate). We have chosen a threshold of at least 50% gridpoints in which an improvement is achieved in a given scenario in order to consider this model for the data. Although the choice of 50% is inevitably somewhat subjective, the results support that it is a reasonable one; the percentages are very low (below 1%) for models that are too complex for given data, and relatively large (around 70%) for the generally most parsimonious model. The differences between estimates of high quantiles based on two competing statistical models (in the present application, the model with a trend in logarithm of the scale parameter as the more complex one, and the model with constant scale and shape parameters as the less complex one) are small in those gridpoints in which the more complex model is not supported by the likelihood ratio test, which justifies the use of the statistical model based on the above criteria for the estimation in all gridpoints in the area. We also note that a POT methodology for modeling extremes should be preferred over block maxima due to the increase in the amount of data that enter the estimation (e.g. Coles, 2001). The advantage of the POT approach, consisting in a more efficient utilization of data on extremes, is even more pronounced for non-stationary extreme value models, owing to the increase in the number of parameters that have to be estimated.

References

[1] Coles, S., *An Introduction to Statistical Modeling of Extreme Values*, Springer Verlag, London, 2001.

[2] Koenker, R. and Bassett, G., Regression Quantiles. *Econometrica*, **Vol. 46**, 33-50, 1978.

Jan Picek,
Technická univerzita v Liberci,
Studentská 2, 461 17 Liberec

e-mail: jan.picek@tul.cz

PROCESSES OF FAILURES AND MODELS OF PARTIAL REPAIRS

PETR VOLF, JAROSLAV ŠEVČÍK

Key words: Hazard rate, reliability, repairable system, time to failure.
Particular aim: V03

This research activity deals with models for reliability and for analysis of failure times, their distributions and processes. As regards the impact of a repair actions on processes of failures, various concepts has already been introduced. We concentrated to several of them. Dorado et al. (1997) derived a model which allows simultaneously to shift the virtual age of the system after a repair and also to change the shape of the failure intensity in the forthcoming cycle, by accelerating (or retarding) the lifetime 'clock':

For any CDF F, $\theta \in (0,1]$ and $v \in [0,\infty)$, consider the family of distribution functions

$$\bar{F}_v^{\theta}(t) = \frac{\bar{F}(\theta t + v)}{\bar{F}(v)},\ t > 0. \tag{1}$$

The family of distributions $\left\{F_v^{\theta}\right\}$ are stochastically ordered in θ, that is, $\theta \le \theta'$ implies $F_v^{\theta}(t) \le F_v^{\theta'}(t)$, for all v and t. Then the survival function $\bar{F}_v^{\theta}(t)$ can be viewed as the life of a functioning item of age v which has been scaled by a factor θ, with lower values of θ representing longer remaining life. Authors mentioned above refer to $F_v^{\theta}(t)$ as the life distribution of an item with an *effective age* v and a *life supplement* θ.

Basic scheme of partial repairs. Consider two sequences $\{V_i\}_{i\ge 0}$ and $\{\Theta_i\}_{i\ge 0}$ called the effective ages and life supplements, respectively, satisfying

$$\begin{gathered} V_0 = 0,\ \Theta_0 = 1,\ V_i \ge 0,\ \Theta_i \in (0,1] \text{ and} \\ V_i \le V_{i-1} + \Theta_{i-1} T_i \text{ for } i > 0. \end{gathered} \tag{2}$$

The general model of virtual age defines the joint distributions of the inter-failure times T_i as follows

$$P(T_i \le t | V_{i-1},\ \Theta_{i-1},\ T_1, \dots, T_{i-1}) = F_{V_{i-1}}^{\Theta_{i-1}}(t) \tag{3}$$

for $t > 0,\ i \ge 1$, where $F_{V_{i-1}}^{\Theta_{i-1}}(t)$ is defined according to (1).

It is easy to see that T_j defined by distribution $F_{V_{i-1}}^{\Theta_{i-1}}(t)$ is stochastically larger than T_j defined by $F_{V_{i-1}}^{1}$, i.e. better than the working item of age V_{j-1}. Furthermore, we can see that for each $i \ge 1$ the effective age V_i of the system after the i-th repair is less than its effective age $X_i := V_{i-1} + \Theta_{i-1} T_i$ just before the i-th repair, which in turn is less than the actual age S_i. Thus the general repair model defined above can be considered as a better-than-minimal repair model.

Such a model includes several well known special cases, namely:

Supported by the project of MŠMT ČR No. LC06024.

1.: Perfect repair model. $\Theta_i = 1$, $V_i = 0$
2.: Minimal repair model. $\Theta_i = 1$, $V_i = S_i$, $S_i =$ time of failure
3.: Kijima's model I. $\Theta_i = 1$, $V_{i+1} = V_i + \xi_i T_i$, $V_i = \sum_{k=1}^{i} \xi_k T_k$, where $\{\xi_i\}_{i \geq 1}$ constants or (independent) random variables in $[0, 1]$.
4.: Kijima's model II. Similarly, $V_i = \xi_i(V_{i-1} + T_i)$ and $\xi_i \leq 1$, $\Theta_i = 1$, then $V_i = \sum_{k=1}^{i}(\prod_{l=k}^{i} \xi_l)T_k$.
5.: Accelerated lifetime model $\Theta_i > 1$, $V_i = 0$

