INFORMAČNÍ BULLETIN

České statistické společnosti

Ročník 26, číslo 1–2, červen, 2015

WRONG-WAY RIZIKO – KALIBRACE KORELAČNÍHO KOEFICIENTU

WRONG-WAY RISK – CORRELATION COEFFICIENT CALIBRATION

Jakub Černý[1], **Jiří Witzany**[2]

Address: [1]Faculty of Mathematics and Physics, Charles University in Prague, Sokolovska 83, 186 75, Prague 8, Czech Republic,
[2]Faculty of Finance and Accounting, University of Economics, Prague, W. Churchill Sq. 4, 130 67, Prague 3, Czech Republic

E-mail: [1]jcerny@karlin.mff.cuni.cz, [2]jiri.witzany@vse.cz

Abstrakt: Dle nové bankovní regulace Basel III by banky měly zahrnout do výpočtu kreditní přirážky k tržnímu ocenění (CVA) mimoburzovních derivátů také tzv. wrong-way riziko (WWR). WWR představuje nepříznivou závislost mezi expozicí a časem selhání (defaultu) protistrany. Za předpokladu, že propojení mezi sazbou úrokového swapu (IRS), tj. finančního derivátu, spočívajícího v opakující se výměně dvou plateb na základě hodnoty úrokové sazby, a časem selhání je reprezentováno gaussovskou kopulou s konstantním korelačním koeficientem, lze toto riziko vyjádřit právě pomocí tohoto korelačního koeficientu. Vzhledem k tomu, že pozorování času selhání znamená zánik společnosti, nelze tuto korelaci jednoduše odhadnout pomocí pozorovaných dat na rozdíl od intenzity selhání, která přímo odpovídá sazbě swapu úvěrového selhání (CDS). Na základě dostupných denních IRS a CDS sazeb České republiky jsme korelaci odhadli metodou maximální věrohodnosti za předpokladu, že systematický faktor se řídí AR(1) procesem, abychom mohli dekorelovat (eliminovat autokorelaci) obě časové řady. Výsledky ukazují, že korelace kalibrovaná na denní data je poměrně vysoká, a proto by WWR nemělo být v tomto případě zanedbáváno.

Klíčová slova: Wrong-way riziko, WWR, kalibrace korelačního koeficientu, metoda maximální věrohodnosti, CVA.

Abstract: Under the new Basel III banking regulation banks should include wrong-way risk (WWR) into the calculation of the credit valuation adjustment (CVA) of the OTC derivatives. WWR takes place when the exposure to a counterparty is adversely correlated with the credit quality of that counterparty. Assuming a link between the interest rate swap (IRS), i.e. financial derivative in which two counterparties repeatedly exchange cash flows based on interest rate value and that the default time is represented by a Gaussian copula with a constant correlation coefficient, the WWR is expressed by

this correlation coefficient. Because the observation of the default time means bankruptcy of the company, the correlation cannot be simply estimated using the observed data in contrast to the credit default swap (CDS) rate which is related to the intensity of default. Based on available daily Czech Republic government IRS and CDS rates we estimated the correlation using maximum likelihood method assuming that the systematic factor is governed by the AR(1) process, so we can decorrelate (eliminate autocorrelation) both time series. The results show that the correlation calibrated on the daily data is relatively high, and therefore the WWR should not be neglected in this case.

Keywords: Wrong-way risk, WWR, correlation coefficient calibration, maximum likelihood estimation, CVA.

1. Introduction

The new Basel III banking regulation requires banks to calculate credit valuation adjustment (CVA) of OTC derivatives[1] including wrong-way risk (WWR), i.e. when the exposure to a counterparty is adversely correlated with the credit quality of the counterparty.

In Černý and Witzany [4] is introduced a semi-analytical formula for interest rate swap (IRS) CVA approximation including WWR (for details see Theorem 1 and 2 in Černý and Witzany [4]) if the following assumptions are satisfied: The swap rate is described by the following relationship

$$s_{t+\Delta t} = s_t \exp\left\{-\sigma^2 \Delta t/2 + \sigma\sqrt{\Delta t} Y\right\}, \quad Y \sim N(0,1) \tag{1}$$

where Y is decomposed into a systematic factor U and an idiosyncratic factor ε_1

$$Y = aU + \sqrt{1-a^2}\varepsilon_1, \quad a \in [-1,1].$$

In addition, the default time is defined as $\tau = S^{-1}(1-\Phi(Z))$ where $S(t) = \mathrm{e}\,\mathrm{e}^{-ht}$ is the exponential survival probability function with the constant hazard rate h (with respect to the annuity risk-neutral measure), and Φ is the cumulative distribution function of the standard Gaussian distribution. Z is again decomposed into the systematic factor U and an idiosyncratic factor ε_2

$$Z \ = bU + \sqrt{1-b^2}\varepsilon_2, \quad b \in [-1,1].$$

[1] Over-The-Counter derivatives are financial derivatives which are not exchange-traded, i.e. there is an additional credit risk of the counterparty. Conversely, exchange-traded derivatives are completely settled by the exchange.

U, ε_1 and ε_2 are independent standard Gaussian random variables.

In other words we assume that the random component of the interest rate and default time variables can be decomposed into a common systematic (e.g., macroeconomic) and different specific factors. These factors are independent with each other, however, we admit the dependence between the default time and interest rate which is expressed by the constant correlation coefficient.

Because the observation of default time means bankruptcy of the company, the correlation $\rho = ab$ can not be simply estimated using the observed data in contrast to IRS rate and the credit default swap (CDS) rate (or spread) which is closely related to the hazard rate (or default intensity). In the next section we introduce correlation coefficient calibration based on the historical data using the maximum likelihood estimation (MLE) method.

2. Calibration

As we noted earlier, available data from the market are IRS and CDS rates. So we denote the historical data observations of IRS rates as s_i and CDS rates as c_i for $i = 1, \ldots, n$, n is the number of daily (equidistant in general) observations and the maturity is fixed for both rates, e.g. 5Y. Then the market implied risk neutral probability of default to the time T can be approximated[2] from CDS rates as[3]

$$\mathbb{Q}\left[\tau \leq T | U_i\right] \approx 1 - \exp\left\{-\frac{c_i}{\mathrm{LGD}} T\right\} \tag{2}$$

assuming that the default time has approximately exponential distribution with the hazard rate $h_i \approx c_i / \mathrm{LGD}$ and survival function $S_i(t) \approx \mathrm{e}^{-h_i t}$. For simplicity, we consider a constant and deterministic LGD although, generally, it should be stochastic. Given $b \in (-1, 1)$ we can calculate implied values of the systematic factor U_i using market implied probability $\mathbb{Q}$ and Gaussian distribution of ε_2 as follows:

[2]We have to assume that the CDS premium is paid continuously but, in practice, the premium is paid in discrete times (typically semi-annually or quarterly).

[3]LGD is the Loss Given Default measured as a percentage of the exposure at default.

$$\begin{aligned}
\mathbb{Q}\left[\tau \leq T | U_i\right] &= \mathbb{Q}\left[S_i^{-1}(1-\Phi(Z)) \leq T | U_i\right] \\
&= \mathbb{Q}\left[Z \leq \Phi^{-1}(1-S_i(T)) \mid U_i\right] \\
&= \mathbb{Q}\left[bU_i + \sqrt{1-b^2}\varepsilon_2 \leq \Phi^{-1}(1-S_i(T)) \mid U_i\right] \qquad (3) \\
&= \mathbb{Q}\left[\varepsilon_2 \leq \frac{\Phi^{-1}(1-S_i(T)) - bU_i}{\sqrt{1-b^2}} \middle| U_i\right] \\
&= \Phi\left(\frac{\Phi^{-1}(1-S_i(T)) - bU_i}{\sqrt{1-b^2}}\right).
\end{aligned}$$

By combining (2) and (3) and assuming that $b \neq 0$ we obtain the following expression for the systematic factor

$$U_i = \frac{(1-\sqrt{1-b^2})\Phi^{-1}\left(1-\exp\left\{-c_i T/\mathrm{LGD}\right\}\right)}{b}. \qquad (4)$$

IRS and CDS rates time series are typically highly autocorrelated, and therefore Us will be autocorrelated too. We will assume that U_i follows simple AR(1) process, i.e.

$$U_i = \rho_U U_{i-1} + \sqrt{1-\rho_U^2}\varepsilon_{U,i}, \quad \rho_U \in [-1,1] \qquad (5)$$

with autocorrelation ρ_U corresponding to a mean reverting process for the systematic factor and idiosyncratic factor $\varepsilon_{U,i} \sim N(0,1)$ i.i.d. which is also independent of U_{i-1}. This time series can be easily decorrelated (see Chapter 4.3 in Cipra [3]) as

$$\varepsilon_{U,i} = \frac{U_i - \rho_U U_{i-1}}{\sqrt{1-\rho_U^2}} \sim N(0,1) \text{ i.i.d.}, \quad \rho_U \in (-1,1) \qquad (6)$$

The calibration will be done, as we noted, using the maximum likelihood method. First, we have to set up likelihood, resp. log-likelihood, functions. Although we can not assume the independence between IRS and CDS rates, the data can be transformed into mutually independent variables (see Witzany [5]).

The likelihood function (excluding transformation adjustment) for the idiosyncratic factor $\boldsymbol{\varepsilon}_U = (\varepsilon_{U,1}, \ldots, \varepsilon_{U,n})^T$ is

$$\mathcal{L}^*(\boldsymbol{\varepsilon}_U | \rho_U, b) = \prod_{i=2}^{n} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(U_i - \rho_U U_{i-1})^2}{2(1-\rho_U^2)}\right\}. \qquad (7)$$

Given U_i and Y_i (Y_i can be calculated from the IRS model, see equation (1)) we calculate implied ε_1, i.e.

$$\varepsilon_{1,i} = \frac{Y_i - aU_i}{\sqrt{1-a^2}} \sim N(0,1) \text{ i.i.d.}, \quad a \in (-1,1). \tag{8}$$

The likelihood function (excluding transformation adjustment) for $\boldsymbol{\varepsilon}_1 = (\varepsilon_{1,1}, \ldots, \varepsilon_{1,n})^T$ can be then set up as

$$\mathcal{L}^*(\boldsymbol{\varepsilon}_1 | a, b, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{(Y_i - aU_i)^2}{2(1-a^2)} \right\}. \tag{9}$$

When transforming an observed value x to another value $y = g(x)$ where we know the density function $f(y)$, an adjustment should be used to get a correct likelihood function of the observed value (see Theorem 3.7 in Anděl [1])

$$\mathcal{L}(x) = f(y) \left| \frac{\mathrm{d}y}{\mathrm{d}x} \right| = f(y) \left| g'(x) \right|, \tag{10}$$

in our case the first data transformation functions are

$$g_{\mathrm{CDS}}(x) = \frac{\Phi^{-1}\left(1 - \mathrm{e}^{-xT/\,\mathrm{LGD}}\right)\left(1 - \sqrt{1-b^2}\right)}{b} \tag{11}$$

and

$$g_{\mathrm{IRS}}(x) = \frac{\log x - \log s_t + \sigma^2 \Delta t/2}{\sigma\sqrt{\Delta t}}. \tag{12}$$

The second transformation functions follow from equations (6) and (8).

Joint likelihood function including all transfomations adjustments can be then written into conditional marginal likelihood functions in the following form

$$\mathcal{L}(\mathbb{c}, \mathbb{s} | a, b, \rho_U, \sigma) = \mathcal{L}^*(\boldsymbol{\varepsilon_U} | \rho_U, b) \left| J(c) \right| \left| J(U) \right| \mathcal{L}^*(\boldsymbol{\varepsilon_1} | a, b, \sigma) \left| J(s) \right| \left| J(Y) \right|$$

using the determinants of Jacobi matrices for the transformation $c_i \to U_i$

$$J(c) = \prod_{i=1}^{n} \frac{\partial U_i}{\partial c_i} = \prod_{i=1}^{n} \frac{\left(1 - \sqrt{1-b^2}\right) T \mathrm{e}^{-c_i T/\,\mathrm{LGD}}}{\varphi\left(\Phi^{-1}\left(1 - \mathrm{e}^{-c_i T/\,\mathrm{LGD}}\right)\right) b\,\mathrm{LGD}}$$

for the transformation $U_i \to \varepsilon_{U,i}$

$$J(U) = \prod_{i=1}^{n} \frac{\partial \varepsilon_{U,i}}{\partial U_i} = \prod_{i=1}^{n} \frac{1}{\sqrt{1-\rho_U^2}} = \left(\frac{1}{\sqrt{1-\rho_U^2}} \right)^n,$$

for the transformation $s_i \to Y_i$

$$J(s) = \prod_{i=1}^{n} \frac{\partial Y_i}{\partial s_i} = \prod_{i=1}^{n} \frac{1}{s_i \sigma \sqrt{T}},$$

and for the transformation $Y_i \to \varepsilon_{1,i}$

$$J(Y) = \prod_{i=1}^{n} \frac{\partial \varepsilon_{1,i}}{\partial Y_i} = \left(\frac{1}{\sqrt{1-a^2}} \right)^n,$$

where φ is the probability density function of the standard Gaussian distribution.

Now we can estimate the parameter vector $\theta = (a, b, \rho_U, \sigma)^T$ by maximazing logarithm of $\mathcal{L}(\mathbb{c}, \mathbb{s}|a, b, \rho_U, \sigma)$ with respect to θ.

Note that the correlation required for the CVA calculation should be estimated with respect to the risk-neutral measure $\mathbb{Q}$. The parameters estimated from the real data using MLE usually correspond to the real-world measure $\mathbb{P}$ (e.g. see Brigo and Mercurio [2]). However, the coefficient b corresponds to the risk-neutral probabilities of default and coefficient a is a swap rate diffusion parameter which is the same for both measures $\mathbb{P}$ and $\mathbb{Q}$. Therefore their product $\rho = ab$ corresponds to the risk-neutral measure $\mathbb{Q}$.

3. Application

We have calibrated the correlation coefficient based on Czech Republic government 5Y IRS and 5Y CDS rates obtained from *Thomson Reuters Eikon*. On the following figure IRS rates (full line) and CDS rates (dotted line) are observed from 9. 10. 2008 to 9. 4. 2013 (1166 daily observations).

The next figure shows an empirical correlation coefficient calculated on the moving window containing 200 observations.

The second figure shows that these two rates have been negatively or positively correlated in different periods and therefore we should take into account a non-zero correlation between the IRS rates and the default time. The correlation coefficient can be, generally, positive or negative[4]. In both cases it may cause the WWR depending on the type of the swap contract, i.e. whether it is fix-paying or fix-receiver swap. As in Černý and Witzany [4], we will assume that coefficients a and b are equal up to the sign, i.e. $|a| = |b|$.

[4]Negative, resp. positive, correlation between IRS and CDS rates indicates deterioration, resp. improvement, of the credit quality of the counterparty in case that interest rates are decreasing and vice versa.

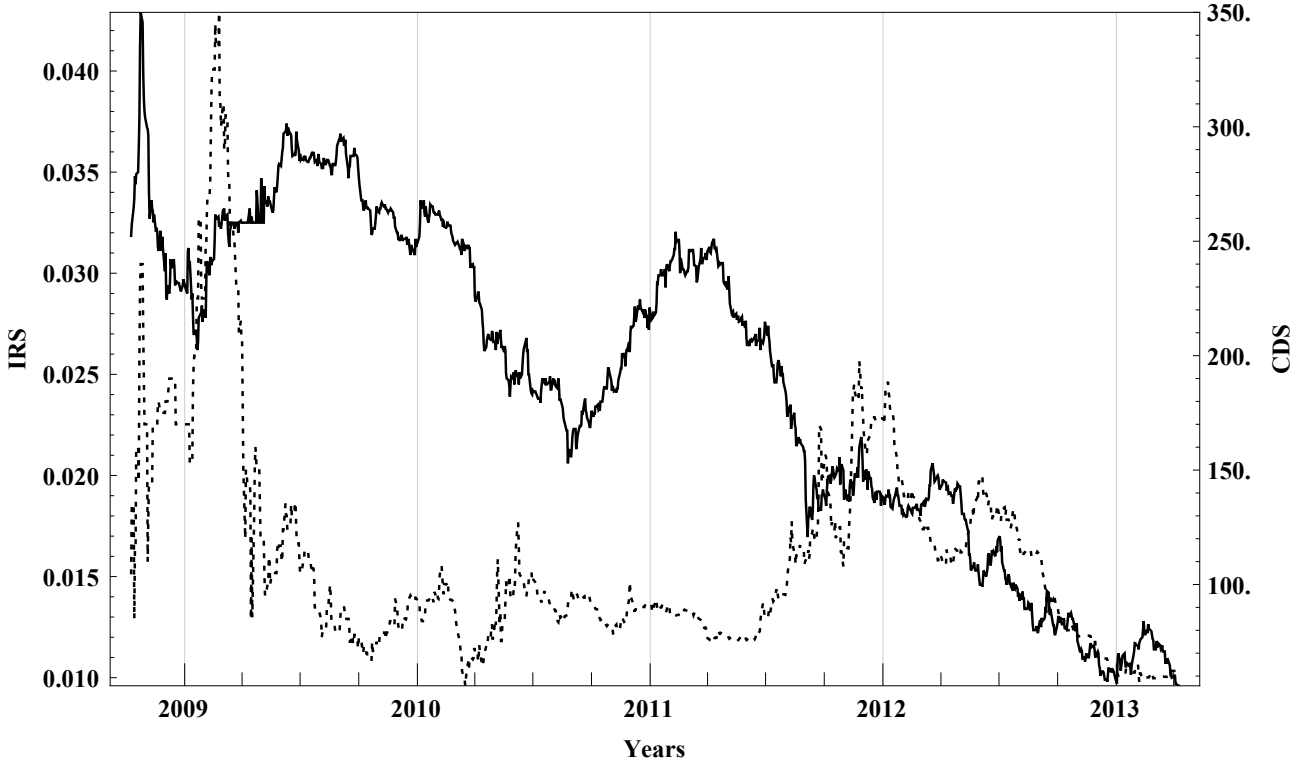


Figure 1: IRS rate (full line) and CDS rate (dotted line) development.

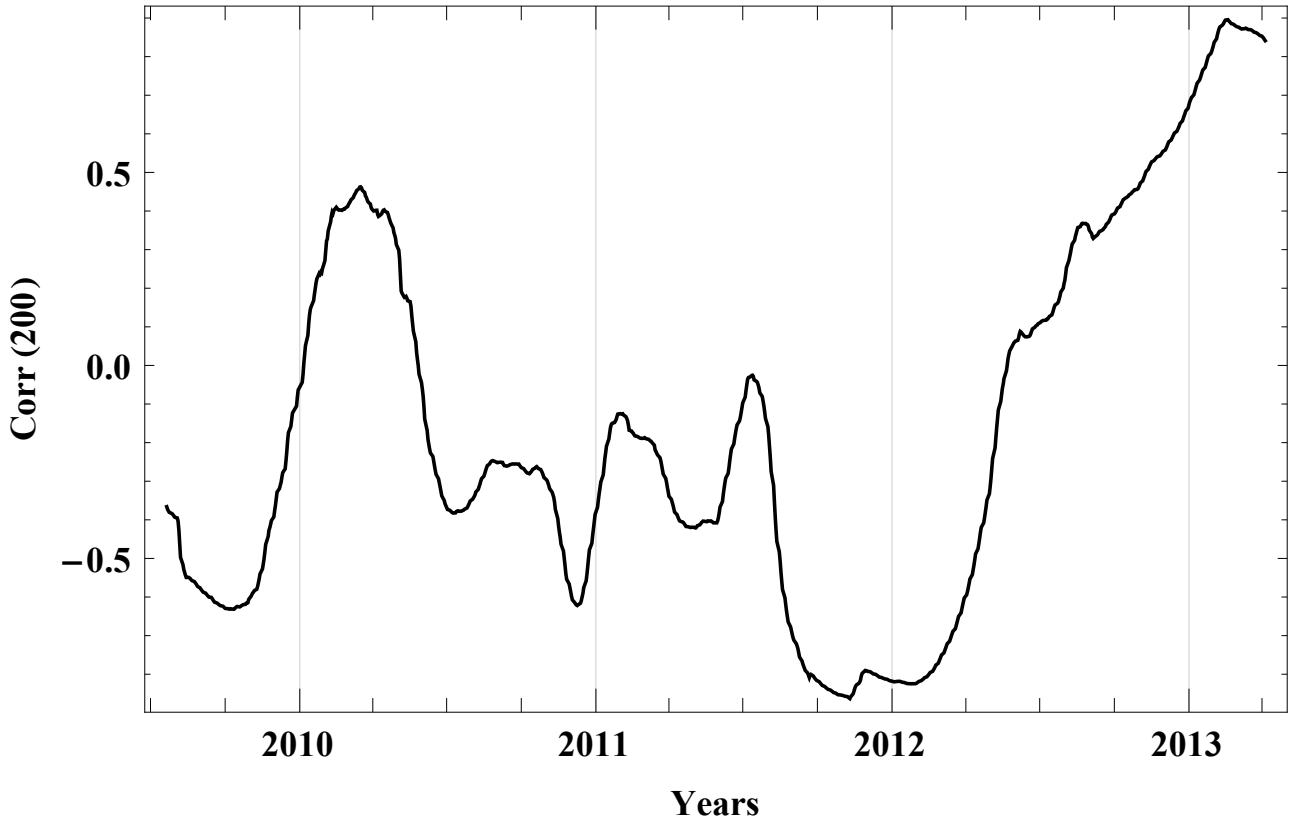


Figure 2: Empirical correlation between IRS and CDS rated on moving window of 200 observations.

The maximum likelihood estimation of the vector parameter θ, described in the previous section, was performed in *Wolfram Mathematica*® *9* software using function `NMaximize`. The table below shows the final estimates and an asymptotic error (standard deviation) of the estimates using observed Fisher information matrix.

Table 1: Parameter estimates by MLE.

Parameter	a	b	ρ_U	σ	$\rho = ab$
Estimate	-0.58296	-0.58296	0.999971	0.31828	0.33984
(**Error**)	(0.01975)	(0.02712)	(3.14×10^{-6})	(0.00704)	(0.03355)

From the results it is clear that we cannot assume independence between the IRS rates and the default time because the estimated correlation $\rho = 0.33984$ is quite high and definitely not equal to zero.

An unexpected secondary result is the value of the autocorrelation coefficient ρ_U, which is very close to one. A possible solution is to choose a different model for the decorrelation of the latent systematic factor.

Consider a 5Y IRS entered into on 9. 4. 2013 where the risky counterparty is the Czech Republic. CVA of this IRS at the trade date, i.e. the price of the swap, has substantially increased from 0.463 bps to 1.10752 bps when including the WWR and therefore it should not be neglected.

4. Conclusion

In this paper, we have discussed calibration of the correlation coefficient between the default time and the interest rates, which expresses the WWR for the calculation of IRS CVA using semi-analytical formulas presented in Černý and Witzany [4]. The calibration on the IRS and CDS rates is based on the well-known maximum likelihood estimation method.

In the application part of the paper, we estimated the correlation based on IRS and CDS rates of the Czech Republic government observed from 9. 10. 2008 to 9. 4. 2013 and calculated the estimation error using the observed Fisher information matrix. The results show that when interest rates fall, the default time decreases, i.e. the credit quality of the Czech Republic decreases.

We calculated CVA using estimated correlation and also the impact of the WWR on the resulting IRS CVA, where the risky counterparty is the Czech Republic. The price of the swap incorporating the WWR has significantly increased, and therefore WWR should not be ignored.

We realize that the AR(1) process for the decorrelation of the latent systematic factor is not the best choice due to the high autocorrelation coefficient ρ_U (see the results in Table 1). This problem can be solved by choosing a different model which is a subject of further research.

Acknowledgement

This research has been supported by the Czech Science Foundation Grant P402/12/G097 "Dynamical Models in Economics" and by SVV grant No. SVV-2014-260105.

References

[1] Anděl J.: *Základy matematické statistiky.* Praha, Matfyzpress, 2005.

[2] Brigo D., Mercurio F.: *Interest Rate Models: Theory and Practice – with Smile, Inflation and Credit.* Second edition, Springer Verlag, 2006.

[3] Cipra T.: *Finanční ekonometrie.* Praha, Ekopress, 2008.

[4] Černý J., Witzany J.: *Interest Rate Credit Valuation Adjustment.* 2014. URL at SSRN: `http://ssrn.com/abstract=2302519`.

[5] Witzany J.: *A Two-Factor Model for PD and LGD Correlation.* February 7, 2011. URL at SSRN: `http://ssrn.com/abstract=1476305`.

POUŽITÍ LOGISTICKÉ REGRESE PRO DIAGNOSTIKU VÝSKYTU RAKOVINY PROSTATY

USE OF LOGISTIC REGRESSION IN PROSTATE CANCER DIAGNOSIS

Kamila Fačevicová

Adresa: Katedra matematické analýzy a aplikací matematiky, PřF, Univerzita Palackého v Olomouci, 17. listopadu 12, 771 46 Olomouc

E-mail: kamila.facevicova@upol.cz

Abstrakt: Příspěvek za pomoci modelu logistické regrese a záznamů z vyšetření, která proběhla na Urologické klinice Fakultní nemocnice v Olomouci, hledá faktory, které mají významný vliv na výsledek rebiopsie, tento vliv kvantifikuje a následně také porovnává s vlivem týchž faktorů na výsledek první biopsie. Výsledkem biopsie, resp. rebiopsie, je v tomto případě myšleno, zda byl odhalen karcinom prostaty či nikoliv. Hlavní otázkou potom je, zda výsledek rebiobsie závisí na diagnóze, jež byla stanovena při biopsii první.

Klíčová slova: Logistická regrese, rakovina prostaty, rebiopsie.

Abstract: The aim of this paper is to find out factors, that have important influence on result of prostate rebiopsy, using the logistic regression model and records of examination from Urologic clinic of University hospital in Olomouc. This influence is quantified and compared with influence of the same factors on the result of first biopsy. Result of biopsy or rebiopsy means whether prostate cancer was found or not. The main question is then whether the result of rebiopsy depends on diagnosis from the first biopsy.

Keywords: Logistic regression, prostate cancer, rebiopsy.

1. Úvod

Tento příspěvek se věnuje problematice včasného odhalení karcinomu prostaty prostřednictvím správného vyhodnocení hodnot ukazatelů, měřených při běžných urologických prohlídkách. Zatímco publikace [2] se zabývala situací při prvním odběru vzorku tkáně (biopsii), nyní se zaměřujeme na odběr následný, tedy první rebiopsii. Pomocí modelu logistické regrese chceme nalézt faktory, které mají významný vliv na výsledek rebiopsie, tento vliv kvantifikovat a porovnat s vlivem týchž ukazatelů na výsledek biopsie první. Dále nás také zajímá, zda popř. jak souvisí výsledek rebiopsie s diagnózou stanovenou při první biopsii.

K dispozici byly záznamy z vyšetření celkem od 126 pacientů, z nichž 18 bylo při následné rebiopsii odhaleno zhoubné onemocnění prostaty. Tato vyšetření byla provedena v období od června 2006 do konce roku 2010 na Urologické klinice Fakultní nemocnice Olomouc a zahrnovala jen pacienty ve věku od 45 do 80 let s hladinou prostatického specifického antigenu (PSA) v krvi menší než 20 ng/ml a objemem prostaty do 150 ml.

Pro hledání vhodného modelu byly k dispozici tyto vysvětlující proměnné (výběrová střední hodnota; výberová směrodatná odchylka): věk pacienta (63,06; 6,09), hodnota PSA (7,32; 3,42), volné frakce PSA (fPSA) (1,16; 0,71) a PSA indexu – poměr volného a celkového PSA (16,05; 7,14), celkový objem prostaty (50,33; 22,10). Všechny tyto veličiny byly měřeny při první rebiopsii. Poslední sledovanou proměnnou byl výsledek předchozí biopsie, kdy jsme rozlišovali jen zda byla pacientovi diagnostikována hyperplazie či nikoliv, četnosti zbylých diagnóz totiž nebyly dostatečné. Případy, kdy byla již při první biopsii odhalena rakovina, nebyly do souboru zařazeny.

Vysvětlovanou proměnnou pak byl výsledek rebiopsie, ten byl považován za pozitivní (značíme 1) v případě, kdy odhalil karcinom prostaty, a za negativní (značíme 0) ve zbylých případech. Data byla zpracována za pomoci softwaru R.

Tabulka 1: Rozdělení pozitivních a negativních výsledků rebiopsie v závislosti na výsledku první biopsie.

Biopsie \ Rebiopsie	Negativní	Pozitivní	Celkem
Hyperplazie	63	13	76
Zbylé	45	5	50
Celkem	108	18	126

2. Model logistické regrese

Obecný zápis modelu logistické regrese je

$$\ln \frac{\pi}{1-\pi} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k,$$

kde π je pravděpodobnost výskytu sledované události, tedy že vysvětlovaná proměnná bude rovna jedné a regresory $x_1, x_2, \ldots, x_k$ pak zastupují jednotlivé vysvětlující proměnné. Pomocí logistické regrese tedy namísto pravděpodobnosti výskytu dané události odhadujeme logaritmus poměru šancí na tuto

Tabulka 2: Bodové a intervalové odhady parametrů ($\alpha = 0{,}05$).

Par.	Regresor	Bod. odhad	Int. odhad
β_0	—	−2,1435	$\langle -6{,}1931;\ 1{,}9061\rangle$
β_1	PSA	0,4390	$\langle 0{,}2252;\ 0{,}6529\rangle$
β_2	fPSA	−2,7036	$\langle -5{,}5837;\ 0{,}1764\rangle$
β_3	Hyperplazie	3,3443	$\langle 0{,}5539;\ 6{,}1348\rangle$
β_4	Objem prostaty	−0,0884	$\langle -0{,}1655;\ -0{,}0113\rangle$
β_5	Interakce fPSA a hyperplazie	−2,0247	$\langle -4{,}0313;\ -0{,}0181\rangle$
β_6	Interakce fPSA a objemu	0,0467	$\langle 0{,}0081;\ 0{,}0853\rangle$

událost. Hledané parametry $\beta_1, \beta_2, \ldots, \beta_k$ tak lze interpretovat jako logaritmus tohoto poměru šancí v případě, kdy příslušný regresor zvýšíme o jedničku a zbylé zůstanou neměnné. Absolutní člen β_0 je pak logaritmus šance na výskyt události v situaci, kdy jsou všechny regresory nulové.

Vztah mezi výsledkem rebiopsie a výše uvedenými regresory nejlépe popisuje model (AIC = 85,04)

$$\begin{aligned}\ln \frac{\pi}{1-\pi} = {} & -2{,}14 + 0{,}44 \cdot \mathrm{PSA} - 2{,}70 \cdot \mathrm{fPSA} + 3{,}34 \cdot \mathrm{hp} - \\ & -0{,}09 \cdot \mathrm{objem} - 2{,}03 \cdot \mathrm{fPSA} \cdot \mathrm{hp} + 0{,}05 \cdot \mathrm{fPSA} \cdot \mathrm{objem}.\end{aligned}$$

V tomto modelu π značí pravděpodobnost, že bude výsledek rebiopsie pozitivní a bude tedy odhalen karcinom prostaty. Regresor „hp“ je roven 1, pokud byla pacientovi při první biopsii zjištěna hyperplazie prostaty, ve zbylých případech je nulový. Za všechny zbylé regresory dosazujeme hodnoty naměřené pacientovi při vyšetření, jež předcházelo rebiopsii.

Bodové i intervalové odhady jednotlivých parametrů jsou uvedeny v Tabulce 2.

3. Interpretace výsledků

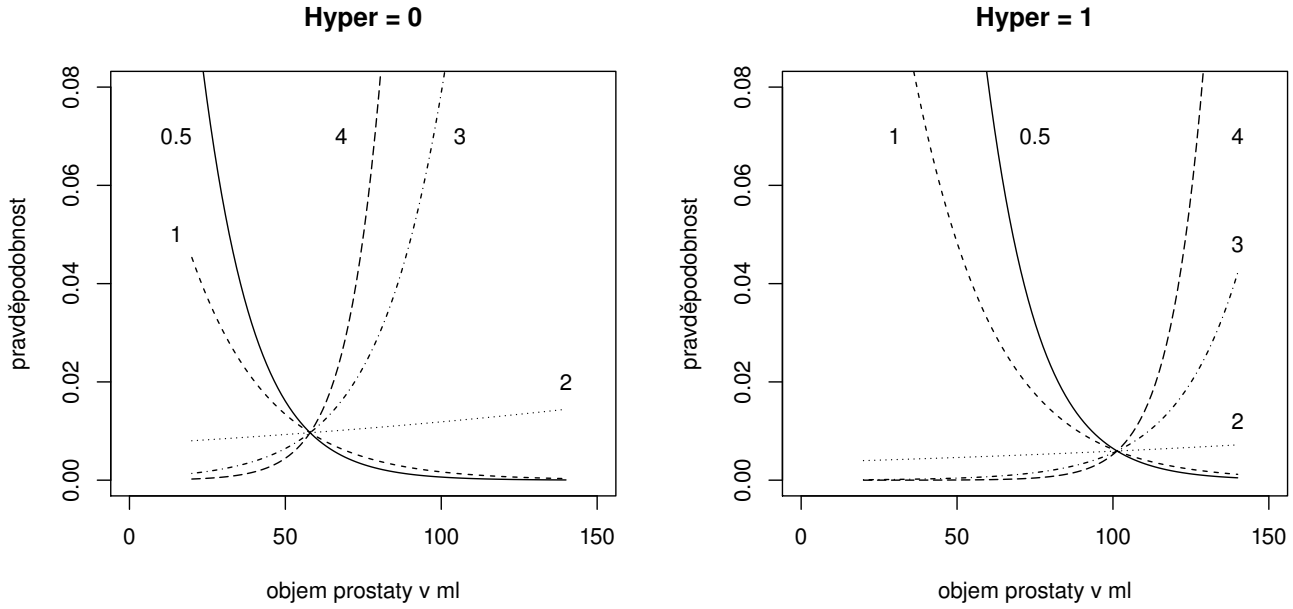
Pro správnou interpretaci modelu je potřeba rozlišit, zda pacient při první biopsii trpěl hyperplazií, nebo ne. Zabývejme se nejprve situací, kdy pacientovi při první biopsii byla diagnostikována hyperplazie, v tomto případě platí:

1. Pokud je při rebiopsii naměřeno fPSA menší než asi 1,9 ng/ml, pak s rostoucím objemem klesá pravděpodobnost diagnostiky rakoviny, pokud je ale fPSA naopak větší než 1,9 ng/ml, roste s objemem i daná

pravděpodobnost a vliv objemu je tak opačný než v předchozím případě.

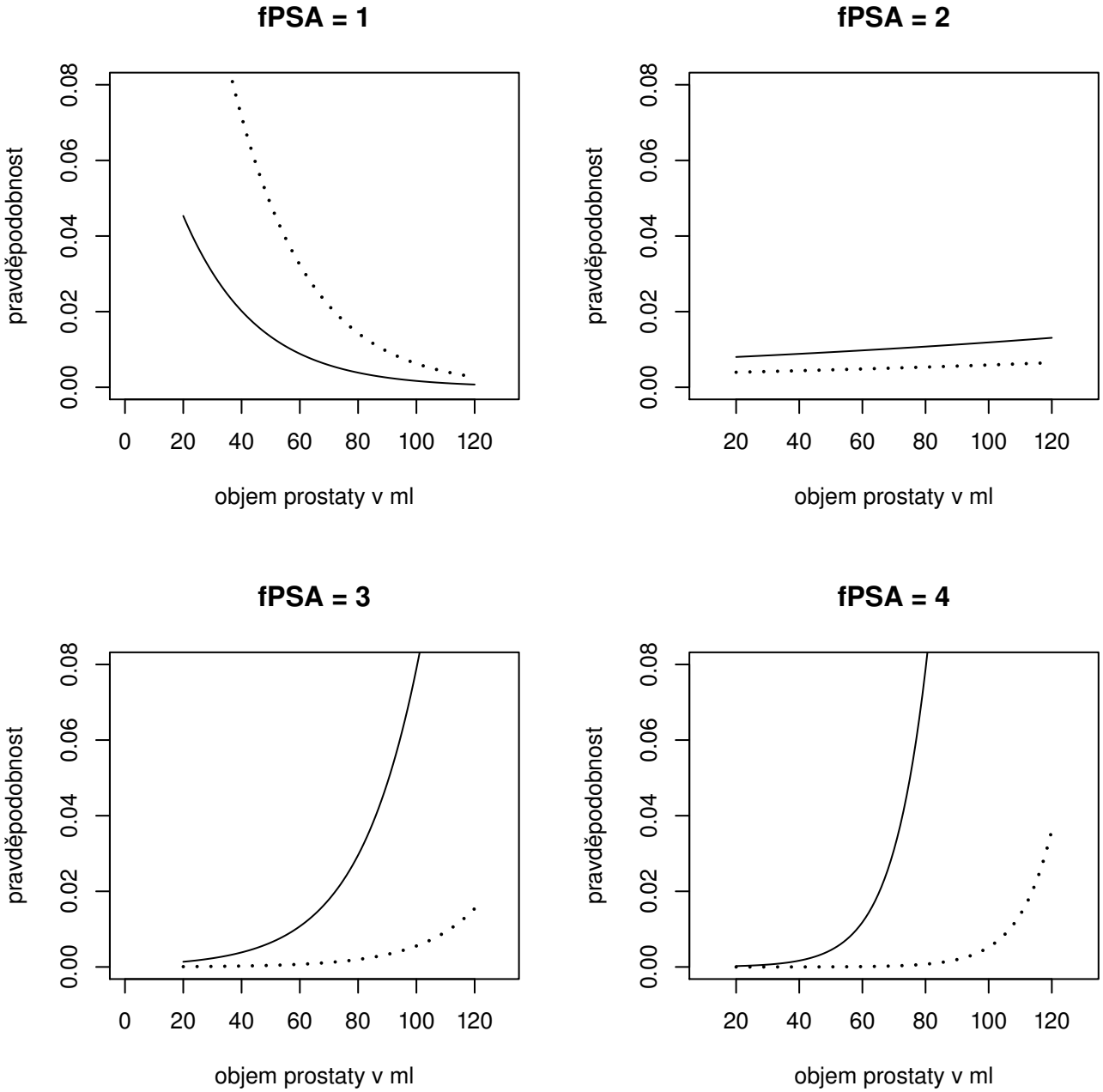
2. Obdobně platí, že při objemu menším než asi 100 ml, znamená větší fPSA menší pravděpodobnost pozitivní biopsie a při překročení hranice 100 ml je tomu opět naopak a s rostoucím fPSA roste i pravděpodobnost.