Incomplete repair reducing the system deterioration. Let us consider a function $S(t)$ (or a latent random process) evaluating the level of degradation after a time t of system usage. In certain cases we can imagine $S(t) = \int_0^t s(u)du$ with $s(u) \geq 0$ is a stress at time u. We further assume that the failure occurs when $S(t)$ crosses a random level X. Recall also that (in the non-repaired system) the cumulated hazard rate $H_0(t)$ of random variable $T =$ time-to failure has the similar meaning, namely the failure occurs when $H_0(t)$ crosses a random level given by Exp(1) random variable,

Hence, as $T > t <=> X > S(t)$, i.e. $\bar{F}_0(t) = \bar{F}_X(S(t))$, where by $\bar{F}$ we denote the survival function, then

$$H_0(t) = -log\bar{F}_X(S(t)).$$

We can again have some special cases, for instance:

– $X \sim$ Exp(1), then $H_0(t) = S(t)$,

– $S(t) = c \cdot t^d$, $d \geq 0$, and X is Weibull (a, b), then T is also Weibull ($\alpha = ac^b, \beta = b \times d$), i.e. $H_0(t) = \alpha \cdot t^\beta$.

Let us now imagine that the repair reduces $S(t)$ as in the Kijima II model, to $S^*(t) = \delta \cdot S(t)$. In the Weibull case considered above we are able to connect such a change with the reduction of virtual time from t to some t^*:

$$S(t^*) = S^*(t) => t^* = \delta^{\frac{1}{d}} \cdot t,$$

so that the virtual time reduction follows the Kijima II model, too, with $\delta_t = \delta^{\frac{1}{d}}$. It can be shown that each selection of δ, Δ leads (converges) to a stable ('constant' intensity) case.

For other forms of function $S(t)$, e.g. if it is of exponential form, $S(t) \sim e^{ct} - 1$, such a tendency to a constant intensity does not hold in general. Nevertheless, it is possible to select convenient δ and Δ stabilizing the failure intensity.

Degradation as a random process. In the case we cannot observe the function $S(t)$ directly, and it is actually just a latent factor influencing the lifetime of the system, it can be modelled as a random process. There are several possibilities, for instance:

1. $S(t) = Y \cdot S_0(t)$, $Y > 0$ is a random variable, $S_0(t)$ a function as above.
2. Diffusion with trend function $S_0(t)$ and Brown process $B(t)$, $S(t) = S_0(t) + B(t)$.
3. $S(t)$ cumulating a random walk $s(t) \geq 0$.
4. Compound Poisson process (and its generalizations).

Again, it is assumed that failure occurs when the process $S(t)$ crosses a level x. Hence, $S(t) < x <=> t < T$, therefore $\bar{F}_0(t) = F_{S(t)}(x)$, where $F_{S(t)}(x)$ is the compound distribution at t. If X is a random level, then the right side has the form $\int_0^\infty F_{S(t)}(x)dF_X(x)$.

Random generation shows that the system behaves similarly as in the non-randomized case, and has the tendency to stabilize the intensity.

Cost optimization problem. The goal is to find optimal δ and Δ, for given other parameters, if costs of failure and preventive repair are given. In our examples, it was mostly possible to fix an optimal Δ to given δ, while optimal δ to selected Δ lied often close to complete or minimal repair degree.

Degradation process as a part of intensity model. When the degradation process is just one of factors influencing the survival of the system, it is quite natural to use it as a covariate in a regression model of failure intensity. In the situation when such a factor is not observed directly, it is more appropriate to use a model with latent component. In every case, it has a sense to consider the intensity of failure having several parts, one of them expressing the influence of the degradation process of the interest. Moreover, if the additive form of the intensity model is used (as in the case of Aalen regression model for intensity), the components stay separated even when integrated to cumulated intensity. Let us therefore recall several basic regression models for intensities of failures:

1. In the additive (also Aalen's) model, the total intensity is the sum of the intensities of components, e.g. $h(t) = h_1(t) + h_2(t)$.

2. In the multiplicative model $h(t) = h_1(t) \cdot h_2(t)$. The Cox's model uses the form $h(t) = h_0(t) \cdot \exp(B(z(t))$, where $z(t)$ is the regressor, i.e. in our case some characteristics of the deterioration.

3. Accelerated failure-time model $H(t) = H_0(V(t))$ was already briefly recalled here, too, in the connection with the model of growing virtual age.

The schemes of regression mentioned above offer different possibilities how to model the impact of degradation and then of repairs.

References

[1] C. Dorado, M. Hollander, and J. Sethuraman, "Nonparametric estimation for a general repair model," *The Annals of Statistics* 25, 1997, 1140-1160.

[2] M. Kijima, "Some results for repairable systems with general repair,"*J. Appl. Prob.* 26, 1989, 89-102.

Petr Volf[1], Jaroslav Ševčík[2],
[1]TU Liberec, [2]KPMS MFF UK Praha.

e-mail: petr.volf@tul.cz

ANALYSIS OF BREAKS IN NON-EQUALLY LOAD SHARING SYSTEM

PETR VOLF, ALEŠ LINKA AND MAROŠ TUNÁK

Key words: Breaking strength, load sharing model, latent variable, MCMC.
Particular aim: V03

The research deals with the reliability (survival, breaking strength) of a parallel, non-equally load-sharing system. The objective is to model and estimate the latent, unobserved variable characterizing this inequality, i. e. departures from equal load sharing. The inspiration is the analysis of breaking strength of a textile yarn, a bundle of fibers, where as a rule the fibers are warped so that the tension in them differs.

We shall consider 3 different cases, analyzing the critical extension before the break, critical breaking strength, and the relationship of both described by the stress-strain curve.