Druhou situací je, že výsledek první biopsie byl jiný než hyperplazie. I v tomto případě jsou ale závěry 1. a 2. platné s tím rozdílem, že hraniční hodnota objemu je nyní jen 57,5 ml.



Obrázek 1: Grafy zachycují vývoj pravděpodobnosti pozitivního výsledku rebiopsie v situaci, kdy je pacientova hladina PSA v krvi rovna 6 ng/ml, v závislosti na hladině fPSA a objemu prostaty. Zatímco levý graf znázorňuje vývoj v případě, kdy byla pacientovi při první biopsii diagnostikována hyperplazie prostaty, pravý graf se věnuje případům zbylým.

Výše uvedené závěry lze shrnout tak, že nezávisle na výsledku první biopsie je vždy riziková kombinace vysokého fPSA a vysokého objemu prostaty nebo nízkého fPSA a malého objem prostaty. Pokud je ale jeden z těchto faktorů vysoký a druhý nízký, je predikce výsledku rebiopsie příznivější. Tuto vlastnost lze snadno pozorovat i v grafech na Obrázku 1, kde je zachycen vývoj pravděpodobnosti pozitivní biopsie v závislosti na hladině fPSA a objemu prostaty. Hodnoty jsou počítány pro modelovou situaci, kdy je PSA rovno 6 ng/ml.



Obrázek 2: Grafy zachycují vývoj pravděpodobnosti pozitivní rebiopsie v závislosti na objemu prostaty a hladině fPSA, kterou uvažujeme postupně 1, 2, 3 a 4 ng/ml. Ve všech případech zároveň předpokládáme, že hladina PSA v krvi je 6 ng/ml. Přerušovaná křivka značí situaci, kdy byla pacientovi nejprve diagnostikována hyperplazie prostaty, plná křivka pak zachycuje případy, kdy byla původní diagnóza jiná.

Analýza dále ukázala rozdílnost pravděpodobností výskytu karcinomu mezi skupinou pacientů, jimž byla při první biopsii diagnostikována hyperplazie a skupinou pacientů s diagnózou odlišnou. Pokud je fPSA menší než 1,65 ng/ml byla tato pravděpodobnost vyšší pro první skupinu, je-li ale fPSA

vyšší, je tomu právě naopak. I tuto skutečnost lze pozorovat v grafech na Obrázku 2. Parametr příslušející hladině PSA v krvi je roven 0,439 což znamená, že srovnáváme-li dvě skupiny pacientů lišící se pouze v množství PSA v krvi, kdy jedna z nich jej má o 1 ng/ml vyšší, pak má tato skupina $e^{0,439} = 1,55$krát vyšší šanci na záchyt karcinomu prostaty než skupina s nižší hladinou PSA. Obecně lze tedy říct, že vyšší hodnoty PSA v krvi znamenají vyšší riziko pozitivní rebiopsie.

Vliv věku pacienta a PSA indexu na výsledek rebiopsie se ukázal jako nevýznamný.

3.1. Senzitivita a specificita

Dosazením vstupních dat do modelu a stanovením pevné hodnoty π, při jejímž překročení předpokládáme, že bude výsledek biopsie pozitivní, lze určit senzitivitu a specificitu tohoto modelu. Senzitivita je pravděpodobnost, že bude model predikovat pozitivní výsledek rebiopsie v případě, kdy se karcinom na prostatě skutečně nachází. Její odhad získáme ze vstupních dat jako podíl počtu pacientů s pozitivním výsledkem rebiopsie a odhadem pravděpodobnosti větším než stanovené π a počtu všech pacientů, u nichž byl při rebiopsii odhalen karcinom.

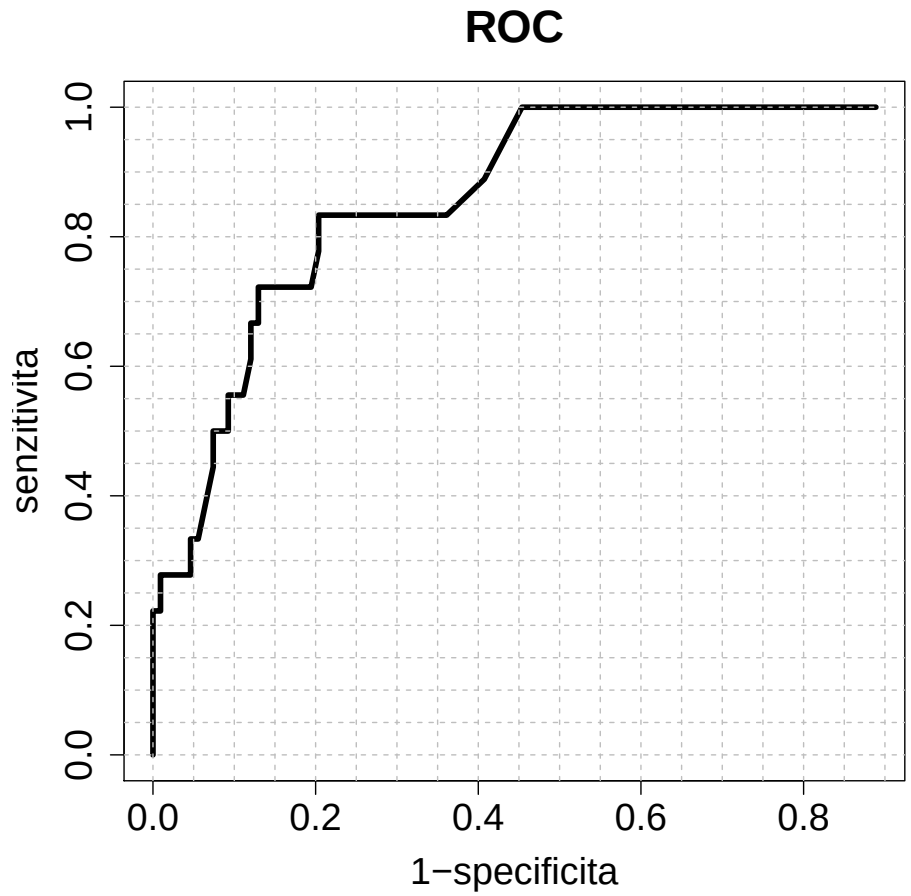
Oproti tomu specificita značí pravděpodobnost, že bude model u zdravých pacientů predikovat negativní výsledek rebiopsie. Tentokrát získáme její odhad jako podíl počtu všech pacientů s negativní rebiopsií a odhadem pravděpodobnosti menším než π a počtu všech pacientů s negativním výsledkem rebiopsie. Vztah mezi senzitivitou a specificitou, při různých hraničních hodnotách π, lze znázornit pomocí ROC křivky, viz Obrázek 3.

3.2. Srovnání s modelem pro 1. biopsii

Pro první biopsii byl v [2] nalezen odlišný model:

$$\ln \frac{\pi}{1-\pi} = -3,917 + 0,053 \cdot \text{vk} - 0,027 \cdot \text{objem} + 0,144 \cdot \text{PSA} - 0,305 \cdot \text{fPSA}.$$

Při odhadu pravděpodobnosti byl tedy nově významný vliv věku pacienta. Ze srovnání intervalových odhadů parametrů v tomto modelu a v modelu pro rebiopsii navíc vyplývá, že vliv hladiny PSA v krvi je v případě rebiopsie významně vyšší. Zatímco při první biopsii je intervalový odhad příslušného parametru $\langle 0,0888; 0,1992\rangle$, v modelu pro rebiopsii je to $\langle 0,2252; 0,6529\rangle$. Rozdíly mezi odhady zbylých porovnatelných parametrů nebyly statisticky významné.



Obrázek 3: Graf zachycující hodnoty senzitivity a 1-specificity modelu, při různých hraničních hodnotách pravděpodobnosti pozitivní rebiopsie π.

4. Závěr

V příspěvku je nalezen vhodný model pro odhad pravděpodobnosti, že bude při rebiopsii odhalen karcinom prostaty. Pro tuto predikci se jako významné faktory ukázaly hladina celkového a volného PSA v krvi pacienta, objem jeho prostaty a také výsledek předchozí biopsie, přičemž platí, že rizikovými kombinacemi jsou malý objem prostaty a nízká hladina fPSA nebo naopak velký objem a vysoká hladina fPSA v krvi, a to bez ohledu na předchozí diagnózu.

Pro správnou interpetaci výsledků je však potřeba brát v úvahu i šíři intervalových odhadů regresních parametrů, které jsou v některých případech poměrně široké. Vyšší relevanci výsledků by napomohla analýza většího souboru pacientů, jako tomu je například v práci [4]. Zároveň je nutné si uvědomit, že se prezentované závěry vztahují pouze na skupinu pacientů uvedenou v úvodu příspěvku, tedy na pacienty ve věku od 45 do 80 let s hladinou PSA menší než 20 ng/ml a objemem prostaty menším než 150 ml.

Poděkování

Příspěvek byl podpořen Operačním programem vzdělávání pro konkurenceschopnost – Evropský sociální fond (číslo projektu CZ.1.07/2.3.00/20.0170 Ministerstva školství mládeže a tělovýchovy České republiky) a také grantem PrF_2014_028 interní grantové agentury Univerzity Palackého v Olomouci.

Literatura

[1] Agresti A.: *Categorical Data Analysis*. 2. vyd., Wiley, 2002.

[2] Fačevicová K.: *Použití logistické regrese pro diagnostiku výskytu rakoviny prostaty*, diplomová práce, Univerzita Palackého v Olomouci, 2012.

[3] Grepl M., Študent V., Fürst T., Fürstová J.: Prostate cancer detection yield in repeated biopsy is independent of the diagnosis of earlier biopsies, *Biomedical papers* **4**, 297–305, 2009.

[4] Vencálek O., Fačevicová K., Fürst T., Grepl M.: When less is more: A simple predictive model for repeated prostate biopsy outcomes, *Cancer Epidemiology* **37** (6), 864–869, 2013.

ROVNICE NA ČASOVÝCH ŠKÁLÁCH A NÁHODNÉ PROCESY

EQUATIONS ON TIME SCALES AND STOCHASTIC PROCESSES

Michal Friesl

Adresa: Katedra matematiky, Fakulta aplikovaných věd, Západočeská univerzita v Plzni, Univerzitní 8, 306 14, Plzeň

E-mail: friesl@kma.zcu.cz

Abstrakt: Matematická analýza se v poslední době zabývá některými typy dynamických rovnic, sjednocujícím pohledem zahrnujícím prostřednictvím obecné časové škály jak diferenciální rovnice ve spojitém čase, tak diferenční v čase diskrétním, a zkoumá jejich vlastnosti. Ve speciálních případech lze řešení některých rovnic chápat jako marginální rozdělení markovského řetězce. Ukážeme, že některé výsledky tak lze přenést či dovodit ze známých vlastností odpovídajícího náhodného procesu.

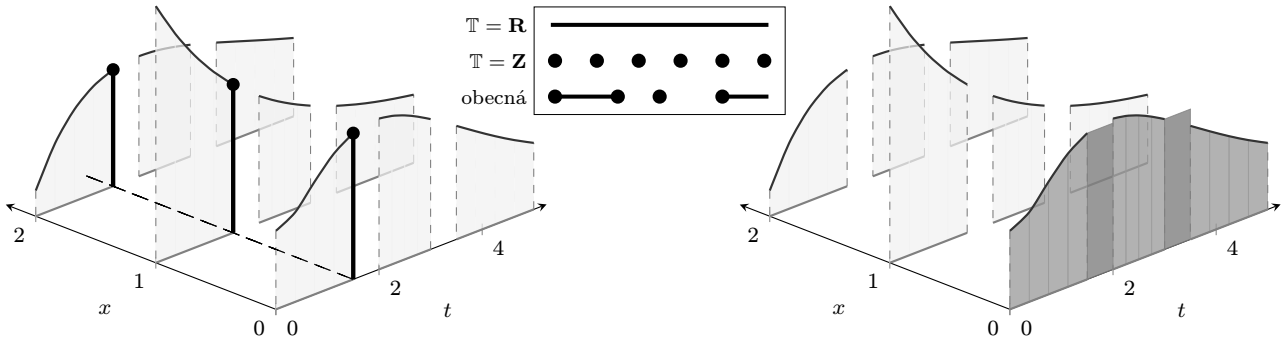
Klíčová slova: Markovský řetězec, časová škála, dynamická rovnice.

Abstract: Mathematical analysis has recently dealt with dynamic equations, unifying by a general time scale both differential equations in continuous time and difference equations in discrete time, and examined their properties. In special cases, solutions of such equations can be understood as marginal distributions of a Markov chain. We show that this way some results can be transferred or inferred from the known properties of the corresponding random process.

Keywords: Markov chain, time scale, dynamic equation.

1. Úvod

Tématem, ke kterému se tento text vztahuje, jsou dynamické rovnice na časových škálách. *Časovou škálou* se rozumí libovolná uzavřená množina časových okamžiků $\mathbb{T} \subset \mathbf{R}$. Speciálním případem časové škály je interval či množina diskrétních časů, ale obecně může mít množina $\mathbb{T}$ strukturu složitější. Funkce zrnitosti $\mu_t = \inf\{s \in \mathbb{T}; s > t\} - t$ udává pro jednotlivé časy $t \in \mathbb{T}$ vzdálenost k nejbližšímu dalšímu okamžiku časové škály; pokud $t = \max \mathbb{T}$, rozumí se $\mu_t = 0$. Body t s $\mu_t = 0$ označujme jako zprava spojité, body s $\mu_t > 0$ jako zprava diskrétní. V bodech časové škály definujeme zobecněnou delta derivaci



Obrázek 1: Příklady časové škály (uprostřed) a graf funkce $u = u(t, x)$ s naznačeným prostorovým součtem $\sum_x u_x(t)$ pro $t = 1{,}5$ (vlevo) a plochou časového integrálu $\int_0^\infty u_x(t)\,\Delta t$ pro $x = 0$ (vpravo).

(spojité) funkce u předpisem

$$u^{\Delta_t}(t) = \begin{cases} \dfrac{u(t+\mu_t) - u(t)}{\mu_t}, & \text{je-li } \mu_t > 0, \\ \lim_{h\to 0} \dfrac{u(t+h) - u(t)}{h}, & \text{je-li } \mu_t = 0. \end{cases}$$

V případě bodů zprava diskrétních se tedy pracuje s diferencí, v případě bodů spojitých s obyčejnou derivací. Analogicky se definuje zpětná nabla derivace u^{∇_t}.

Dále uvažujeme diskrétní stavový prostor X, kterým v dalším bude množina celých, případně nezáporných celých čísel, $X = \mathbf{Z}$, popř. $X = \mathbf{N}_0$, a funkce $u : \mathbb{T} \times X \to \mathbf{R}$ definované na kartézském součinu. Podle potřeby budeme používat také alternativní označení $u(t) := (u(t, x), x \in X) = (u_x(t))_{x\in X} : \mathbb{T} \to \ell^\infty(X)$. Pro lineární omezený operátor $A : \ell^\infty(X) \to \ell^\infty(X)$ budeme zkoumat omezená řešení u rovnice

$$u^{\Delta_t}(t) = A\,u(t), \quad t \in \mathbb{T},$$

s „počáteční“ podmínkou $u(t_0) = u^0$ v nějakém čase $t_0 \in \mathbb{T}$. Tuto rovnici můžeme chápat po složkách jako soustavu *dynamických* rovnic (pro každé x jedna). Dynamické rovnice jsou probírány např. v [1, 3]. Řešení soustavy lze za předpokladu, že $I + A\mu_t$ je pro každé $t \in \mathbb{T}$ invertibilní, psát pomocí hodnot zobecněné exponenciely e_{A,t_0} jako $u(t) = \big(e_{A,t_0}(t)\big)(u^0)$, blíže viz [6].

Mezi význačné příklady uvedené rovnice se řadí rovnice transportní a difuzní, tedy rovnice tvaru

$$u^{\Delta_t} = -ku^{\nabla_x} \qquad \text{neboli} \qquad u_x^{\Delta_t} = -k(u_x - u_{x-1}), \quad x \in X,$$
$$u^{\Delta_t} = (u^{\nabla_x})^{\Delta_x} \qquad \text{neboli} \qquad u_x^{\Delta_t} = u_{x+1} - 2u_x + u_{x-1}, \quad x \in X,$$

či obecněji tvaru

$$u_x^{\Delta_t} = au_{x-1} + bu_x + cu_{x+1}, \quad x \in X,$$

kterou nazývejme rovnice difuzního typu. Koeficienty k, a, b, c v rovnicích jsou reálné konstanty, prostorem X myslíme $X = \mathbf{Z}$. Jestliže však u transportní rovnice uvažované níže se ukazuje, že $u_x = 0$ pro $x < 0$, můžeme u ní počítat s efektivním stavovým prostorem $X = \mathbf{N}_0$. Rovnice spadají do oblasti parciálních dynamických rovnic, diskutované v [5] či [7]. V článcích [6, 7, 8] se mimo jiné vyšetřuje, zda a za jakých podmínek mají omezená řešení těchto rovnic některé užitečné vlastnosti. Mezi zkoumané vlastnosti patří

(i) zachování znaménka: pokud $u(t_0) \geqq 0$, pak $u(t) \geqq 0$ pro všechna $t \in \mathbb{T}$,
(ii) princip maxima: pokud $u(t_0) \leqq K$, pak $u(t) \leqq K$ pro všechna $t \in \mathbb{T}$,
(iii) zachování prostorového integrálu (součtu): $\sum_x u_x(t) = \sum_x u_x(t_0)$,
(iv) konečnost časových integrálů: $\int_{\mathbb{T}} u_x(t)\,\Delta t$ či dokonce jejich rovnost pro všechna $x \in X$.

Vzhledem k linearitě problému stačí v principu tyto vlastnosti vyšetřovat při jednotkové počáteční podmínce

$$u(t_0, x) = \begin{cases} 1, & x = 0, \\ 0, & x \neq 0. \end{cases} \tag{1}$$

V následujícím textu zdůvodníme či odvodíme vlastnosti řešení z pravděpodobnostního pohledu. V tom, co následuje, budeme předpokládat

- nezápornou časovou škálu $\mathbb{T} \subset \langle 0, \infty)$, pro kterou $\min \mathbb{T} = 0$, $\sup \mathbb{T} = \infty$, navíc s vlastností, že na každém omezeném intervalu je počet diskrétních bodů časové škály $\mathbb{T}$ konečný,
- počáteční čas $t_0 = 0$ a jednotkovou počáteční podmínku (1).

Náš přístup nejprve ukážeme na jednodušším příkladu transportní rovnice, pak rozebereme rovnici difuzního typu. Využívat při tom budeme základních poznatků o markovských řetězcích.

2. Transportní rovnice

Na časové škále $\mathbb{T}$ a se stavovým prostorem $X = \mathbf{N}_0$ (či $X = \mathbf{Z}$) uvažujeme soustavu rovnic

$$u_x^{\Delta_t} = -k(u_x - u_{x-1}), \quad x \in X, \tag{2}$$

přičemž předpokládáme, že $k > 0$ a také že $k\mu_t \leqq 1$ pro všechna $t \in \mathbb{T}$ (podmínka svazující velikost mezer v časové škále s koeficientem k); při $X = \mathbf{N}_0$ značí u_{-1} nuly. V této kapitole ukážeme odvození následujícího tvrzení.

Tvrzení 2.1. *Řešení u rovnice* (2) *při libovolné časové škále má za daných předpokladů vlastnosti* (i)–(iv)*; hodnota časového integrálu je vždy* $1/k$.

Zabývejme se nejprve diskrétní časovou škálou s jednotkovými kroky, $\mathbb{T} = \{0, 1, 2, \dots\}$. Tedy případem $\mu_t = 1$, $t \in \mathbb{T}$ (a vzhledem k podmínce $k\mu_t \leqq 1$ tak zde $k \leqq 1$). Soustavu (2), která má v diskrétních časech t obecně tvar

$$\frac{u_x(t+\mu_t) - u_x(t)}{\mu_t} = -k\big(u_x(t) - u_{x-1}(t)\big), \quad x \in X,$$

můžeme přepsat do tvaru

$$u_x(t+1) = (1-k)u_x(t) + ku_{x-1}(t), \quad x \in X.$$

neboli maticově jako $u(t+1) = P^{\mathsf{T}}u(t)$. Ta je dobře známa z teorie pravděpodobnosti, řešením jsou $u_x(t) = \mathrm{P}(X_t = x)$, pravděpodobnosti homogenního markovského řetězce $(X_0, X_1, \dots)$ s diskrétním časem a maticí pravděpodobností přechodu

$$P = \begin{pmatrix} 1-k & k & & \\ & 1-k & k & \\ & & \ddots & \ddots \end{pmatrix}.$$

V každém čase řetězec s pravděpodobností k přejde do stavu o 1 vyššího a s pravděpodobností $1-k$ setrvá v současném stavu. V čase 0 dle počáteční podmínky začíná ve stavu 0. Jde tedy o Bernoulliův proces.

Poznámka. Pokud by nás zajímal tvar řešení soustavy, připomeňme, že vychází $u_x(t) = \binom{t}{x} k^x (1-k)^{t-x}$ pro $x = 0, 1, \dots, t$, $u_x(t) = 0$ jinde.

Jelikož posloupnost $(u_x(t))_{x \in X}$ tvoří pro každé t rozdělení, platí automaticky $0 \leqq u_x(t) \leqq 1$ a také $\sum_x u_x(t) = 1$, automaticky tedy dostáváme vlastnosti zachování znaménka a prostorového integrálu. Jednoduše dostaneme i princip maxima: pravděpodobnosti přechodu $p_{xy}(t)$ po t krocích závisí na x a y jen prostřednictvím rozdílu $x - y$, dostáváme tak

$$u_x(t) = \sum_k u_{x-k}(0)p_{x-k,x}(t) = \sum_k u_{x-k}(0)p_{x,x+k}(t) \leqq \sup_x u_x(0) \cdot 1.$$

Při zjišťování časových integrálů (resp. součtů) si uvědomíme, že vyjadřují celkovou očekávanou dobu strávenou řetězcem ve stavu x. Doba τ do opuštění stavu má geometrické rozdělení. Zároveň každý stav $x \geqq 0$ je navštíven jednou, můžeme tedy přímo spočítat

$$\int_0^\infty u_x(t)\,\Delta t = \sum_{t=0}^{\infty} u_x(t) = \sum \mathrm{E}\, I_{X_t=x} = \mathrm{E} \sum I_{X_t=x} = \mathrm{E}\,\tau = 1/k,$$

což je hodnota konečná a pro všechny stavy $x \geqq 0$ stejná. Všechny zkoumané vlastnosti jsou tak pro $\mathbb{T} = \{0, 1, 2, \dots\}$ vysvětleny.

Nyní se podívejme na případ spojité časové škály $\mathbb{T} = \langle 0, \infty)$. V soustavě (2) teď ve všech časech $t \in \mathbb{T}$ vystupují obyčejné derivace, její tvar

$$u'_x(t) = -ku_x + ku_{x-1}, \quad x \in X,$$

můžeme číst jako z teorie pravděpodobnosti známou soustavu $u'(t) = Q^{\mathsf{T}}u(t)$ Kolmogorovových prospektivních rovnic pro pravděpodobnosti $u_x(t)$ homogenního markovského řetězce $(X_t, t \geqq 0)$ se spojitým časem a s maticí intenzit přechodu

$$Q = \begin{pmatrix} -k & k & & \\ & -k & k & \\ & & \ddots & \ddots \end{pmatrix}.$$

Celková intenzita výstupu z každého stavu je k a možný je přechod vždy jen do stavu o 1 vyššího. Rovnice tak popisuje Poissonův proces.

Poznámka. Řešením soustavy tedy jsou pravděpodobnosti $u_x(t) = \mathrm{e}^{-kt}\,\frac{(kt)^x}{x!}$, $x = 0, 1, \dots$

Odvození vlastností (i)–(iv) může proběhnout analogicky jako v diskrétním případě. U časových integrálů tentokrát využijeme, že doba do opuštění stavu τ má u markovského řetězce se spojitým časem exponenciální rozdělení, konkrétně tedy

$$\int_0^\infty u_x(t)\,\Delta t = \int_0^\infty u_x(t)\mathrm{d}t = \int \mathrm{E}\, I_{X_t=x}\mathrm{d}t = \mathrm{E}\int I_{X_t=x}\mathrm{d}t = \mathrm{E}\,\tau = 1/k.$$

Integrály jsou tedy opět konečné a stejné pro všechny stavy x, jejich hodnota se i shoduje s hodnotou z diskrétního případu.

Nakonec jsme si nechali případ obecné časové škály $\mathbb{T}$. V závislosti na tom, je-li časový okamžik $t \in \mathbb{T}$ zprava spojitý ($\mu_t = 0$), nebo diskrétní ($\mu_t > 0$), je rovnice z (2) tvaru

$$u'_x(t) = -ku_x(t) + ku_{x-1}(t) \quad \text{či} \quad u_x(t + \mu_t) = (1 - k\mu_t)u_x(t) + k\mu_t\, u_{x-1}(t).$$

Tedy rovnice pro $u_x(t)$, které jsou pravděpodobnostmi výskytu stavu x v čase t u markovského řetězce se *smíšeným časem*. Tento řetězec chápeme tak, že doba čekání na výstup ze stavu se ve spojitých časech řídí intenzitou k, zatímco v diskrétních pravděpodobností $k\mu_t$, úměrnou délce následující

mezery (tuto mezeru počítáme ještě do doby pobytu ve stavu). Po výstupu ze stavu řetězec přejde do stavu o 1 vyššího (ve vnořeném řetězci přechod $x \to x+1$ nastává s pravděpodobností 1).

Analogicky jako v předchozích případech dostaneme vlastnosti řešení (i)–(iii). Platí i vlastnost (iv), shoda časových integrálů. Ty totiž budou stejné jako např. ve spojitém případě, protože časové integrály řešení na struktuře časové škály $\mathbb{T}$ nezávisí, jak je ukazáno v [2]. Jsme-li totiž v určitém okamžiku $t \in \mathbb{T}$ ve sledovaném stavu x, pak

- je-li t spojitý, tak následující interval spojitých bodů (za předpokladů o $\mathbb{T}$ skutečně následuje interval), řekněme délky s, do střední hodnoty do opuštění stavu přispěje hodnotou $\int_0^s \mathrm{e}^{-kt}\,\mathrm{d}t$ a s pravděpodobností e^{-ks}, s níž přechod během něj nenastane, budou započteny i příspěvky následujících intervalů,
- je-li t diskrétní, tak následující interval délky $\mu = \mu_t$ (tj. následující mezera v časové škále) do střední hodnoty přispěje délkou μ a s pravděpodobností $(1 - k\mu)$ se přidají příspěvky následujících časů.

Porovnáním zjišťujeme, že diskrétní mezera délky μ přispívá stejně jako spojitý interval délky $s_\mu = -\frac{1}{k}\ln(1-k\mu)$. Vložíme-li fiktivně místo diskrétních mezer do časové škály $\mathbb{T}$ intervaly spojitých bodů příslušných délek s_μ (bude vždy $s_\mu > \mu$), pak řešení u se přechodem ke spojité škále sice změní (na rozdělení Poissonova procesu), ale střední doby strávené v jednotlivých stavech zůstanou stejné.

3. Rovnice difuzního typu

Na časové škále $\mathbb{T}$ a s prostorem $X = \mathbf{Z}$ nyní uvažujeme soustavu rovnic

$$u_x^{\Delta_t} = a u_{x-1} + b u_x + c u_{x+1}, \quad x \in X, \tag{3}$$

kde předpokládáme, že $a, c \geqq 0$, $a + c > 0$, a také že $-b\mu_t \leqq 1$ pro všechna $t \in \mathbb{T}$. Tato rovnice zahrnuje jako speciální případ jak transportní rovnici ($a = 0$, $b = -c$), tak difuzní rovnici ($a = c = 1$, $b = -2$).

Nejprve spočteme „pravděpodobnostní“ případ $a + b + c = 0$, dále pak obecný případ, kdy tato rovnost platit nemusí. Výsledky v obecném případě lze dostat převodem z případu pravděpodobnostního. Tento převod však nakonec ukážeme níže z důvodu složitosti zápisu pouze pro škálu $\mathbb{T} = \{0, 1, 2, \dots\}$ a $\mathbb{T} = \langle 0, \infty)$.

Tvrzení 3.1. *Nechť $a + b + c = 0$ a $a = c$. Pak řešení u rovnice* (3) *při libovolné časové škále má (za předpokladů z konce úvodní kapitoly) vlastnosti* (i)–(iv) *až na časový integrál, který je vždy nekonečný.*

Nechť $a+b+c=0$ a $a>c$. Pak řešení u má navíc i konečné časové integrály, jejich hodnota je $-b/(a-c)$ pro $x \geqq 0$ a $-b(c/a)^x/(a-c)$ pro $x<0$. Analogicky pro $a<c$.

Pro škálu $\mathbb{T}=\{0,1,2,\dots\}$ (zde s $-b \leqq 1$), resp. $\mathbb{T}=\langle 0,\infty)$ mají rovnice tvar

$$t \text{ diskrétní:} \qquad u_x(t+1)=au_{x-1}(t)+(1+b)u_x(t)+cu_{x+1}(t), \tag{4}$$

$$t \text{ spojitý:} \qquad u'_x(t)=au_{x-1}(t)+bu_x(t)+cu_{x+1}(t). \tag{5}$$

Za předpokladu $a+b+c=0$ tedy rovnice pro pravděpodobnosti markovského řetězce s maticí pravděpodobností resp. s maticí intenzit přechodu

$$P=\begin{pmatrix} \ddots & \ddots & \ddots & & \\ & c & 1+b & a & \\ & & c & 1+b & a \\ & & & \ddots & \ddots & \ddots \end{pmatrix} \qquad \text{resp.} \qquad Q=\begin{pmatrix} \ddots & \ddots & \ddots & & \\ & c & b & a & \\ & & c & b & a \\ & & & \ddots & \ddots & \ddots \end{pmatrix}.$$

Jde o náhodnou procházku s velikostí kroku $-1,0,1$ a příslušnými pravděpodobnostmi c, $1+b$, a, podobně ve spojitém čase. Z vlastností rozeberme více jen časové integrály. Vnořeným řetězcem $(X_t^*,\, t=0,1,\dots)$ těchto řetězců je náhodná procházka, která odpovídá diskrétnímu modelu se škálou $\mathbb{T}=\{0,1,2,\dots\}$ a upravenými koeficienty

$$a^*=\frac{a}{a+c}, \quad 1+b^*=0, \quad c^*=\frac{c}{a+c}.$$

O této procházce je ale vše známo:

- Je-li $a=c$, její stavy jsou trvalé nulové, neboli pro pravděpodobnosti u^* tohoto vnořeného řetězce $\sum_t u_x^*(t)=\infty$.
- Je-li $a>c$, jsou její stavy přechodné. Do stavů $x \geqq 0$ se procházka dostane s pravděpodobností 1 a $\sum_t u_x^*(t)=(a^*-c^*)^{-1}$. Do stavů $x<0$ se dostane s pravděpodobností $\left(\frac{c^*}{a^*}\right)^x$ a $\sum_t u_x^*(t)=\left(\frac{c^*}{a^*}\right)^x(a^*-c^*)^{-1}$.

Původní řetězec při návštěvě stavu v něm do jeho opuštění setrvá po dobu v průměru $\frac{1}{-b}$. Při $a=c$ zůstává časový integrál nekonečný, $\int u_x(t)\,\Delta t=\infty$. Při $a>c$ pak

$$\int_0^\infty u_x(t)\,\Delta t=\frac{1}{-b}\sum_t u_x^*(t)=\begin{cases} \dfrac{1}{a-c} & x \geqq 0, \\ \dfrac{(c/a)^x}{a-c} & x<0. \end{cases}$$

Podobně jako v případě transportní rovnice se nahlédne, že stejný výsledek platí i pro ostatní časové škály.

Nakonec se podíváme na rovnici difuzního typu (3), kde případně není splněna rovnost $a + b + c = 0$. Označme hodnotu, která ke splnění rovnosti chybí, jako $d = -(a + b + c)$. V diskrétním případě vzhledem ke znaménkům koeficentů a, c a vzhledem k $b \geqq -1$ bude $d < 1$.

K vyšetření vlastností můžeme použít souvislost řešení této rovnice s řešením $\widetilde{u}$ pravděpodobnostní rovnice se „znormovanými“ koeficienty $\widetilde{a}$, $\widetilde{b}$, $\widetilde{c}$, pro něž $\widetilde{a} + \widetilde{b} + \widetilde{c} = 0$. Tato souvislost je zřejmá přímo dosazením do rovnic.

Tvrzení 3.2. *Nechť $\mathbb{T} = \{0, 1, 2, \dots\}$, $b \geqq -1$ a u je řešení rovnice* (3)*. Pak $u(t) = (1 - d)^t\, \widetilde{u}(t)$, kde $\widetilde{u}$ je řešením pravděpodobnostní rovnice* (4) *s*

$$\widetilde{a} = \frac{a}{1 - d}, \quad 1 + \widetilde{b} = \frac{1 + b}{1 - d}, \quad \widetilde{c} = \frac{c}{1 - d}. \tag{6}$$

Nechť $\mathbb{T} = \langle 0, \infty)$ a u je řešení rovnice (3)*. Pak $u(t) = \mathrm{e}^{-dt}\, \widetilde{u}(t)$, kde $\widetilde{u}$ je řešením pravděpodobnostní rovnice* (5) *s*

$$\widetilde{a} = a, \quad \widetilde{c} = c, \quad \widetilde{b} = -(a + c). \tag{7}$$

V případě $d > 0$ $(a + (1 + b) + c < 1)$ můžeme $u_x(t)$ interpretovat jako pravděpodobnosti pro řetězec, jehož stavový prostor obsahuje oproti pravděpodobnostnímu případu navíc dodatečný absorpční stav, do nějž řetězec v každém čase t může přejít s pravděpodobností, resp. intenzitou d.

Tvrzení 3.3. *Za podmínek předchozího tvrzení je $0 \leqq u_x(t) \leqq (1-d)^t$, resp. e^{-dt}, a také $\sum_x u_x(t) = (1 - d)^t$, resp. e^{-dt}. Bez ohledu na časovou škálu jsou časové integrály konečné právě tehdy, když $\sqrt{4ac} < -b$.*

Až na časový integrál vlastnosti plynou přímo z vlastností pravděpodobnostního případu a vyjádření dle předchozího tvrzení. Pokud $d > 0$, dostáváme oproti (ii) a (iii) vylepšené vlastnosti, jelikož na pravé straně rovnosti/nerovnosti máme číslo menší než 1. V případě $d < 0$ jde při $t \to \infty$ pravá strana $(1 - d)^t$ resp. e^{-dt} do nekonečna.

Za účelem zjištění časových integrálů uvažujme řetězec $(\widetilde{X}_t,\ t \geqq 0)$, jehož pravděpodobnosti $\widetilde{u}_x(t)$ nechť splňují pravděpodobnostní rovnici (4) resp. (5) s koeficienty $\widetilde{a}$, $\widetilde{b}$, $\widetilde{c}$ z (6) resp. (7). Pro řetězec $(\widetilde{X}_t)$ označme

$$\begin{aligned} T_n &= \text{okamžik } n\text{-té změny stavu; } T_0 = 0, \\ \tau_n &= \text{doba strávená ve stavu po } n\text{-té změně stavu, } T_n = \sum_{k=0}^{n-1} \tau_k. \end{aligned}$$

Vyšetřovaný časový integrál poskládáme z příspěvků jednotlivých přechodů vnořeného řetězce (X_n^*) (příslušného řetězci $(\widetilde{X}_t)$) a jeho pravděpodobností $u_x^*(n)$. Vyjádříme jej jako

$$\int_0^\infty u_x(t)\,\Delta t = \int D(0,t)\widetilde{u}_x(t)\,\Delta t = \mathrm{E}\int D(0,t) I_{\widetilde{X}_t=x}\,\Delta t$$
$$= \sum_n \underbrace{\mathrm{E}\Big(\int_{T_n}^{T_{n+1}} D(0,t)\,\Delta t\Big)}_{I_n} \cdot u_x^*(n) = \sum_n I_n u_x^*(n) \qquad (8)$$

kde $D(t_1,t_2) = (1-d)^{t_2-t_1}$ resp. $\mathrm{e}^{-d(t_2-t_1)}$ označuje „diskontní faktor“ a I_n je střední diskontovaná doba strávená řetězcem $(\widetilde{X}_t)$ po n-té změně stavu do příští změny. Příspěvek n-tého přechodu můžeme zapsat jako

$$I_n = \mathrm{E}\int_{T_n}^{T_{n+1}} D(0,t)\,\Delta t = \underbrace{\mathrm{E}\,D(0,T_n)}_{D_n}\,\underbrace{\mathrm{E}\Big(\int_0^{\tau_n} D(T_n,t)\,\Delta t \mid T_n\Big)}_{e_n}$$

kde D_n je střední diskontní faktor do n-té změny stavu a e_n je střední diskontovaná doba do opuštění stavu (diskontovaná k okamžiku vstupu do stavu). Vzhledem k nezávislosti a stejnému rozdělení přírůstků procesu (procházky) platí $D_n = (D_1)^n$ a $e_n = e_1$. V diskrétním resp. spojitém případě vychází při $-b > 0$

$$D_1 = \mathrm{E}(1-d)^{T_1} = \sum_1^\infty (1-d)^k(-\widetilde{b})(1+\widetilde{b})^{k-1} = \frac{-b-d}{-b} = \frac{a+c}{-b},$$
$$D_1 = \mathrm{E}\,\mathrm{e}^{-dT_1} = \int_0^\infty \mathrm{e}^{-dt}\cdot(-\widetilde{b})\,\mathrm{e}^{\widetilde{b}t}\,\mathrm{d}t = \frac{-\widetilde{b}}{-\widetilde{b}+d} = \frac{a+c}{-b}.$$

Jelikož $e_1 = \mathrm{E}\sum_0^{\tau_0-1}(1-d)^k$ resp. $\mathrm{E}\int_0^{\tau_0}\mathrm{e}^{-dt}\,\mathrm{d}t$, máme hned také $e_1 = \frac{1}{d}(1-D_1) = \frac{1}{-b}$. Při $-b \leqq 0$ jsou $I_n = \infty$.