Critical extension. In this part we assume that the break of fiber is connected with its extension (measured e. g. in percents) and that from the experiments with individual fibers we know the distribution (density function, distribution function) of their critical extension $t_1 : f_1(t)$, $F_1(t)$, say. Now, when parallel system of M fibers is extended, we measure a variable t – extension of the bundle, but the actual extension of individual fibers is $T_j = t - Z_j$, Z_j is i.i.d. sample from a random variable $Z \geq 0$. We observe critical extension (when the 1-st break occurs) T, which in the equal load sharing case with $Z \equiv 0$ should equal $\min_j T_{1j}$ $(j = 1, \ldots, M)$, but in our case it is $\min_j(T_{1j} + Z_j)$, where T_{1j} are i.i.d. copies of T_1. Assume that on $i = 1, \ldots, n$ bundles we measure the extensions to the first break, $T(i)$, we know the distribution of variable T_1, and we should find an "optimal" (best-fitting) distribution of Z. Denote it f_Z, F_Z, resp., further denote by f_Y, F_Y, density and distribution function of variable $Y = T_1 + Z$, by f_T, F_T the same for variable T, by $\overline{F} = 1 - F$ the survival functions.

Then

$$\overline{F}_Y(y) = \int_0^\infty f_1(t)\,\overline{F}_Z(y-t)\,\mathrm{d}t = \int_0^\infty f_Z(z)\cdot\overline{F}_1(y-z)\,\mathrm{d}z, \tag{1}$$

where $\overline{F}$ of negative argument is taken as 1. Hence, for instance the first term can also be written as $\int_0^y f_1(t)\,\overline{F}_Z(y-t)\,\mathrm{d}t + \overline{F}_1(y)$.

Corresponding density is then

$$f_Y(y) = \int_0^y f_1(t)\,f_Z(y-t)\,\mathrm{d}t = \int_0^y f_Z(z)\,f_1(y-z)\,\mathrm{d}z, \tag{2}$$

and for the density of the minimum, T, we obtain

$$f_T(t) = M\overline{F}_Y(t)^{M-1}f_Y(t). \tag{3}$$

Supported by the project of MŠMT ČR No. LC06024.

Hence, we can construct the likelihood $\mathcal{L} = \prod_{i=1}^{n} f_T(T(i))$, depending on known distribution of T_1 and unknown distribution of Z. In the Bayes setting, we can consider also a convenient prior.

Essentially, three cases can be distinguished. First, let us assume that the type of distribution of Z is known (e. g. exponential), its parameter $\boldsymbol{\theta}$ unknown. We then deal with standard optimization problem, maximizing either $\mathcal{L}(\boldsymbol{\theta}; \{T_i\})$ for MLE or $\mathcal{L}(\boldsymbol{\theta}; \{T_i\}) \cdot g(\boldsymbol{\theta})$ for Bayes MAPE with prior $g(\boldsymbol{\theta})$.

In the case of unknown distribution type for Z the simplest way is to consider a certain form of nonparametric curve as unknown density. A frequent choice is the mixture of Gauss distributions, for the distribution, or a linear combination of Gauss curves for a curve modeling, with different means μ_k and equal variance σ^2, either cut off at zero (because $Z \geq 0$) or used for the distribution of $\log Z$. We used a mixture of gamma distributions instead, for modeling the distribution of Z.

The parameters of used units, i.e. of normal or gamma densities, and also the number K of them should be optimized The case can be solved either by repeating optimization for different K, then selection of optimal penalized one, or by randomized search allowing for adding and deleting of units – this is one of typical application of MH algorithm with changes of dimension (RJMH).

The third approach can try to generate a representation of variable Z directly (i.e. the data augmentation approach). In such a case we repeat the following steps:

i) Select randomly an index of bundle, i, and generate (from a prior, e.g. rhat from method 2) some new values Z_{ji}.

ii) Compute likelihood (just for i-th bundle: The survival function of the extension at the first break is $\overline{F}_T(t) = \prod_{j=1}^{M} \overline{F}_1(t + Z_j)$, hence the density has the form $f_T(t) = \sum_{k=1}^{M} \prod_{j \neq k} \overline{F}_1(t + Z_j) \cdot f_1(t + Z_k)$.

iii) Those densities, with old and new values of Z_j, are now compared and randomly accepted or not - this is a regular MH step, as the proposals were generated from prior distribution. Naturally, sequential innovation of priors parameters (in the case of other than uniform prior) can be considered, too.

We can also imagine that the distribution of T_1 is given just empirically, by observed values. Then, in (1) and (2) the first formulas, actually representing the expectations w.r. to T_1, will use arithmetic means instead.

Optimization with constraint. The situation is, as a rule, more complicated, because the extension is measured from the moment when at least one fiber is already taut, i. e. $\min Z_j = 0$. In that case we try to identify the distribution of the remaining variables $Z_j^* = Z_j - \min Z_j$, assuming that there still remain $M - 1$ random variables $Z_j^* > 0$, which represent certain distribution above minimum of $\{Z_j,\, j = 1, \ldots, M\}$.