Vidíme, že nezávisle na časové škále (v prezentované ukázce pro škály $\mathbb{T} = \{0,1,2,\dots\}$ i $\mathbb{T} = \langle 0,\infty)$) vycházejí stejné hodnoty. V řadě pro časový integrál (8) je n-tý člen při $-b > 0$ a $n \to \infty$

$$I_n\,u_x^*(n) = e_1\,u_x^*(n)\,D_1^{n-1} = \frac{1}{-b}\cdot\binom{n}{\frac{n+x}{2}}(a^*)^{\frac{n+x}{2}}(c^*)^{\frac{n-x}{2}}\cdot\left(\frac{a+c}{-b}\right)^n$$
$$\asymp \text{konst}\cdot\frac{1}{\sqrt{n}}(\sqrt{4a^*c^*})^n\left(\frac{-b-d}{-b}\right)^n = \text{konst}\cdot\frac{1}{\sqrt{n}}\left(\frac{\sqrt{4ac}}{-b}\right)^n,$$

součet řady je tak konečný, právě když $\sqrt{4ac} < -b$.

4. Shrnutí

Ukázali jsem, že při zkoumání otázek o vlastnostech řešení dynamické rovnice difuzního typu na dané časové škále lze dobře použít známé výsledky z markovských řetězců. V pravděpodobnostním případě vlastnosti vyplynou zcela automaticky, pomocí transformace z tvrzení 3.2 dostaneme vlastnosti i v nepravděpodobnostním případě. Pravděpodobnostním přístupem lze také jednoduše zdůvodnit, že časové integrály nezávisí na konkrétní časové škále. Naproti tomu je třeba uvést, že některé vlastnosti platí i v případech, kdy a nebo c je záporné, a případně i pro $t < t_0$, kterými jsme se nezabývali.

Analogicky lze postupovat v případě obecnějších lineárních rovnic, kdy rovnici interpretujeme jako rovnici pro pravděpodobnosti přechodu náhodné procházky s více možnostmi kroků než $-1, 0, 1$. Podobně i v případě vícerozměrných stavových prostorů (místo $X = \mathbf{Z}$ mít $\mathbf{Z}^2$, $\mathbf{Z}^3$).

Uvedený postup nelze přímočaře aplikovat v nelineárním případě, pro rovnice $u^{\Delta_t}(t) = A\,u(t)$ s nelineárním operátorem A. Jejich řešení by sice šlo interpretovat jako pravděpodobnosti markovského řetězce, avšak nehomogenního a navíc jiného pro každou počáteční podmínku.

Literatura

[1] Bohner M., Peterson A.: *Dynamic equations on time scales. An introduction with applications.* Birkhäuser, Boston, 2001.

[2] Friesl M., Slavík A., Stehlík P.: Partial dynamic equations on time scales and applications to discrete-state stochastic processes. *Appl. Math. Lett.* **37**, 86–90, 2014.

[3] Hilger S.: Analysis on measure chains—a unified approach to continuous and discrete calculus. *Results Math.* **18** (1–2), 18–56, 1990.

[4] Hoffacker J.: Basic partial dynamic equations on time scales. *J. Difference Equ. Appl.* **8** (4), 307–319, 2002.

[5] Jackson B.: Partial dynamic equations on time scales. *J. Comput. Appl. Math.* **186** (2), 391–415, 2006.

[6] Slavík A., Stehlík P.: Dynamic diffusion-type equations on discrete-space domains. Submitted. URL: `http://www.karlin.mff.cuni.cz/~slavik/publications.html`.

[7] Slavík A., Stehlík P.: Explicit solutions to dynamic diffusion-type equations and their time integrals. *Appl. Math. Comput.* **234**, 486–505, 2014.

[8] Stehlík P., Volek J.: Transport equation on semidiscrete domains and Poisson-Bernoulli processes. *J. Difference Equ. Appl.* **19** (3), 439–456, 2013.

METODA HLAVNÍCH KOMPONENT APLIKOVANÁ NA ROZŠÍŘENÝ QUERMASS-INTERAKČNÍ PROCES

PRINCIPAL COMPONENTS METHOD APPLIED IN THE EXTENDED QUERMASS-INTERACTION PROCESS

Kateřina Helisová[1], **Jakub Staněk**[2]

Adresa: [1]FEL ČVUT v Praze, Technická 2, 166 27 Praha 6,
[2]MFF UK v Praze, Sokolovská 83, 186 75 Praha 8

E-mail: [1]helisova@math.feld.cvut.cz, [2]stanekj@karlin.mff.cuni.cz

Abstrakt: Příspěvek se zabývá redukcí dimenze v rozšířeném Quermass--interakčním procesu pomocí metody hlavních komponent. Cílem této redukce je zvýšení efektivity odhadů parametrů tohoto modelu. Je zde uveden jak teoretický popis, tak aplikace na simulovaná a reálná data. Tento článek je rešerší článku [7].

Klíčová slova: Hlavní komponenty, MCMC maximální věrohodnost, redukce dimenze, Quermass-interakční proces.

Abstract: The contribution concerns dimension reduction in extended Quermass-interaction process using principal components method in order to make estimating parameters of QI process more effective. Theoretical description as well as the application to the both simulated and real data are shown. The paper is a research of [7].

Keywords: Dimension reduction, MCMC maximum likelihood, principal components, Quermass-interaction process.

1. Úvod

Spousta objektů studovaných v biologii, medicíně nebo materiálních vědách tvoří prostorové formace náhodného tvaru, v nichž můžeme pozorovat vzájemné interakce mezi těmito objekty. K analýze takových shluků používáme metody prostorové statistiky. Nedávno byl studován jeden z modelů pro tyto shluky, tzv. rozšířený Quermass-intrakční proces, který byl teoreticky popsán, nasimulován (viz [3]) a následně statisticky analyzován pomocí maximální věrohodnosti s využitím MCMC simulací (viz [4]). Nicméně tyto analýzy s sebou nesou také spoustu komplikací. Tou první je časová náročnost, tou druhou skutečnost, že v některých speciálních případech mohou být odhady podhodnocovány. V tomto článku ukážeme, jak je možné tyto problémy řešit

pomocí redukce dimenze, přičemž aplikovaná bude metoda hlavních komponent.

2. Model a jeho dřívější analýzy

2.1. Definice modelu

Mějme náhodnou množinu $\mathbf{X} \subset \mathbb{R}^2$ danou sjednocením kruhů se vzájemnými interakcemi, středy náhodně rozprostřenými v omezené množině $S \subset \mathbb{R}^2$ a náhodnými poloměry. Pro každou konečnou konfiguraci $\mathbf{x} = (x_1, \ldots, x_n)$ kruhů $x_1, \ldots, x_n$ je množina $\mathbf{X}$ popsaná hustotou $f_\theta(\mathbf{x})$ vzhledem k pravděpodobnostní míře tzv. Booleovského modelu, tj. procesu kruhů bez jakýchkoliv interakcí, viz [8]. Předpokládejme, že hustota je tvaru

$$f_\theta(\mathbf{x}) = c_\theta^{-1} \exp\{\theta \cdot T(U_\mathbf{x})\}, \tag{1}$$

kde $T(U_\mathbf{x})$ je m-dimenzionální vektor geometrických charakteristik sjednocení $U_\mathbf{x}$ kruhů z konfigurace $\mathbf{x}$, $\theta = (\theta_1, \ldots, \theta_m)$ je vektor parametrů, $\cdot$ značí skalární součin a c_θ je normující konstanta.

V publikaci [1] je Quermass-interakční proces definovaný jako proces s hustotou (1), v níž $T(U_\mathbf{x}) = (A(U_\mathbf{x}), L(U_\mathbf{x}), \chi(U_\mathbf{x}))$, kde A je plocha, L obvod a χ Euler-Poincarého charakteristika ($\chi = N_{cc} - N_h$, počet spojitých komponent mínus počet děr).

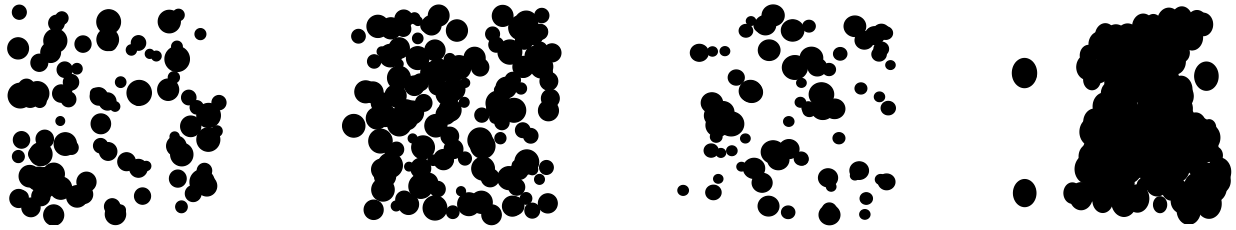
V publikaci [3] autoři rozšířili Quermass-interakční proces položením $T(U_\mathbf{x}) = (A(U_\mathbf{x}), L(U_\mathbf{x}), \chi(U_\mathbf{x}), N_h(U_\mathbf{x}), N_{id}(U_\mathbf{x}), N_{bv}(U_\mathbf{x}))$, kde N_{id} je počet izolovaných kruhů a N_{bv} počet vnějších vrcholů (tj. bodů na hranici $U_\mathbf{x}$, v němž se protínají dva kruhy). Jak dále zmiňují v publikaci [4], data většinou bývají v digitální podobě, v níž je těžko rozeznat hlavně vnější vrcholy. Proto zde budeme uvažovat hustotu s pěti charakteristikami ve tvaru

$$\begin{aligned} f_\theta(\mathbf{x}) \quad = \quad & c_\theta^{-1} \exp\{\theta_1 A(U_\mathbf{x}) + \theta_2 L(U_\mathbf{x}) + \theta_3 N_{cc}(U_\mathbf{x}) \\ & + \theta_4 N_h(U_\mathbf{x}) + \theta_5 N_{id}(U_\mathbf{x})\}, \end{aligned} \tag{2}$$

tedy $\theta = (\theta_1, \ldots, \theta_5)$ bude 5-dimenzionální parameter, který budeme odhadovat.

Interpretace parametrů je taková, že proces s kladným parametrem θ_j upřednostňuje konfigurace s větší j-tou geometrickou charakteristikou T_j, zatímco proces se záporným koeficientem produkuje spíše realizace s menší odpovídající charakteristikou. Např. na Obrázku 1, pozorujeme, že realizace procesu s $\theta_1 = 1$ a $\theta_1 = -1$ mají větší, resp. menší, plochu než vztažný Booleovský model. Interpretace procesu s více nenulovými parametry je stále

stejná, avšak hodně záleží na konkrétních hodnotách parametru, např. na Obrázku 1 pozorujeme realizaci procesu s $(\theta_1, \theta_2) = (5, -3)$, tj. s větší plochou a menším obvodem, tj. realizaci tvořící jeden velký shluk.



Obrázek 1: Realizace Booleovského modelu se středy v okně S velikosti 10×10 a poloměry s rovnoměrným rozdělením na $(0.2, 0.6)$ (1. obrázek), A-interakční proces s $\theta_1 = 1$ (2. obrázek), $\theta_1 = -1$ (3. obrázek) a (A, L)-interakční proces s $(\theta_1, \theta_2) = (5, -3)$ (4. obrázek).

2.2. MCMC maximální věrohodnost

Předpokládejme, že pozorujeme sjednocení $U_{\mathbf{x}}$. Metoda je založena na klasické maximalizaci logaritmicko-věrohodnostní funkce $l(\theta) = \log f_\theta(\mathbf{x})$, přičemž v našem případě dostáváme

$$l(\theta) = \log f_\theta(\mathbf{x}) = \theta_1 A(U_{\mathbf{x}}) + \ldots + \theta_5 N_{id}(U_{\mathbf{x}}) - \log c_\theta.$$

Problém však je, že c_θ nemá explicitní vyjádření, proto místo f_θ maximalizujeme f_θ / f_{θ^0} pro pevný vektor parametrů θ^0, neboť můžeme použít tzv. importance sampling (viz [2]) k aproximaci c_θ / c_{θ^0}, a to tak, že označíme-li

$$h_\theta(\mathbf{x}) = \exp\{\theta_1 A(U_{\mathbf{x}}) + \ldots + \theta_5 N_{id}(U_{\mathbf{x}})\},$$

pak

$$\begin{aligned} \log \frac{f_\theta}{f_{\theta^0}} &= l(\theta) - l(\theta^0) = \log(h_\theta(\mathbf{x})/h_{\theta^0}(\mathbf{x})) - \log(c_\theta/c_{\theta^0}) \qquad (3) \\ &\approx \log(h_\theta(\mathbf{x})/h_{\theta^0}(\mathbf{x})) - \log \frac{1}{R} \sum_{i=1}^{R} h_\theta(\mathbf{Z}_i)/h_{\theta^0}(\mathbf{Z}_i) =: l_{\theta^0}(\theta), \end{aligned}$$

kde $\mathbf{Z}_i$ jsou realizace z hustoty f_{θ^0} získané MCMC simulacemi a R je daný, dostatečně velký, počet těchto realizací. Cílem je tedy maximalizovat

$$\begin{aligned} l_{\theta^0}(\theta) =&(\theta_1 - \theta_1^0)A(U_{\mathbf{x}}) + \ldots + (\theta_5 - \theta_5^0)N_{id}(U_{\mathbf{x}}) \\ &- \log \frac{1}{R}\sum_{i=1}^{R} \exp\{(\theta_1 - \theta_1^0)A(U_{\mathbf{Z}_i}) + \ldots + (\theta_5 - \theta_5^0)N_{id}(U_{\mathbf{Z}_i})\}. \end{aligned}$$

Bohužel se objevují komplikace v případě, kdy některá z geometrických charakteristik v datech je příliš malá nebo příliš velká. Dá se dokázat (viz Tvrzení 1 v publikaci [7]), že když pro některou složku T_j vektoru geometrických charakteristik platí $T_j(U_{\mathbf{Z}_i}) \geq T_j(U_{\mathbf{x}})$ pro všechna $i \in \{1, \ldots, R\}$ a $T_j(U_{\mathbf{Z}_i}) > T_j(U_{\mathbf{x}})$ pro alespoň jedno i, pak $l_{\theta^0}(\theta)$ je klesající v j-té složce, a tedy maximálně věrohodný odhad je $\widehat{\theta}_j = -\infty$. To znamená, že pro data s uvažovanou charakterisikou na dolní hranici (např. data tvořená pouze jednou komponentou nebo data bez děr) MCMC MLE podhodnocuje odhady v odpovídajících složkách.

Další problém zmíněný již v publikaci [3] je, že MCMC MLE je velice časově náročná, a to ze dvou důvodů:

1. Počet R realizací $\mathbf{Z}_i$ potřebných k aproximaci podílu konstant $\mathbf{Z}_i$ je celkem velký.
2. Aproximaci (3) lze použít pouze pro θ^0 dostatečně blízko θ, takže je nutné použít metodu postupného odhadování, tzv. bridge sampling (viz [2]).

Tyto dva problémy lze částečně vyřešit pomocí redukce dimenze, která je popsaná v následujícící kapitole.

3. Metoda hlavních komponent

3.1. Popis metody

Uvažujme m-dimenzionální náhodný vektor $T(U_{\mathbf{X}})$ geometrických charakteristik množiny $U_{\mathbf{X}}$. Označme $\mathbf{V} = (\sigma^2_{i,j})_{i,j=1}^m$ kovarianční matici $T(U_{\mathbf{X}})$. Předpokládejme, že $\mathbf{V}$ má $k > 0$ vzájemně různých kladných vlastních čísel, označme je $\lambda_1 > \lambda_2 > \ldots > \lambda_k$ a příslušné vlastní vektory označme $\mathbf{v}_1, \ldots, \mathbf{v}_k$.

Hledáme-li $\mathbf{u}$ takové, že $\mathbf{u}^T\mathbf{u} = 1$, přičemž $\mathbf{u}T(U_{\mathbf{X}})$ má největší možný rozptyl, lze dokázat, že $\mathbf{u} = \mathbf{v}_1$ a navíc $\mathrm{var}(\mathbf{v}_1 T(U_{\mathbf{X}})) = \lambda_1$.

Označme $C_1(U_{\mathbf{X}}) = \mathbf{v}_1 T(U_{\mathbf{X}})$ a hledejme vektor $\mathbf{u}$ takový, že $\mathbf{u}^T\mathbf{u} = 1$, přičemž $\mathbf{u}T(U_{\mathbf{X}})$ má největší možný rozptyl za podmínky, že $\mathbf{u}T(U_{\mathbf{X}})$ není

korelováno s $C_1(U_{\mathbf{X}})$, tj.

$$\operatorname{cov}(\mathbf{u}T(U_{\mathbf{X}}), \mathbf{v}_1 T(U_{\mathbf{X}})) = \mathbf{u}^T \mathbf{V} \mathbf{v}_1 = \mathbf{u}^T \lambda_1 \mathbf{v}_1 = \lambda_1 \mathbf{u}^T \mathbf{v}_1 = 0.$$

To je splněno tehdy, když $\mathbf{u}^T \mathbf{v}_1 = 0$. Lze dokázat, že takový vektor je $\mathbf{u} = \mathbf{v}_2$ a navíc označíme-li $C_2(U_{\mathbf{X}}) = \mathbf{v}_2 T(U_{\mathbf{X}})$, dostáváme $\operatorname{var} C_2(U_{\mathbf{X}}) = \lambda_2$.

Tímto způsobem obdržíme k náhodných veličin $C_1(U_{\mathbf{X}}), \ldots, C_k(U_{\mathbf{X}})$, které se nazývají hlavními komponentami vektoru $T(U_{\mathbf{X}})$ a plně vysvětlují chování $T(U_{\mathbf{X}})$.

V praxi obvykle máme $k = m$, avšak pro nějaké p již vektory $C_{p+1}(U_{\mathbf{X}}), \ldots, C_m(U_{\mathbf{X}})$ nejsou významné. Přesněji označíme-li

$$\sigma^2 = \sum_{i=1}^{m} \sigma_{ii}$$

celkovou variabilitu $T(U_{\mathbf{x}})$, lze dokázat, že

$$\operatorname{var} C_1(U_{\mathbf{X}}) + \ldots + \operatorname{var} C_k(U_{\mathbf{X}}) = \lambda_1 + \ldots + \lambda_k = \sigma^2,$$

a tedy $C_1(U_{\mathbf{X}}), \ldots, C_p(U_{\mathbf{X}})$ takové, že relativní kumulativní variabilita

$$\mathrm{RVC}_p = \frac{\lambda_1 + \ldots + \lambda_p}{\sigma^2} \doteq 1,$$

pokrývá dostatečně variabilitu dat, tudíž může dostatečně vysvětlovat chování vektoru $T(U_{\mathbf{X}})$.