Hence, the only change is that in (3) we obtain

$$f_T(t) = f_1(t)\, \overline{F}_Y(t)^{M-1} + (M-1)\, f_Y(t)\, \overline{F}_Y(t)^{M-2} \cdot \overline{F}_1(t).$$

Critical strength. In other instances we measure directly the breaking strength or, equivalently, the resistance of a fiber, the stress in it, Again, let us assume that the distribution of random variable S_1 – the strengths breaking individual fibers, is known. During the experiment with the bundle, we measure the sum of strengths influencing the components. Let S be critical level, the overall strength when the first break occurs. This $S = \sum_{j=1}^{M} S_j$, where S_j are strengths per component. In the equal load sharing case $S_j = S/M$, here we assume that $S_j = S/M + V_j$. It follows that $\sum_{j=1}^{M} V_j = 0$. Similarly as in the preceding part, we, at the moment of the first break, measure S, and it holds that $\overline{S} = S/M = \min_j(S_{1j} - V_j)$, with constraint $\sum_{j=1}^{M} V_j = 0$. Let us assume certain prior distribution for r.v. V (with $EV = 0$). Let us take for instance a "constrained" distribution derived from the Gauss one $\mathcal{N}(0, \sigma^2)$ in such a way that if we wish to generate one representation of variables V_j, we sample $V_1 \cdots V_{M-1}$ and shift them by their average. Then, for given values of V_j, the distribution of $S_{1j} - V_j$ is now just the shifted distribution of S_1. The distribution of $\overline{S}$ follows. Thus, for given $\{V_{ji}\}$ and given distribution of S_1 we are able to derive the probability of observed variables $\overline{S}_i$, $i = 1, \dots, n$. The procedure sequentially generates and accepts or rejects the values V_{ji}, together with updating parameter σ^2 of their prior. Finally, we obtain a sample representing distribution V.

Relationship between strength and extension. From experiments with breaks of ductile fibers the stress–strain curves are available. They show the development of stress measured in the fiber and its dependence on the extension. In the region of the fiber break, the curve is, as a rule, approximately linear, $s_1 = \alpha + \beta t_1 + \varepsilon$.

When the bundle of fibers is examined, we measure the extension of the bundle, t, which actually is $t_{1j} + Z_j$ for j-th fiber, and the total stress $s = \sum s_1 j$. Hence, we have that

$$s = \sum_j s_{1j} = M\alpha + M\beta t - \beta \sum_j Z_j + \sum_j \varepsilon_j$$

from which $\alpha + \beta t = \overline{s} + \beta\overline{z} - \overline{\varepsilon}$, where $\overline{s}$, $\overline{z}$, $\overline{\varepsilon}$ denote arithmetic means. It follows that the stress per j-th component in the bundle can be expressed as

$$\begin{aligned} s_{1j} &= \overline{s} + \beta\overline{z} - \overline{\varepsilon} - \beta z_j + \varepsilon_j = && (4)\\ &= \overline{s} + (\beta(\overline{z} - z_j) + \varepsilon_j - \overline{\varepsilon}) = \overline{s} + V_j^*. && (5) \end{aligned}$$

The first break occurs when the left side equals the right side first time, i. e. for

$$\overline{S} = \min_j \left(S_{1j} - V_j^*\right). \qquad (6)$$

Variables V_j^* represents the variables V_j from part 2, and really $\sum V_j^* = 0$. It is assumed that parameters α, β as well as distributions of S_1 and ε are known from the analysis of individual fibers, while distribution of Z is identified by the methods described in the part 1.

While in the case 1 the critical extension for the bundle was (stochastically) larger than simple $\min_j (T_{1j})$, in the case 2, quite logically, the critical strength breaking the bundle is stochastically smaller than $\min_j (S_{1j})$. Except the analysis, the method could serve for the prediction of breaking strength, on the basis of individual stress – strain curves.

References

[1] Belyaev Yu. K. and Rydén P., *Non-parametric estimators of the distribution of tensile strengths for wires.* Research report, University of Umea, 1997.

[2] Volf P., Linka A., On Reliability of System Composed of Parallel units subject to increasing load. *International Journal of Reliability, Quality and Safety Engineering* 7, 2000, 271-284.

Petr Volf[1], Aleš Linka[2] and Maroš Tunák[2],
Technical University of Liberec, Czech Republic,
[1]Dept. of Applied Mathematics, [2]Dept. of Textile Materials.

e-mail: petr.volf@tul.cz

TESTS FOR EXTREME VALUE DOMAIN OF ATTRACTION IN REGRESSION MODEL

JAN PICEK, JAN DIENSTBIER

Keywords: domain of attraction, regression quantile
Particular aim: VO1

Neves, Picek and Alves in 2006 constructed test for the null hypothesis that the distribution comes from the Gumbel domain of attraction in location model. Let $X_{n-k:n} \le X_{n-k+1:n} \le ... \le X_{n:n}$ is the selection of the largest observations from a random sample. Then they suggested the following test statistic

$$T_{k,n} := \frac{X_{n:n} - X_{n-k:n}}{\frac{1}{k}\sum_{i=1}^{k}\left(X_{n-i+1:n} - X_{n-k:n}\right)} \tag{1}$$

where an intermediate sequence $k = k_n$ is thesequence of positive integers, $k = k_n \to \infty$ as $k_n/n \to 0$, as the sample size n tends to infinity.

In this contribution we try to generalize that setup in regression context.

We consider the linear regression model

$$\mathbf{Y}_n = \mathbf{X}_n\boldsymbol{\beta} + \mathbf{E}, \tag{2}$$

where $\mathbf{Y}_n = (Y_1, \ldots, Y_n)^\top$ is a vector of observations, $\mathbf{X}_n$ is an $(n\times p)$ known design matrix, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^\top \in R^p$ is an unknown parameter and $\mathbf{E}_n = (E_1, \ldots, E_n)^\top$ is the vector of *i.i.d.* errors with (generally unknown) distribution function F with unknown shape, location and scale parameters, belonging to some max-domain of attraction. We assume that β_1 is an intercept, that is, the first column of $\mathbf{X}_n$ is $(1, \ldots, 1)^\top$.