V našem případě to znamená, že hustota (1) může být přepsaná do tvaru

$$f_\varphi(\mathbf{x}) = c_\varphi^{-1} \exp\{\varphi \cdot C(U_{\mathbf{x}})\} = c_\varphi^{-1} \exp\{\varphi_1 C_1(U_{\mathbf{x}}) + \ldots + \varphi_p C_p(U_{\mathbf{x}})\}, \quad (4)$$

a jelikož vektor neznámých parametrů φ má nižší dimenzi než původní vektor θ, jeho odhad je rychlejší. Navíc jelikož $C_j(U_{\mathbf{x}})$ pro $j = 1, \ldots, p$ je lineární kombinací původních (pouze kladných) geometrických charakteristik $T_j(U_{\mathbf{x}})$ pro $j = 1, \ldots, m$, znamená to, že pokud je alespoň jeden koeficient lineární kombinace záporný, může být charakteristika $C_j(U_{\mathbf{x}})$ také záporná, což eliminuje patologickou situaci zmíněnou v Tvrzení 1 v publikaci [7] (viz výše), takže je tím vyřešen i problém podhodnocování parametrů.

V publikaci [6] jsou zmíněny čtyři způsoby, jak určit vhodné p. Zde použijeme ten nejjednodušší, a to vzít p takové, že relativní kumulativní variabilita (RVC_p) je větší než 80 %, tj. $p = \min\{\widetilde{p} : \mathrm{RVC}_{\widetilde{p}} > 0.8\}$.

3.2. Simulační studie

Metodu popsanou v sekci 3.1. jsme aplikovali na model s hustotou (2). Jelikož potřebujeme odhadnout kovarianční matici $\mathbf{V}$ vektoru $T(U_{\mathbf{X}})$, musíme mít na vstupu více realizací náhodné množiny. K tomuto účelu jsme nasimulovali N různých realizací procesu s hustotou (2) vzhledem k Booleovskému modelu s intenzitou středů $\rho = 1$ v okně S o velikosti 10×10, rozdělením poloměrů rovnoměrným na $[0.2, 0.7]$ a 5-dimenzionálním parametrem $\theta = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5) = (1.5, -1, 1, -0.25, -0.5)$ (ukázka takové realizace je na Obrázku 2). Metodu jsme zkoumali pro $N = 100$, resp. $N = 10$. Odhady lze nalézt v Tabulce 1.

Tabulka 1: Vlastní čísla, vlastní vektory a relativní kumulativní variabilita (významné hodnoty vyznačeny tučně) a příslušné maximálně věrohodné odhady parametrů $(\widehat{\varphi}_1, \widehat{\varphi}_2)$ pro $N = 100$ a $N = 10$.

N	Vl. čísla	Vlastní vektory	RVC	$(\widehat{\varphi}_1, \widehat{\varphi}_2)$
	117.40	**(−0.41,0.82,0.28,−0.25,0.11)**	49%	
	104.15	**(−0.77,−0.55,0.22,−0.23,0.05)**	93%	
100	8.35	(0.06,−0.02,0.64,0.63,0.44)	97%	(−1.08,−0.32)
	5.71	(0.43,−0.15,0.30,−0.69,0.47)	99%	
	1.57	(0.20,−0.05,0.61,−0.11,−0.75)	100%	
	98.53	**(0.61,−0.47,−0.52,0.29,−0.21)**	63%	
	46.70	**(0.61,0.78,0.01,0.09,0.11)**	93%	
10	7.16	(0.06,0.09,−0.43,−0.88,−0.19)	98%	(0.91,−0.15)
	2.58	(−0.49,0.39,−0.60,0.37,−0.34)	99%	
	0.78	(−0.10,−0.04,−0.43,0.02,0.89)	100%	

Dle výše uvedeného kritéria je zřejmé, že pro obě zkoumaná N jsou první dva vektory mnohem významnější než ty zbývající. Znamená to tedy, že data mohou být popsaná pouze dvěma charakteristikami místo původních pěti, přičemž tyto vysvětlující charakteristiky jsou dány lineární kombinací původních charakteristik $A(U_{\mathbf{x}})$, $L(U_{\mathbf{x}})$, $N_{cc}(U_{\mathbf{x}})$, $N_h(U_{\mathbf{x}})$, $N_{id}(U_{\mathbf{x}})$ s koeficienty lineární kombinace dané vlastními vektory. Např. při analýze s $N = 10$ to znamená, že označením

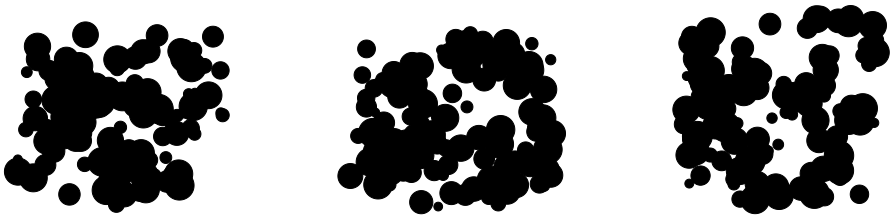
$$LC_1 = 0.61A(U_{\mathbf{x}}) - 0.47L(U_{\mathbf{x}}) - 0.52N_{cc}(U_{\mathbf{x}}) + 0.29N_h(U_{\mathbf{x}}) - 0.21N_{id}(U_{\mathbf{x}}),$$

$$LC_2 = 0.61A(U_{\mathbf{x}}) + 0.78L(U_{\mathbf{x}}) + 0.01N_{cc}(U_{\mathbf{x}}) + 0.09N_h(U_{\mathbf{x}}) + 0.11N_{id}(U_{\mathbf{x}}),$$

má nový model hustotu

$$f_\varphi(\mathbf{x}) = c_\varphi^{-1} \exp\{\varphi_1 LC_1 + \varphi_2 LC_2\}.$$

Dvoudimenzionální parametr $\varphi = (\varphi_1, \varphi_2)$ byl pak odhadnutý metodou MCMC MLE, přičemž za hodnoty charakteristik $A(U_\mathbf{x}), \ldots, N_{id}(U_\mathbf{x})$ jsme vzali průměry z N vstupních realizací. Odhady $\widehat{\varphi}_1$ a $\widehat{\varphi}_2$ jsou v posledním sloupci Tabulky 1 a příklady realizací z nafitovaného modelu na Obrázku 2.



Obrázek 2: Porovnání realizací modelu s původními parametry $(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5) = (1.5, -1, 1, -0.25, -0.5)$ (vlevo) s realizacemi nafitovaných modelů při použití $N = 100$ (uprostřed) a $N = 10$ (vpravo).

3.3. Kontrola modelu

Uvažujme náhodnou množinu $\mathbf{Z}$. Pro kruh $b(0, r)$ se středem v počátku a poloměrem r definujme $D = \inf\{r \geq 0 : \mathbf{Z} \cap b(0, r) \neq \emptyset\}$. Jestliže $P(D > 0) > 0$, pak sférická kontaktní distribuční funkce pro stacionární množinu $\mathbf{Z}$ je definovaná jako

$$H_B(r) = P(D \leq r | D > 0).$$

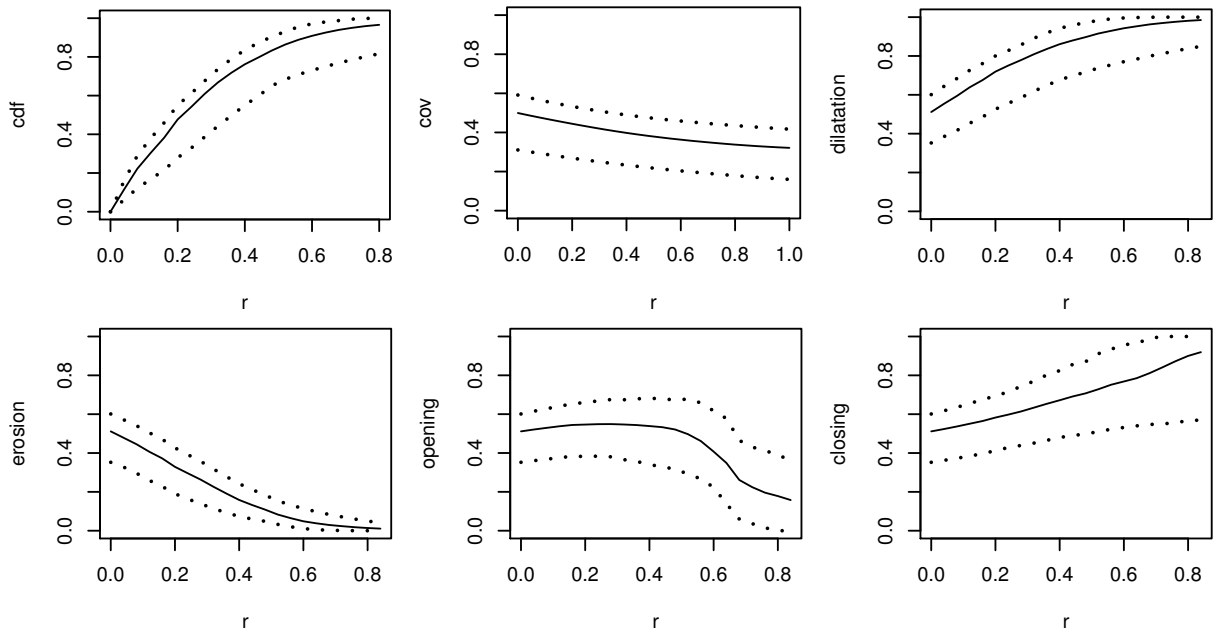
Dále uvažujme body u a v takové, že $\|u - v\| = r$, pak kovarianční funkce stacionární a izotropní množiny $\mathbf{Z}$ je definovaná jako

$$C(r) = P(u \in \mathbf{Z}, v \in \mathbf{Z}).$$

Obě tyto funkce mohou být z dat odhadnuty použitím vhodné diskretizace, viz [8].

Dále nechť $|\mathbf{Z}| = A(\mathbf{Z})$ a pro všechna $r > 0$ nechť $\mathbf{Z}_{\oplus r} = \cup_{u \in \mathbf{Z}} b(u, r)$ a $\mathbf{Z}_{\ominus r} = \{u \in \mathbb{R}^2 : b(u, r) \subseteq \mathbf{Z}\}$. Dilatace d, eroze e, otevření o a zavření c množiny $\mathbf{Z}$ použitím kruhu $b(0, r)$ je definováno (viz [8]) jako

$$d(r) = \frac{|\mathbf{Z}_{\oplus r} \cap W_{\ominus r}|}{|W_{\ominus r}|}, \quad e(r) = \frac{|\mathbf{Z}_{\ominus r}|}{|W_{\ominus r}|}, \tag{5}$$



Obrázek 3: Grafy sférické kontaktní distribuční funkce, kovarianční funkce, dilatace, eroze, otevření a zavření množiny zprůměrované z $N = 10$ realizací vstupních dat (plná čára) porovnané s 95% obálkami získanými z 39 realizací nafitovaného modelu (tečkované čáry).

$$o(r) = \frac{|(\mathbf{Z}_{\ominus r})_{\oplus r} \cap W_{\ominus 2r}|}{|W_{\ominus 2r}|}, \quad c(r) = \frac{|(\mathbf{Z}_{\oplus r})_{\ominus r} \cap W_{\ominus 2r}|}{|W_{\ominus 2r}|}.$$

Obrázek 3 srovnává těchto šest konktrolních funkcí s nasimulovanými 95% obálkami získanými z 39 realizací nafitovaného modelu (viz [2]) pro $N = 10$. Je zde vidět, že grafy pro data leží uvniř obálek, takže závěr je, že zredukovaný model popisuje data dostatečně dobře, a to i při relativně malém počtu vstupních realizací.

3.4. Aplikace na reálná data

Metoda popsaná výše byla dále aplikovaná také na reálná data, jimiž je $N = 10$ obrázků prsní tkáně postižené invazivním duktálním karcinomem (viz Obrázek 4). Data byla poskytnuta autory publikace [5], v níž jsou detailněji popsána.

Výsledky redukce dimenze pro tato data lze nalézt v Tabulce 2, kontrolu modelu pak na Obrázku 5. Z výsledků můžeme usoudit, že redukovaný model dobře popisuje i tato reálná data.

Tabulka 2: Vlastní čísla, vlastní vektory a relativní kumulativní variabilita (významné hodnoty vyznačeny tučně) a příslušný maximálně věrohodný odhad parametru $\widehat{\varphi}$.

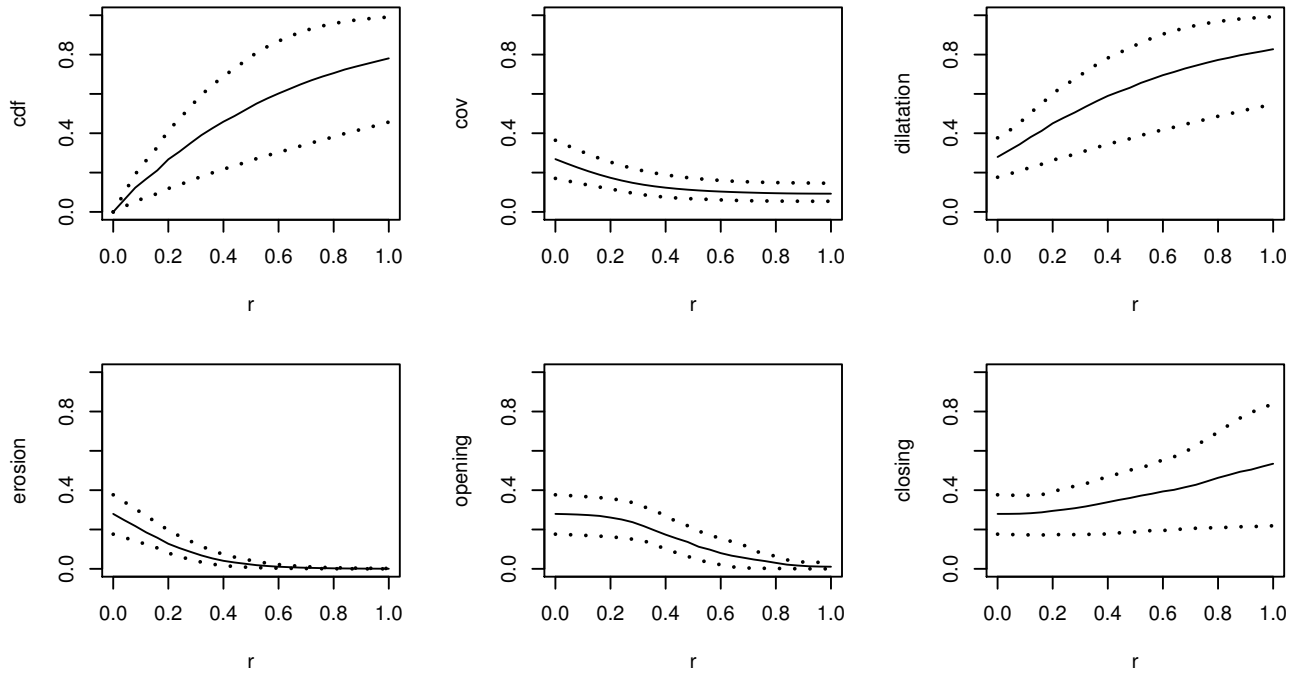
Vlastní čísla	Vlastní vektory	RVC	$\widehat{\varphi}$
2714.25	**(−0.09,−0.88,−0.44,−0.02,−0.12)**	94.00%	0.64
156.98	(0.30,0.42,−0.82,−0.11,−0.23)	99.00%	
23.84	(0.00, −0.01,−0.20,−0.43,0.88)	99.80%	
4.19	(−0.24,0.03,0.15,−0.88,−0.39)	99.99%	
0.88	(0.92,−0.22,0.26,−0.20,−0.04)	100.00%	

Obrázek 4: Reálná data srovnána s nafitovaným modelem.

4. Závěr

Na začatku jsme si dali dva úkoly: zrychlit metodu MCMC MLE a vyřešit problém s podhodnocováním těchto odhadů v případě malých hodnot geometrických charakteristik vstupních dat. Prezentovaná metoda hlavních komponent řeší obojí. Odhad pětidimenzionálního parametru θ, tj. hledání maxima v pětidimenzionálním prostoru, se zredukoval na odhad parametru v mnohem nižší dimenzi, čímž se odhadovací procedura významně urychlila. Navíc složky nových vstupních charakteristik jsou dané lineární kombinací původních pěti charakteristik, takže v případě jediného záporného koeficientu lineární kombinace mohou být nové vstupní charakteristiky záporné, a tudíž nedochází k situaci způsobující podhodnocování parametrů. Stručná simulační studie i aplikace na reálná data ukázaly, že redukovaný model popisuje data dostatečně dobře, a tudíž je zmíněná metoda velice užitečná při aplikaci rozšířeného Quermass-interakčního procesu.

Více detailů a širší simulační studii lze nalézt v publikaci [7].



Obrázek 5: Grafy sférické kontaktní distribuční funkce, kovarianční funkce, dilatace, eroze, otevření a zavření množiny zprůměrované z $N = 10$ realizací vstupních dat (plná čára) porovnané s 95% obálkami získanými z 39 realizací nafitovaného modelu (tečkované čáry).

Poděkování

Výzkum byl podpořen granty GA ČR P201/10/0472 a GA ČR 13-05466P.

Literatura

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

[2] Møller J., Waagepetersen R.: *Statistical inference and simulations for spatial point processes*. Boca Raton, Chapman and Hall/CRC, 2004.

[3] Møller J., Helisová K.: Power diagrams and interaction process for unions of discs. *Advances in Applied Probability* **40**, 321–347, 2008.

[4] Møller J., Helisová K.: Likelihood inference for unions of interacting discs. *Scandinavian Journal of Statistics* **37**, 365–381, 2010.

[5] Mrkvička T., Mattfeldt T.: Testing histological images of mammary tissues on compatibility with the Boolean model of random sets. *Image Analysis and Stereology* **30**, 11–18, 2011.

[6] Rencher A. C.: *Methods of Multivariate Analysis*, 2. vyd. New York, Wiley & Sons, 2002.

[7] Staňková Helisová K., Staněk J.: Dimension reduction in extended Quermass-interaction process. *Methodology and Computing in Applied Probability* **16**, 355–368, 2014.

[8] Stoyan D., Kendall W. S., Mecke J.: *Stochastic Geometry and Its Applications*. Chichester, Wiley & Sons, 1995.

LIMITNÍ VĚTY PRO SLABĚ ZÁVISLÁ NÁHODNÁ POLE

LIMIT THEOREMS FOR WEAKLY DEPENDENT RANDOM FIELDS

Jana Klicnarová

Adresa: Ekonomická fakulta, Studentská 13, 370 05, České Budějovice

E-mail: klicnarova@ef.jcu.cz

Abstrakt: Existuje mnoho výsledků na téma limitních vět pro slabě závislá náhodná pole. My se v tomto příspěvku zaměříme na problematiku martingalových aproximací pro náhodná pole a na výsledky využívající technik aproximací m-závislými poli. Ukážeme také výsledky při sčítání přes obecné množiny.

Klíčová slova: Limitní věty, slabá závislost, náhodná pole.

Abstract: There are many results on the topic of limit theorems for weakly dependent random fields. We are interested in the methods based on a martingale approximation and an approximation by m-dependent random fields. We will also present results for weakly dependent random fields in case of summation over general sets.

Keywords: Limit theorems, weak dependence, random fields.

1. Introduction

The aim of this paper is to present some results on limit theorems for weakly dependent random fields. At first, we briefly introduce results for processes and then we discuss the generalization of these results for random fields.

At the beginning of our paper, let us introduce the used notation. We consider a probability space $(\Omega, \mathcal{A}, \mu)$ which is equipped with a bimeasurable and measure preserving transformation $T : \Omega \to \Omega$.

We will suppose $f : \Omega \to \mathbb{R}$ to be a regular function with zero mean and finite second moment. We denote $X_i = f \circ T^i$ so then $(X_i)_{i\in\mathbb{Z}}$ is a strictly stationary process. It is well-known that every stationary process can be presented in this way. Then, U is a unitary operator on L_2 such that $Uf = f \circ T$. (Recall that the space L_2 is a space of all square integrable functions.) By $\mathcal{F}_0$, we denote a σ-field $\mathcal{F}_0 \subset \mathcal{A}$ such that $\mathcal{F}_0 \subset T^{-1}(\mathcal{F}_0)$ and by $\mathcal{F}_i$ we denote $T^{-i}(\mathcal{F}_0)$. Hence, we have a filtration $(\mathcal{F}_i)_{i\in\mathbb{Z}}$ such that $\mathcal{F}_i \subset \mathcal{F}_{i+1}$. Further, we can denote $\mathcal{F}_\infty = \bigvee_{i\in\mathbb{Z}} \mathcal{F}_i$ and $\mathcal{F}_{-\infty} = \bigcap_{i\in\mathbb{Z}} \mathcal{F}_i$. We suppose f to be regular, it means that $f \in L_2(\mathcal{F}_\infty)$ and $\mathrm{E}(f|\mathcal{F}_{-\infty}) = 0$ a.s.

Through the whole paper, we use projection criteria. To define the projection, we need first to recall the following notation. Let H_i be a Hilbert space, $H_i \subset L_2(\mu)$, such that it contains all functions from the space $L_2(\mathcal{F}_i) \ominus L_2(\mathcal{F}_{i-1})$. We also need to define a projection operator $P_i(X)$ of the function X onto the space H_i, respectively $P_i(X) = \mathrm{E}(X|\mathcal{F}_i) - \mathrm{E}(X|\mathcal{F}_{i-1})$. It is easily seen that for any regular function f we have $f = \sum_{i=-\infty}^{\infty} P_i f$.

In this paper, we are interested in limit theorems. It means, that we focus on the limit behaviour of the term $S_n(f) = \sum_{i=1}^{n} f \circ T^i$, or more generally $S_{\Gamma_n}(f) = \sum_{i \in \Gamma_n} f \circ T^i$ normalized by any term. More precisely, we study a convergence in distribution of these normalized terms.

2. Limit Theorems for weakly dependent processes

There are many results on limit theorems for weakly dependent processes. There are three main approaches to this problem. First one, and probably the most well-known, is the approach based on so called mixing conditions. Most results on this topic can be found in [2], but there are also recent results, too. See, for example, Tone (2010).

The second approach uses so called associated random variables, see, for example, [3] and others.

We are interested in the third possibility, which are approaches based on martingale approximations. The idea of this methods comes from the paper given by Gordin in 1969, see [4].

Let us recall the definition of the martingale approximation. We say that a process $(f \circ T^i)_i$ has a martingale approximation with respect to a filtration $(\mathcal{F}_i)_{i \in \mathbb{Z}}$, if there exists a martingale difference sequence $(m \circ T^i)_{i \in \mathbb{Z}}$ such that $P_0 m = m$ and $||S_n(f-m)||_2^2 = o(n)$.

Since the time, there are many results which use this approximation, we can mention, for example, the central limit theorem given by Hannan in 1979, [6]. Hannan proved the central limit theorem under his condition and some more extra conditions. Hannan's result was later extended by Dedecker, Merlevéde and Volný (2007), who proved the central limit theorem under only Hannnan's condition and also the invariance principle under this condition. For completeness, let us recall Hannan's condition:

$$\sum_{i=1}^{\infty} ||P_0(f \circ T^i)||_2 < \infty. \tag{1}$$

Another interesting result was given by Maxwell and Woodroofe, see [8]. They proved the central limit theorem under the following condition

$$\sum_{k=1}^{\infty} \frac{||\mathrm{E}(S_k(f)|\mathcal{F}_0)||_2}{k^{3/2}} < \infty.$$

Later, Peligrad and Utev (2005) proved the invariance principle under Maxwell-Woodroofe's condition.

3. Random fields

Now, there is the question, if it is possible to extend the results which are known for weakly dependent processes to similar results for weakly dependent random fields. There are well-known results based on mixing conditions for random fields (Bolthausen (1982), Goldie and Morrow (1986), Bradley (1989) and many others). Results based on associated random variables and on martingale approximations are extended to multi-dimension version, too. At this paper, we are interested in the last type of the problems. So, our aim is to study the results based on approximation methods. Sure, it is an old problem, how to extend approximation results for processes to results for random fields. The reason why the generalization is not direct is the problem with a definition of a martingale structure in the multi-dimension case. The definition of a martingale in the multi-dimension case depends on the ordering which we choose on the space $\mathbb{Z}^d$. There are several possibilities of defining the martingale. Some of results on limit theorems for random fields based on a martingale approximation, we can find in [1], [9], [10] and others.

We will be interested in the results which can be obtained by using technics of approximation by m-dependent random fields. At first, let us recall the notation which we use in the case of random fields.

Let d be the dimension of our index space $\mathbb{Z}^d$. Then by $\mathbf{i}$ we denote an arbitrary element of $\mathbb{Z}^d$, more precisely $\mathbf{i}$ is the point from $\mathbb{Z}^d$ with coordinates $(i_1, i_2, \ldots, i_d)$. So, by $\mathbf{0}$ we also denote $(0, 0, \ldots, 0)$ point in Z^d and so on.

Let us consider a probability space $(\mathbb{R}^{\mathbb{Z}^d}, \mathcal{B}^{\mathbb{Z}^d}, \mathrm{P}^{\mathbb{Z}^d})$ and write $\epsilon_{\mathbf{k}}(\omega) = \omega_{\mathbf{k}}$. Then, the σ-field $\mathcal{F}_{\mathbf{k}}$ is defined as $\sigma\{\epsilon_{\mathbf{l}} : \mathbf{l} \leq \mathbf{k}, \mathbf{l} \in \mathbb{Z}^d\}$, for all $\mathbf{k} \in \mathbb{Z}^d$ (we use $\mathbf{k} \leq \mathbf{l}$ in case that each coordinate of $\mathbf{k}$ is less or equal to a corresponding coordinate of $\mathbf{l}$). Then we use again a transformation T on $\mathbb{R}^d$: $(T^{\mathbf{k}}\omega)_{\mathbf{l}} = \omega_{\mathbf{k}+\mathbf{l}}$, then $(f \circ T^{\mathbf{k}})_{\mathbf{k}\in\mathbb{Z}^d}$ is a stationary random field. We still suppose f to be regular, $f \in L_2$ and f to have a zero mean.

Now, let us mention some recent results. Wang and Woodroofe, see [13], proved the invariance principle based on Maxwell-Woodroofe's type of con-

dition. They denote $V_n = \Pi_{i=1}^d \{1, \ldots, m_i^{(n)}\} \subset \mathbb{N}^d$, where $m_i^{(n)} \to \infty$. Futher

$$W_{d,p}(f) = \sum_{\mathbf{k} \in \mathbb{N}^d} \frac{||\mathrm{E}(f \circ T^{\mathbf{k}} | \mathcal{F}_{\mathbf{1}})||_p}{\Pi_{i=1}^d \sqrt{k_i}},$$

$$B_{n,t}(f) = \sum_{\mathbf{k} \in \mathbb{N}^d} \lambda(V_n(t) \cap R_{\mathbf{k}}) f \circ T^{\mathbf{k}}$$

and $S_n(f) = \sum_{\mathbf{k} \in V_n} f \circ T^{\mathbf{k}}$.

The result given by Wang and Woodroofe, see [13], is as follows.

Theorem 3.1 (Wang, Woodroofe [13]). *Let us have a probability space $(\mathbb{R}^{\mathbb{Z}^d}, \mathcal{B}^{\mathbb{Z}^d}, P^{\mathbb{Z}^d})$ with a filtration $(\mathcal{F}_{\mathbf{k}})_{\mathbf{k} \in \mathbb{N}^d}$. If f has zero mean and finite second moment, $f \in \mathcal{F}_{\mathbf{0}}$ and $W_{d,2}(f) < \infty$,*

then

$$\sigma^2 = \lim_{n \to \infty} \frac{||S_n||^2}{|V_n|} < \infty$$

and

$$\frac{S_n}{\sqrt{|V_n|}} \Rightarrow N(0, \sigma^2).$$

Moreoever, if $f \in L_0^p$ and $W_{d,p}(f) < \infty$ for some $p > 2$, then

$$\frac{B_{n,\cdot}(f)}{\sqrt{|V_n|}} \Rightarrow \sigma \mathbb{B}(\cdot)$$

in a space $C([0,1]^d)$.

Later, Volný and Wang, see [12], used a different approach, they used Hannan's condition and proved the following result.

Theorem 3.2 (Volný, Wang [12]). *Under Hannan's condition:*

$$\sum_{\mathbf{i} \in \mathbb{Z}^d} ||P_{\mathbf{0}} X_{\mathbf{i}}||_2 < \infty,$$

we have

$$\left\{ \frac{S_{\lfloor nt \rfloor}}{\sqrt{n}} \right\}_{t \in [0,1]^d} \Rightarrow \{\mathbb{B}_t\}_{t \in [0,1]^d}$$

as $n \to \infty$ in $D([0,1]^d)$.

Both of the above mentioned results are limit theorems with the summation over rectangles. Now, let us mention some results for more general case – the summation over more general sets – let us denote them $(\Gamma_n)_{n\in\mathbb{N}}$.

Some of the results for general sets can be found in the paper given by El Machkouri, Volný, Wu, see [5]. Their results are formulated under the condition of so called p-stability. The definition of p-stability is based on a finiteness of a physical dependence measure.

Let us briefly recall the idea of the physical dependence measure. We put $X_{\mathbf{i}} = g(\varepsilon_{\mathbf{i}-\mathbf{j}}; \mathbf{j} \in \mathbb{Z}^d),\ \mathbf{i} \in \mathbb{Z}^d$, where $(\varepsilon_{\mathbf{i}})_{\mathbf{i}}$ are i.i.d., and $(\varepsilon'_{\mathbf{i}})_{\mathbf{i}}$ are i.i.d. copies of $(\varepsilon_{\mathbf{i}})_{\mathbf{i}}$. By $X^*_{\mathbf{i}}$ we denote a version of $X_{\mathbf{i}}$ such that $X^*_i = g(\varepsilon^*_{\mathbf{i}-\mathbf{j}}; \mathbf{j} \in \mathbb{Z}^d),\ \mathbf{i} \in \mathbb{Z}^d$, $\varepsilon^*_{\mathbf{i}} = \varepsilon_{\mathbf{i}}$ for all $\mathbf{i} \neq \mathbf{0}$ and $\varepsilon^*_{\mathbf{0}} = \varepsilon'_{\mathbf{0}}$.

Further, we can define $\delta_{i,p} = ||X_{\mathbf{i}} - X^*_{\mathbf{i}}||_p$ and $\Delta_p = \sum_{\mathbf{i}\in\mathbb{Z}^d} \delta_{\mathbf{i},p}$. Hence, we say that the process is p-stable if $\Delta_p < \infty$.

Theorem 3.3 (El Machkouri, Volný, Wu [5])**.** *Let $(X_{\mathbf{i}})$ be a stationary random field with $\Delta_2 < \infty$ and a $\sigma_n^2 = ||S_{\Gamma_n}||^2 \to \infty$. If $(\Gamma_n)_{n\in\mathbb{N}}$ is a sequence of subsets of $\mathbb{Z}^d$ such that $|\Gamma_n| \to \infty$, then Lévy distance:*

$$L\left(S_{\Gamma_n}/\sqrt{|\Gamma_n|}, N(0, \sigma_n^2/|\Gamma_n|)\right) \to 0$$

as $n \to \infty$, where $S_n(A) = \sum_{\mathbf{i}\in\{1,2,\ldots,n\}^d} \lambda(nA \cap R_{\mathbf{i}}) X_{\mathbf{i}}$ and $R_{\mathbf{i}} = (\mathbf{i} - \mathbf{1}, \mathbf{i}]$.

Remark. Under the condition that $|\partial\Gamma_n|/|\Gamma_n| \to 0$ (where $|\cdot|$ denotes the cardinality of the set and ∂ the boundary of the set) and $\sigma^2 = \sum\limits_{\mathbf{k}\in\mathbb{Z}^d} \mathrm{E}(X_{\mathbf{0}}, X_{\mathbf{k}}) > 0$ there follows:

$$\frac{S_{\Gamma_n}}{\sqrt{|\Gamma_n|}} \Rightarrow N(0, \sigma^2).$$

To introduce the invariance principle given by El Machkouri, Volný and Wu we need to generalize p-stability to ψ-stability and also to recall some definitions about entropy and VC-classes. For more details about entropy, see for example, [11].

At first, let us mention the definition of ψ-stability.

A function ψ is a **Young function** if it is a real convex nondecreasing function defined on $\mathbb{R}^+$ which satisfies

$$\begin{aligned} \lim_{t\to\infty} \psi(t) &= \infty \\ \psi(0) &= 0. \end{aligned}$$

The Orlicz space L_ψ is defined as a space of real random variables Z defined on a probability space $(\Omega, \mathcal{A}, P)$ such that

$$\mathrm{E}[\psi(|Z|/c)] < \infty$$

for some $c > 0$. For more details see, for example, [7].

The Orlicz space L_ψ is equipped with a Luxemburg norm $||\cdot||_\psi$ defined for a real random variable Z by

$$||Z||_\psi = \inf\{c > 0;\, \mathrm{E}[\psi(|Z|/c)] \leq 1\}.$$

So, it is possible to generalize the definition of p-stable processes to ψ-stable processes. Then we can put $\delta_{i,\psi} = ||X_{\mathbf{i}} - X^*_{\mathbf{i}}||_\psi$ and $\Delta_\psi = \sum_{\mathbf{i}\in\mathbb{Z}^d} \delta_{\mathbf{i},\psi}$.

We say, that the process is ψ-stable if $\Delta_\psi < \infty$.

Now, we can recall some basic facts about covering numbers. Let us have a collection $\mathcal{A}$ of Borel subsets on $[0,1]^d$. We can equip the collection with the pseudometric ρ: $\rho(A,B) = \sqrt{\lambda(A \Delta B)}$, where Δ denotes a symmetric difference between the sets and λ is the Lebesgue measure. To measure a size of $\mathcal{A}$ it is possible to use a metric entropy. Let us recall, that the entropy $H(\mathcal{A}, \rho, \varepsilon)$ is the logarithm of $N(\mathcal{A}, \rho, \varepsilon)$, where $N(\mathcal{A}, \rho, \varepsilon)$ is so called covering number – it is the smallest number of open balls of radius ε with respect to ρ which cover $\mathcal{A}$.

Let $\mathcal{C}$ be a collection of subsets of a set $\mathcal{X}$. And let $F \subset \mathcal{X}$. We say that $\mathcal{C}$ picks out a certain subset of F if this can be formed as $F \cap C$ for some $C \in \mathcal{C}$. The collection $\mathcal{C}$ is said to shatter F if it picks out each of its $2^{|F|}$ subsets. The VC-index (Vapnik-Chervonenkis index) $V(\mathcal{C})$ of the class $\mathcal{C}$ is the smallest n for which no set of size n can be shattered by $\mathcal{C}$. Formally,

$$V(\mathcal{C}) = \inf\left\{n;\, \max_{x_1,\ldots,x_n\in\mathcal{X}} \Delta_n(\mathcal{C}, x_1, \ldots, x_n) < 2^n\right\},$$

where $\Delta_n(\mathcal{C}, x_1, \ldots, x_n) = |\{C \cap \{x_1, \ldots, x_n\}; C \in \mathcal{C}\}|$.

Now, we can formulate the invariance principle given by El Machkouri, Volný and Wu (2013) in [5].

Theorem 3.4 (El Machkouri, Volný and Wu [5]). *Let $(U_{\mathbf{i}} f)_{\mathbf{i}\in\mathbb{Z}^d}$ be a stationary centered random field and let $\mathcal{A}$ be a collection of regular Borel subsets of $[0,1]^d$. Assume that one of the following conditions holds:*

(i) *The collection $\mathcal{A}$ is a Vapnik-Chervonenkis class with an index V and there exists $p > 2(V-1)$ such that $f \in L_p$ and $\Delta_p < \infty$.*

(ii) *There exists a positive* θ *and* $0 < q < 2$*:* $\mathrm{E}[\exp(\theta|f|^{\beta(q)})] < \infty$*, where* $\beta(q) = 2q/(2-q)$ *and* $\Delta_{\psi(\beta(q))} < \infty$ *and such that the class* $\mathcal{A}$ *satisfies condition*

$$\int_0^1 (H(\mathcal{A}, \rho, \varepsilon))^{1/q} \, \mathrm{d}\varepsilon < \infty,$$

where

$$\psi_\beta(x) = \exp\{(x + h_\beta)^\beta\} - \exp\{h_\beta^\beta\}, \quad x \in \mathbb{R}^+,$$

with $\beta > 0$ *and*

$$h_\beta = ((1-\beta)\beta)^{\frac{1}{\beta}} \mathbb{I}_{\{0<\beta<1\}}.$$

(iii) $f \in L^\infty$, $\int_0^1 (H(\mathcal{A}, \rho, \varepsilon))^{1/2} \, \mathrm{d}\varepsilon < \infty$ *and*

$$\Delta_\infty < \infty.$$

Then the sequence of processes $\{n^{-d/2} S_n(A); A \in \mathcal{A}\}$*, where*

$$S_n(A) = \sum_{\mathbf{i} \in [\mathbf{0}, \mathbf{n}]} \lambda(nA \cap R_\mathbf{i}) U_\mathbf{i} f$$

with $R_\mathbf{i} = (\mathbf{i} - \mathbf{1}, \mathbf{i}]$*, converges in distribution in* $C(\mathcal{A})$ *to* σW*, where* W *is a standard Brownian motion indexed by* $\mathcal{A}$ *and* $\sigma^2 = \sum_{\mathbf{i} \in \mathbb{Z}^d} \mathrm{E}(f U_\mathbf{i} f)$*.*

Remark. We can state the question, what is the relation between the p-stability and Hannan's condition. Wu (2005), see [14], proved, in a 1-dimensional case, that $\Delta_2 \geq \sum_{i=1}^\infty ||P_0(f \circ T^i)||_2$. In other words, that 2-stability of a process implies Hannan's condition for this process. This result can be extend into a high dimensional case, too. Volný and Wang [12] show an example of a process such that it satisfies Hannan's condition but a stability does not take a place.