Koenker and Basset (1978) defined the α-regression quantile $\boldsymbol{\beta}(\alpha)$ $(0 < \alpha < 1)$ for the model (2) as any solution of the minimization

$$\sum_{i=1}^{n} \rho_\alpha(Y_i - \mathbf{x}_i'\mathbf{t}) := \min, \quad \mathbf{t} \in R^p, \tag{3}$$

where

$$\rho_\alpha(x) = x\psi_\alpha(x),\ x \in R^1 \text{ and } \psi_\alpha(x) = \alpha - I_{[x<0]},\ x \in R^1. \tag{4}$$

Let $Y_1, \ldots, Y_n$ are observations obtained from the linear model (2). Define a subsample of exceedances over some "high" regression quantile threshold

$$Z_i := \left(Y_i - \mathbf{X}_{i\bullet}\hat{\boldsymbol{\beta}}(\alpha_k|\mathbf{X}, \mathbf{Y})\right)^+, \tag{5}$$

i.e. for some α_{k_n} and it's associated intermediate order sequence of k_n such that $k_n/n \to 0$ as $n \to \infty$ and where $\mathbf{X}_{i\bullet}$ denotes i-th row of matrix $\mathbf{X}_n$.

Supported by MŠMT LC06024.

Finally, the test statistics is

$$T^*_{k,n} := \max\left\{ \frac{Z_{n-j+1:n} - X_{n-l:n}}{\frac{1}{l}\sum_{i=1}^{l}(Z_{n-i+1:n} - Z_{n-l:n})} : j = 1, \cdots, l(k_n, \mathbf{X}, \mathbf{Y}) \right\}. \tag{6}$$

Since the empirical regression quantiles form for $\alpha \in [0,1]$ a step function in R^d space (or similarly the estimates of intercept form a step function in R), it is advisable to use only unique solutions of regression quantile process $(\hat{\boldsymbol{\beta}}(\alpha))_{\alpha\in[0,1]}$. We obtain a sequence of high regression quantiles for $\alpha_1, , \alpha_2$ using described procedure. Then we follow the univariate i.i.d case.

References

[1] Koenker R. and Bassett G., Regression Quantiles. *Econometrica*, **Vol. 46**, 33-50, 1978.

[2] Neves C., Picek J., Alves F.M.I. The contribution of the maximum to the sum of excesses for testing max-domains of attractions. *J. Statist. Planning Infer.* **Vol. 136** (4), 1281-1301, 2006.

Jan Picek,
Technická univerzita v Liberci,
Studentská 2, 461 17 Liberec

e-mail: jan.picek@tul.cz

ODHAD PARETOVA INDEXU ZALOŽENÝ NA NEPARAMETRICKÉM TESTU

JANA JUREČKOVÁ, MAREK OMELKA

Klíčová slova: Paretův index; silná konzistence; asymptotická normalita.
Dílčí cíl: V01

Nechť $X_1, \dots, X_R$ jsou nezávislá pozorování se společnou distribuční funkcí F, která má nedegenerovaný pravý chvost. Rozlišujme dvě rozsáhlé třídy rozdělení pravděpodobností odpovídající dvěma typům chvostů F:

- *Exponenciální chvosty* F (1. typ): pro F platí konvergence
$$\lim_{a\to\infty} \frac{-\log(1-F(a))}{b\,a^r} = 1$$
pro nějaká $b > 0$ a $r \geq 1$.
- *Těžké chvosty* F (2. typ): pro F platí konvergence
$$\lim_{a\to\infty} \frac{-\log(1-F(a))}{m\log a} = 1$$
pro nějaké $m > 0$.

S rozsáhlými bloky dat, která se dají modelovat rozdělením pravděpodobností 2. typu, se setkáme např. v pojišťovnictví, ve financích, výpočetní technice a telekomunikacích. Proto se mnozí statistikové zabývali možnými odhady a jinými procedurami o parametru m, zvaném *Paretův index nebo index chvostů.* Nejznámější je Hillův odhad (Hill (1975)), ale byla navržena řada dalších odhadů. Překvapující je, že až donedávna prakticky neexistovaly testy hypotézy typu $\mathbf{H}:\ m \leq m_0$, že naše rozdělení je těžší než Paretovo rozdělení s indexem m_0. Některé testy tohoto typu byly navrženy poměrně nedávno (Jurečková and Picek (2001), Jurečková (2003), Fialová, Jurečková and Picek (2004), Jurečková, Koul and Picek (2008)). Tyto neparametrické testy jsou nestranné a dobře rozlišují chvosty distribučních funkcí. Později Jurečková a Picek (2004) zkonstruovali třídu odhadů Paretova indexu inverzí těchto testů způsobem, který užili Hodges a Lehmann (1963). Asymptotické vlastnosti těchto odhadů plynou z asymptotické síly příslušných testů.

Tyto odhady, stejně jako testy, vycházejí z rozkladu množiny pozorování na N malých výběrů velikostí n a na empirických distribučních funkcí $\widehat{F}_N$ vhodných N charakteristik těchto malých výběrů; výše zmíněné články uvažují podvýběrová maxima, podvýběrové průměry a průměry dvou podvýběrových extrémů. Jurečková and Picek (2004) dokázali silnou konzistenci těchto odhadů pro některé volby mezí v příslušném testu. Tuto silnou konzistenci nyní dokazujeme za obecnějších podmínek; dále za mírných podmínek na model dokazujeme asymptotickou normalitu odhadů. Asymptotika se uvažuje pro $N \to \infty$ a pevné (i malé) n. Soustředíme se na odhad založený na N podvýběrových maximech, ale stejným způsobem se mohou uvažovat odhady založené na dalších podvýběrových statistikách.

S podporou M©MT LC06024.

Kromě silné konzistence odhadu a jeho asymptotické normality studujeme jeho vychýlení při konečném N. Dále uvažujeme možná zlepšení vlastností odhadu při konečném N, např. dosažení jeho invariance vzhledem k měřítku pomocí vhodné studentizace, a možnou redukci jeho střední kvadratické odchylky. Účinnost těchto modifikací ilustrujeme na reálných datech.

Reference

[1] A. Fialová, J. Jurečková and J. Picek, Estimating Pareto tail index based on sample means, *REVSTAT* 2/1, 75–100, 2004.

[2] B. M. Hill, A simple general approach to inference about the tail of a distribution *Ann. Statist.*, 3:1163–117, 1975.

[3] J. Jurečková and M. Omelka, Estimator of the Pareto index based on nonparametric test, *Communications in Statistics A*, 2008 (in print).

[4] J. Jurečková, H. L. Koul and J. Picek, Testing the tail index in autoregressive models, *Annals of the Institute of Statistical Mathematics*, DOI 10.1007/s10463-007-0155-z, 2008.

[5] J. Jurečková and J. Picek, A class of tests on the tail index, *Extremes*, 4:165–183, 2001.

[6] J. Jurečková and J. Picek, Estimates of the tail index based on nonparametric tests, In *Statistics for Industry and Technology*, Birkhäuser Verlag Basel, 2004.

Jana Jurečková, Marek Omelka
Univerzita Karlova, Centrum Jaroslava Hájka
Sokolovská 83, 186 75 Praha 8

e-mail: jurecko@karlin.mff.cuni.cz, omelka@karlin.mff.cuni.cz

DVOUVÝBĚROVÉ NESTRANNÉ MNOHOROZMĚRNÉ POŘADOVÉ TESTY

JANA JUREČKOVÁ, JAN KALINA A MAREK OMELKA

Klíčová slova: Kolmogorov-Smirnovův test; Kontiguita; Lehmannovy alternativy; Liu-Singhův test; Nestrannost; Psi-test; Savageův test; Wilcoxonův test.

Dílčí cíl: V01

Nestrannost testů o vektorových parametrech nebo o skalárních parametrech proti oboustranným alternativám při konečných počtech pozorování je otevřená otázka, ne často diskutovaná v literatuře.

Absence nestrannosti je ještě vážnějším problémem při testování v mnohorozměrných modelech, kde musíme uvažovat, proti kterým alternativám je test nestranný, ale které lze zároveň přirozeně interpretovat a jsou ve shodě s experimentem.

Zde navrhujeme mnohorozměrné dvouvýběrové testy, založené na pořadích vzdáleností mezi mnohorozměrnými daty, s důrazem na jejich nestrannost a další vlastnosti při konečných počtech pozorování. Jmenovitě uvažujeme mnohorozměrné verze Wilcoxonova testu, Psi-testu a Savageova testu, které jsou nejen konzistentní, ale také nestranné a nezávislé na rozdělení pravděpodobností (distribution free) při konečných výběrech, za hypotézy a za rozsáhlé třídy alternativ typu poloha/měřítko. Každý z těchto testů je lokálně nejsilnější proti specifickým alternativám Lehmannova typu. Tyto testy srovnáváme s testem Kolmogorov-Smirnovova typu založeného na vzdálenostech dat a s testem Wilcoxonova typu založeným na pořadích hloubek (depths), který navrhli Liu a Singh (1993). Numericky srovnáváme síly testů, včetně Hotellingova, proti některým alternativám.

Reference

[1] Gibbons, J. D., A proposed two-sample test and its properties, *J. R. Statist. Soc. B* 26, 305–312, 1964.

[2] Lehmann, E.L., The power of rank tests, *Ann. Math. Statist.* 24, 23–42, 1953.

[3] Liu, R. and Singh, K. A quality index based on data depth and multivariate rank tests, *J. Amer. Statist. Assoc.* 88, 252–260, 1993.

[4] Savage, I. R., Contributions to the theory of rank order statistics - The two sample case, *Ann. Math. Statist.* 27, 590–615, 1956.

[5] Zuo, Y. and He, X., On the limiting distributions of multivariate depth-based rank sum statistics and related tests, *Ann. Statist.* 34, 2879–2896, 2006.

Jana Jurečková, Jan Kalina, Marek Omelka
Univerzita Karlova, Centrum Jaroslava Hájka
Sokolovská 83, 186 75 Praha 8

e-mail: jurecko@karlin.mff.cuni.cz, kalina@karlin.mff.cuni.cz,omelka@karlin.mff.cuni.cz

S podporou M©MT LC06024.

TESTING THE STABILITY OF THE FUNCTIONAL AUTOREGRESSIVE PROCESS

MARIE HUŠKOVÁ

Particular aim: V03

The talk concerns test procedures for detection of a changes in a functional autoregressive process. Particularly, we consider the functional time series $X_{i+1} = \Psi_i X_i + \varepsilon_i$, $i = 1, \ldots, n$, where X_i are observations, ε_i are i.i.d. errors and Ψ_i are operators. We test develop a test procedure for testing no change in Ψ_i versus Ψ_i is changing at an unknown time point. The developed CUSUM type procedure uses functional principal component analysis it is constructed to have a well-known asymptotic distribution, but asymptotic justification is very delicate. Theoretical properties are investigated.

Finite sample performance is examined by an application to a data set. Particularly, we work with a data set studied in Laukaitis and Rackauskas (2002) that consist of detailed records of transactions made with credit cards issued by Vilnius Bank, Lithuania. The studied functional time series is the count of transactions in a one minute interval starting at minute t on day $n = 1, 2, ...200$.

Joint paper with Lájos Horváth Piotr Kokoszka.

Marie Hušková
Charles University in Prague
Sokolovsk 83, 186 75 Praha 8

e-mail: huskova@karlin.mff.cuni.cz

Supported by MŠMT LC06024.

ESTIMATION OF THE FRACTION OF FALSE HYPOTHESES

LEV B. KLEBANOV, B. SHOKIROV

Keywords: hypothesis testing, false hypothesis, p-value, empirical distribution function, consistent estimator, unstable estimator
Particular aim: V06

1. The problem

Let us have samples $X_1^{(j)}, \dots, X_n^{(j)}$ from distribution function (df) F_j and $Y_1^{(j)}, \dots, Y_n^{(j)}$ from df G_j, $j = 1, \dots, m$. Testing hypotheses $H_0^{(j)} : F_j = G_j$, $j = 1, \dots, m$ we would like to estimate the number k of false hypotheses. Since the total number of hypotheses is equal to m, then the ratio k/m will be equal to some constant π. Then the problem will be reduced to estmating π. If by using some test we obtain p-values, then $p \approx \pi F + (1-\pi)G$.

Let us p-values have df $G(x)$ under the null hypotheses and $F(x)$ under the alternative and let $F(x) \geq G(x)$ $\forall x \in [0,1]$. If the share of true hypotheses is $1-p$ then we can assume that the observed p-values is distributed as $H_p(x) = pF(x) + (1-p)G(x), \quad x \in [0,1], \quad (p \in [0,1])$. Here $H_p(x)$ is a observable df and $G(x)$ is known. We would like to estimate p by the sample $X_1, \dots, X_n$. Obviously, on such setting the problem has no solution. Therefore, we do some suggestions under which the solution whould be possible. Namely, we suppose that

$$(A1) \quad F(x) \geq G(x), \forall x \in [0,1]$$

and

$$(A2) \quad suppF \subset [0, 1-\delta], \quad \text{for some} \quad \delta > 0.$$

2. Different approaches to estimate p-values

Let $(A1)$ and $(A2)$ are satisfied.

I. For all $0 \leq x \leq 1$ we have

$$\sup_x \left| \frac{H_p(x) - F(x)}{1 - G(x)} \right| = p. \tag{1}$$

If $H_{p,m}(x)$ is an empirical df (edf) then we can define

$$p_m^* = \sup_x \left| \frac{H_{p,m}(x) - F(x)}{1 - G(x)} \right|. \tag{2}$$

p_m^* is a consistent estimator of p, however, is not stable.

II. Another similar approach is as following ([2,3]):

$$H_p'(x) = pF'(x) + (1-p)G'(x).$$

By virtue of (A2) $F'(x) = 0$ for $x \in [1-\delta, 1]$, so $H_p'(x) = (1-p)G'(x)$ for $x \in [1-\delta, 1]$. We have to estimate H_p'. For this goal we take $H_{p,m}$ (edf) and smooth

Date: 25-27.09.2008.
S podporou MŠMT LC06024.

it (by convolution or by another process). It is clear that $H_{p,m}(x) = pF_m(x) + (1-p)G_m(x)$. For $x \in [1, 1-\delta]$, $F_m(x) = 1$, but the smooth approximation of F_m may not be equal to 1 for $x \in [1, 1-\delta]$. Usually the smoothing process causes some approximation errors such that the derivative of smoothed F_m may not be equal to zero in the interval $[1, 1-\delta]$. Thus, methods I and II possesses some degree of unstability, which makes their application quite limited. It is clear that ([1]) dependency between tests increases unstability of the estimators.

III. Now consider

$$H_p(x) - G(x) = p(F(x) - G(x)).$$

The quantity

$$\max_x |H_p(x) - G(x)| \tag{3}$$

represents the low bound for the share of false hypotheses and to show this we do not use neither $(A1)$ nor $(A2)$. Of course the validity of $(A1)$ guarantees that

$$\max_x (H_p(x) - G(x)) = \max_x |H_p(x) - G(x)|,$$

but this is not so important. However, the property $(A2)$ now plays important role. Namely, $\max_x |F(x) - G(x)| \geq 1 - G(1-\delta)$, therefore

$$p(1 - G(1-\delta)) \leq \max_x |H_p(x) - G(x)| \leq p. \tag{4}$$

3. The procedure of test performance

Let we have slides with microarray data. Devide slides into K groups. For every group conduct the testing procedure and calculate p-values. Next, for the gene number 1 (i.e., for the first hypothesis) choose randomly (from disctere uniform distribution $\{1, \ldots, K\}$) the group number and prescibe to the first hypothesis (gene number 1) the p-value calculated for this group. For the gene number 2 repeat the same procedure but with new random choice of the group of slides. This procedure will be repeated for all genes. Next, apply (4) to all p-values obtained this way, more exactly, we take

$$\max_x |H_{p,m}(x) - G(x)|$$

as a estimator of p. If we have n slides, then the number of groups $K = [\sqrt{n}]$ and the number of slides in each of group equals to $[\frac{n}{K}]$.

References

[1] Klebanov, L. B., Charactreization of distributions symmetric with respect to a group of transformations and testing of corresponding statistical hypothesis, *Statistics and Probability Letters*, **Vol. 31**, 2000.

[2] Storey, J.D., Estimating false Discovery Rate Under Dependence, with Applications to DNA Microarrays, *Annals of Statistics*, **Vol. 31**, 2003.

[3] Storey, J.D., Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach, *J. R. Statist. Soc.*, **Vol. 66**, 2004.

Bobosharif Shokirov,
Charles University in Prague,
Sokolovska 83, Prague 8

e-mail: bobosari@karlin.mff.cuni.cz

ROBUSTNÍ REGRESE A EKONOMETRICKÉ APLIKACE

JAN KALINA

Klíčová slova: heteroskedasticita, instrumentální proměnné.
Dílčí cíl: V02

Práce v oboru robustní regrese a jejích aplikacích patří do dílčího cíle V02. Pro robustní regresní metodu Least Weighted Squares (nejmenší vážené čtverce, LWS) jsme studovali výpočetní aspekty a rychlost výpočtu pro velké datové soubory.

Dále byly odvozeny asymptotické diagnostické nástroje pro metodu nejmenších vážených čtverců, mezi něž patří Durbinův-Watsonův test nezávislosti náhodných chyb v lineárním regresním modelu a také některé testy heteroskedasticity. Tyto jsou odvozeny z asymptotické reprezentace pro odhad metodou LWS. Ukazuje se, že tyto diagnostické nástroje ve své klasické verzi pro metodu nejmenších čtverců jsou asymptoticky platné a použitelné také pro rezidua odhadu metodou nejmenších vážených čtverců.

Další aplikací jsou potom robustifikace dvou metod pro statistické odhadování parametrů v lineárním regresním modelu, které jsou ve své klasické formě populární v ekonometrii. Jednou z nich je metoda instrumentálních proměnných pro takovou situaci, kdy náhodné chyby sice nejsou nekorelované s regresory, ale jsou k dispozici další instrumenty, které s nimi jsou nekorelované a zároveň jsou dostatečně korelované s regresory. Druhou aplikací je taková modifikace robustní regrese, která zajistí eficienci při heteroskedasticitě a umožňuje také odhadovat rozptyl odhadu regresních parametrů. V obou případech jsme studovali robustní přístupy založené na myšlence přiřadit menší váhy méně spolehlivým pozorováním s malými hodnotami reziduí.

Reference

[1] Kalina J., Computing robust GMM estimators, Submitted to *Computational Statistics and Data Analysis*, 2008.

[2] Kalina J., Diagnostics for instrumental weighted variables, *Preprint KPMS MFF UK 61/2008.*

Jan Kalina,
MFF UK,
KPMS MFF UK, Sokolovská 83, 186 75 Praha 8

e-mail: kalina@karlin.mff.cuni.cz

S podporou MŠMT LC06024.

APLIKACE KLASIFIKAČNÍ ANALÝZY

JAN KALINA, MAREK OMELKA, BOBOSHARIF SHOKIROV

Klíčová slova: klasifikační analýza, výpočetní aspekty.
Dílčí cíl: V01

Pražská pobočka Centra Jaroslava Hájka pro teoretickou a aplikovanou statistiku se podílí také na činnosti v oblasti aplikované statistiky.

Ve spolupráci s dr. Danielem Svozilem a dr. Bohdanem Schneiderem z Ústavu organické chemie a biochemie AV ČR vznikl článek [1] o konformacích DNA a klasifikaci jednotlivých vzorků DNA na základě četností jednotlivých dinukleotidů ve vlákně DNA. V současné době spolupracujeme na klasifikační analýze vzorků DNA opět do jednotlivých konformací, tentokrát na základě torzních úhlů měřených mezi atomy v molekule DNA. Zde se jedná o klasifikaci cirkulárních dat, protože klasifikační pravidlo je založeno na hodnotách úhlů v intervalu od 0 do 360 stupňů.

V rámci činnosti Centra Jaroslava Hájka dokončil J. Kalina dizertační práci [2], která se zabývá automatickým hledáním objektů v obrazech obličejů za pomoci šablon. V práci je navržena a implementována metoda pro optimalizaci šablony a také optimalizaci vah pro vážený korelační koeficient. Smyslem optimalizace je zvýšit diskriminaci mezi částmi obrazu příslušejícími šabloně a ostatními částmi obrazu. Optimalizace bez dodatečných omezení ovšem má tendenci degenerovat, zatímco dodatečné omezení vlivu jednotlivých pixelů vede k získání robustního řešení. Výsledky jsou srovnány také pro různé počáteční velikosti či rotace obličeje. Navržená metoda přitom neužívá speciální vlastnosti obličejů a lze ji popsat jako obecný přístup k robustní neparametrické diskriminaci pomocí šablon.

Reference

[1] Svozil D., Kalina J., Omelka M., Schneider B., DNA conformational space, *Nucleic Acids Research* **36**, No.11, pp. 3690 – 3706, 2008.

[2] Kalina J., *Locating landmarks using templates*, dizertační práce, Universität Duisburg-Essen, Duisburg, 2007.

Jan Kalina,
MFF UK,
KPMS MFF UK, Sokolovská 83, 186 75 Praha 8

e-mail: kalina@karlin.mff.cuni.cz

S podporou MŠMT LC06024.