4. Conclusion

In this paper, we gave a brief introduction to problems on limit theorems for weakly dependent random fields. We briefly introduced some of very interesting results on this topic. We started with results on stationary processes and then we presented some extensions of these results to random fields.

Acknowledgement

Supported by Czech Science Foundation (project no. P201/11/P164). I would like to express thanks to referees and editors for their support and comments.

References

[1] Basu A. K., Dorea C. C. Y.: On functional central limit theorem for stationary martingale random fields. *Acta Mathematica Hungarica*, **33** (3), 307–316, 1979.

[2] Bradley R. C. (2007): *Introduction to strong mixing conditions*. Kendrick Press, Heber City, UT.

[3] Doukhan P., Louchichi S.: A new weak dependence condition and applications to moment inequalitites, *Stoch. Proc. Appl.*, **84**, 313–342, 1999.

[4] Gordin M. I.: A central limit theorem for stationary processes, *Soviet Math. Dokl.*, **10**, 1174–1176, 1969.

[5] El Machkouri M., Volný D., Wu W. B.: A central limit theorem for stationary random fields. *Stochastic Processes and their Applications*, **123** (1), 1–14, 2013.

[6] Hannan E. J.: The central limit theorem for time series regression. *Stochastic Processes and their Applications*, **9** (3), 281–289, 1979.

[7] Ledoux M., Talagrand M.: *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, 1991.

[8] Maxwell M., Woodroofe M.: Central limit theorems for additive functionals of Markov chains, *The Annals of Probability*, **28**, 713–724, 2000.

[9] Nahapetian B.: Billingsley-Ibragimov theorem for mart.-diff. random fields and its appl. to some models of classical statistical physics. *CRAS. Série 1, Mathématique*, **320** (12), 1539–1544, 1995.

[10] Poghosyan S., Roelly S.: Invariance principle for martingale-difference random fields. *Statistics and probability letters*, **38** (3), 235–245, 1998.

[11] Van Der Vaart A. W., Wellner J. A.: *Weak Convergence and Empirical Processes*. Springer, New York, 1996.

[12] Volný D., Wang Y.: An invariance principle for stationary random fields under Hannan's condition. *Stochastic Processes and their Applications*, **124** (12), 4012–4029, 2014.

[13] Wang Y., Woodroofe M.: A new condition on invariance principles for stationary random fields. *Statist. Sinica*, **23** (4), 1673–1696, 2013.

[14] Wu W. B.: Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences of the United States of America*, **102** (40), 14150–14154, 2005.

DIRICHLETOVO ROZDĚLENÍ VZHLEDEM K AITCHISONOVĚ MÍŘE NA SIMPLEXU

THE DIRICHLET DISTRIBUTION WITH RESPECT TO THE AITCHISON MEASURE ON THE SIMPLEX

Petra Kynčlová

Adresa: TU Wien, Wiedner Hauptstrasse 8-10, A-1040 Vienna, Austria; PřF UP, Katedra geoinformatiky, Tř. Svobody 26, 771 46 Olomouc

E-mail: kynclova.petra@gmail.com

Abstrakt: Dirichletovo rozdělení bývá standardně používáno pro parametrické modelování dat s konstantním součtem, mezi něž se řadí i kompoziční data. Geometrická struktura simplexu, výběrového prostoru kompozičních dat, je popsána tzv. Aitchisonovou geometrií, a je tedy odlišná od Euklidovské geometrie v reálném prostoru. Z tohoto důvodu se pro simplexový výběrový prostor zavádí Aitchisonova míra, která je relativní a je definována pomocí transformace Lebesgueovy míry z prostoru ortonormálních souřadnic na simplex. Dirichletovo rozdělení však bývá typicky vyjádřeno vzhledem k Lebesgueově míře. Cílem příspěvku je popsat vlastnosti a číselné charakteristiky Dirichletova rozdělení na simplexu vzhledem k Aitchisonově míře, resp. vzhledem k Lebesgueově míře v prostoru ortonormálních souřadnic, a důsledky volby parametrů na tvar Dirichletova rozdělení.

Klíčová slova: Kompoziční data, Aitchisonova geometrie na simplexu, Aitchisonova míra, Dirichletovo rozdělení na simplexu.

Abstract: The Dirichlet distribution is standardly used for parametric modelling of data with a constant sum constraint including compositional data. The geometric structure of the simplex, the sample space of compositional data, is characterized by the Aitchison geometry, and thus it differs from the Euclidean geometry in the real space. For that reason the Aitchison measure is introduced for the simplex space. The Aitchison measure is relative and it is defined as a transformation of the Lebesgue measure from the space of orthonormal coordinates to the simplex. However, the Dirichlet distribution is typically expressed with respect to the Lebesgue measure. The aim of the paper is to describe properties and characteristics of the Dirichlet distribution on the simplex with respect to the Aitchison measure, or with respect to the Lebesgue measure in the space of ortonormal coordinates, and to show effects of the choice of parameters for the shape of the distribution.

Keywords: Compositional data, Aitchison geometry on the simplex, Aitchison measure, Dirichlet distribution on the simplex.

1. Úvod

Dirichletovo rozdělení patří mezi známá rozdělení pravděpodobnosti definovaná pro data s konstantním součtem. Mezi speciální případy těchto dat patří i tzv. kompoziční data [1, 4]. Tato data ovšem indukují jinou přirozenou geometrickou strukturu výběrového prostoru svých reprezentací, simplexu, než je standardní používaná Euklidovská geometrie pro data z reálného prostoru. Z tohoto důvodu byla na simplexovém prostoru zadefinována alternativní pravděpodobnostní míra, označovaná jako Aitchisonova míra. Dirichletovo rozdělení však bývá typicky vyjádřeno vzhledem k Lebesgueově pravděpodobnostní míře. Otázkou tedy zůstává, jak se bude Dirichletovo rozdělení chovat s ohledem na přirozenou geometrickou strukturu simplexu, tedy vyjádříme-li hustotu Dirichletova rozdělení vzhledem k Aitchisonově míře.

Tento příspěvek se věnuje vlastnostem Dirichletova rozdělení vzhledem k Aitchisonově míře na simplexu. Druhá kapitola pojednává o problematice kompozičních dat a jim odpovídající geometrii. Třetí kapitola je věnována číselným charakteristikám na simplexu a čtvrtá zavedení Aitchisonovy pravděpodobnostní míry. V páté kapitole pak bude představeno Dirichletovo rozdělení vzhledem k Lebesgueově a Aitchisonově míře a jim odpovídající číselné charakteristiky. Závěrem budou pomocí provedených simulací demonstrovány a porovnány základní vlastnosti a odlišnosti Dirichletova rozdělení, budeme-li uvažovat oba případy, tj. Lebesgueovu a Aitchisonovu míru.

2. Kompoziční data a jejich geometrie

Ve statistice se pod pojmem kompoziční data rozumí data nesoucí pouze relativní informaci jako například proporce či procentuální části celku. Od ostatních dat se tedy liší především tím, že informace, kterou nesou, je obsažena pouze v podílech mezi složkami [1]. D-složkový kompoziční vektor, neboli také kompozice, je definován jako kladný reálný vektor $\mathbf{x} = (x_1, \ldots, x_D)'$, jehož složky nesou výhradně relativní informaci. Dalším specifickým znakem těchto dat je možnost reprezentovat je jako data s konstantním součtem, tj.

$$\mathcal{C}(\mathbf{x}) = \left(\frac{\kappa x_1}{\sum_{i=1}^{D} x_i}, \ldots, \frac{\kappa x_D}{\sum_{i=1}^{D} x_i} \right)'.$$

Výběrový prostor reprezentací kompozic při zvoleném $\kappa > 0$ je simplex,

$$\mathcal{S}^D = \left\{ (x_1, x_2, \ldots, x_D)' : x_1 > 0, x_2 > 0, \ldots, x_D > 0; \sum_{i=1}^{D} x_i = \kappa \right\}.$$

Při práci s mnohorozměrnými daty jsme zvyklí pracovat v reálném vektorovém prostoru, na kterém je definovaná standardní Euklidovská geometrie. Euklidovská geometrie není ovšem vhodná pro kompoziční data [8]. Z toho důvodu je nutné zavést jinou geometrii, která by vedla k relevantním výsledkům při statistickém zpracování kompozičních dat. Při práci s kompozicemi tedy používáme Aitchisonovu geometrii na simplexu, která je charakterizována operacemi perturbace, mocnění a Aitchisonovým skalárním součinem

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C}(x_1 y_1, x_2 y_2, \ldots, x_D y_D)', \quad \alpha \odot \mathbf{x} = \mathcal{C}(x_1^\alpha, x_2^\alpha, \ldots, x_D^\alpha)', \quad \alpha \in \mathbb{R},$$

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}.$$

S existencí Aitchisonova skalárního součinu jsou pak zavedeny pojmy Aitchisonovy normy a vzdálenosti

$$\| \mathbf{x} \|_a = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_a}, \quad d_a(\mathbf{x}, \mathbf{y}) = \sqrt{\frac{1}{D} \sum_{i<j} \Big(\ln \frac{x_i}{x_j} - \ln \frac{y_i}{y_j} \Big)^2}.$$

Aitchisonova vzdálenost má standardní vlastnosti jako vzdálenost Euklidovská. Je invariantní vůči perturbaci, invariantní na změnu škály a nezávisí na pořadí složek kompozice.

Zavedení Aitchisonovy geometrie na simplexu zaručuje existenci ortonormální báze $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{D-1}\}$, což znamená, že kompozici $\mathbf{x}$ jsme nyní schopni vyjádřit ve tvaru lineární kombinace

$$\mathbf{x} = (\langle \mathbf{x}, \mathbf{e}_1 \rangle_a \odot \mathbf{e}_1) \oplus (\langle \mathbf{x}, \mathbf{e}_2 \rangle_a \odot \mathbf{e}_2) \oplus \cdots \oplus (\langle \mathbf{x}, \mathbf{e}_{D-1} \rangle_a \odot \mathbf{e}_{D-1}).$$

Souřadnice kompozice $\mathbf{x}$ vzhledem k dané ortonormální bázi $\{\mathbf{e}_1, \ldots, \mathbf{e}_{D-1}\}$ tvoří izometrickou logratio (ilr) transformaci kompozice $\mathbf{x}$ [2], tj.

$$\mathrm{ilr}(\mathbf{x}) = \big(\langle \mathbf{x}, \mathbf{e}_1 \rangle_a, \langle \mathbf{x}, \mathbf{e}_2 \rangle_a, \ldots, \langle \mathbf{x}, \mathbf{e}_{D-1} \rangle_a \big)'.$$

Stejně jako v reálném prostoru, i na simplexu existuje nekonečné množství ortonormálních bází, tedy i ilr transformací jsme schopni vytvořit nekonečně mnoho. Jednou konkrétní volbou ortonormální báze dostaneme ilr souřadnice [3]

$$\mathrm{ilr}(\mathbf{x}) = (z_1, \ldots, z_D)', \quad z_i = \sqrt{\frac{i}{i+1}} \ln \frac{\sqrt[i]{\prod_{j=1}^{i} x_j}}{x_{i+1}}, \quad i = 1, \ldots, D-1.$$

Inverzní ilr transformací vyjádříme ilr souřadnice zpět na simplexu tak, že $\mathbf{x} = \text{ilr}^{-1}(\mathbf{z})$, konkrétně

$$x_i = \exp\left(\sum_{j=i}^{D} \frac{z_j}{\sqrt{j(j+1)}} - \sqrt{\frac{i-1}{i}} z_{i-1}\right),$$

kde $z_0 = z_D = 0$ pro $i = 1, \ldots, D$.

Máme-li souřadnice vzhledem k ortonormální bázi, můžeme ke statistické analýze takto vyjádřené kompozice použít standardní používané metody. Operace perturbace $\oplus$ a mocnění $\odot$ v simplexovém prostoru představují analogii ke klasickým operacím součtu vektorů a násobení vektoru skalárem použitých v prostoru ortonormálních souřadnic vzhledem k libovolné bázi, která nemusí být nutně ortonormální. V případě, že máme souřadnice vzhledem k ortonormální bázi, lze na ně aplikovat standardní skalární součin i Euklidovskou vzdálenost v prostoru $\mathbb{R}^{D-1}$.

3. Číselné charakteristiky na simplexu

Stejně jako u jiných rozdělení nás zajímají odpovídající číselné charakteristiky. Pohybujeme-li se ve výběrovém prostoru kompozičních dat, je potřeba je zadefinovat s ohledem na geometrickou strukturu simplexu. Výhodným postupem se jeví použití geometrické interpretace číselných charakteristik. V případě kompozičních dat tak hovoříme o středu kompozice a metrickém rozptylu [7].

Střed kompozice je odpovídající charakteristikou polohy a představuje kompozici $\text{cen}(\mathbf{x})$, která minimalizuje výraz $\text{E}[d_a^2(\mathbf{x}, \text{cen}(\mathbf{x}))]$. Střed kompozice je tak dán jako

$$\text{cen}(\mathbf{x}) = \mathcal{C}(\exp(\text{E}[\ln \mathbf{x}])).$$

Konkrétní pro izometrickou logratio transformaci platí

$$\text{ilr}(\text{cen}[\mathbf{x}]) = \text{E}[\text{ilr}(\mathbf{x})].$$

To znamená, že jsme schopni spočítat střední hodnotu $\text{E}[\text{ilr}(\mathbf{x})]$ použitím standardní definice a následně aplikací inverzní ilr transformace získat kompozici $\text{cen}(\mathbf{x})$. Můžeme tak tvrdit, že použití standardní statistické metodiky na souřadnice vzhledem k dané ortonormální bázi je ekvivalentní s prací přímo na kompozicích.

Metrický rozptyl popisuje variabilitu náhodné kompozice a je vyjádřen jako střední hodnota čtvercové Aitchisonovy vzdálenosti kompozice od jejího středu, tj.

$$\text{Mvar}[\mathbf{x}] = \text{E}[d_a^2(\mathbf{x}, \text{cen}[\mathbf{x}])].$$

Při použití izometrické logratio transformace pro metrický rozptyl platí

$$\mathrm{Mvar}[\mathbf{x}] = \mathrm{E}[d_e^2(\mathrm{ilr}(\mathbf{x}), \mathrm{ilr}(\mathrm{cen}[\mathbf{x}]))].$$

4. Aitchisonova míra

Většina statistických metod předpokládá, že zkoumaná data pocházejí z reálného prostoru s Euklidovskou geometrií. V tomto případě (uvažujeme-li spojitý náhodný vektor) jsou hustoty rozdělení pravděpodobnosti vyjádřeny vzhledem k Lebesgueově pravděpodobnostní míře. Geometrická struktura daného výběrového prostoru však může být v některých případech odlišná a je tedy nutné pracovat s jinou mírou než právě s Lebesgueovou.

Nechť je dán vektorový prostor E, na kterém je zaveden skalární součin. Zde můžeme zavést pravděpodobnostní míru λ_E, jež bude se strukturou prostoru E kompatibilní, a to prostřednictvím Lebesgueovy míry na ortonormálních souřadnicích. Funkce hustoty f_E, která je definována na E, je pak dána jako Radon-Nikodymova derivace pravděpodobnostní míry P vzhledem k míře λ_E. Míra λ_E má v prostoru E stejné vlastnosti jako Lebesgueova míra v reálném prostoru (tedy v prostoru ortonormálních souřadnic) [6].

Stejným způsobem je zavedena i Aitchisonova míra λ_a, která odpovídá geometrické struktuře simplexu [7]. Míra λ_a je relativní a je absolutně spojitá vzhledem k Lebesgueově míře λ. Vztah mezi mírami λ_a a λ je pak dán pomocí Jakobiánu

$$\frac{\mathrm{d}\lambda_a}{\mathrm{d}\lambda} = \frac{1}{\sqrt{D}x_1 \cdots x_D}. \tag{1}$$

Obecně lze tímto způsobem zadefinovat Lebesgueovu míru libovolného Euklidovského prostoru.

5. Dirichletovo rozdělení na simplexu

Tato kapitola se věnuje Dirichletovu rozdělení vzhledem k Lebesgueově i Aitchisonově míře na simplexu. Odlišné vlastnosti a charakteristiky tohoto rozdělení jsou následně demonstrovány na simulacích.

Definition 5.1. *Náhodný vektor $\mathbf{X} \in \mathcal{S}^D$ má D-rozměrné Dirichletovo rozdělení s parametrem $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_D)' \in \mathbb{R}_+^D$, jestliže jeho hustota pravděpodobnosti má tvar*

$$f(\mathbf{x}) = \frac{\mathrm{d}P}{\mathrm{d}\lambda}(\mathbf{x}) = \frac{\Gamma(\alpha_+)}{\prod_{i=1}^D \Gamma(\alpha_i)} \prod_{i=1}^D x_i^{\alpha_i - 1},$$

kde λ je Lebesgueova míra, $\alpha_+ = \sum_{i=1}^{D} \alpha_i$, a Γ je gamma funkce. Značíme $\mathbf{X} \sim \mathcal{D}^D(\boldsymbol{\alpha})$.

Definice představuje Dirichletovo rozdělení vzhledem k Lebesgueově pravděpodobnostní míře. Jak již bylo řečeno, na simplexu zavádíme alternativní míru, která je této geometrické struktuře vlastní. Vyjádříme-li tedy hustotu vzhledem k Aitchisonově míře λ_a pomocí Jakobiánu (1), dostaneme hustotu Dirichletova rozdělení ve tvaru

$$f_a(\mathbf{x}) = \frac{\mathrm{d}P}{\mathrm{d}\lambda_a}(\mathbf{x}) = \frac{\Gamma(\alpha_+)\sqrt{D}}{\prod_{i=1}^{D}\Gamma(\alpha_i)} \prod_{i=1}^{D} x_i^{\alpha_i}.$$

Explicitní vyjádření hustoty Dirichletova rozdělení vzhledem k Lebesgueově míře v prostoru ilr souřadnic je značně komplikované, tudíž i interpretace jednotlivých parametrů $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_D)'$ je prakticky nemožná.

Odlišnosti jsou patrné i při výpočtu číselných charakteristik Dirichletova rozdělení. Modus a střední hodnota Dirichletova rozdělení vzhledem k Lebesgueově míře λ a Aitchisonově míře λ_a mají následující tvar:

$$\mathrm{modus}(\mathbf{X}) = \left(\frac{\alpha_1 - 1}{\alpha_+ - D}, \ldots, \frac{\alpha_D - 1}{\alpha_+ - D}\right)', \quad \mathrm{E}(\mathbf{X}) = \left(\frac{\alpha_1}{\alpha_+}, \ldots, \frac{\alpha_D}{\alpha_+}\right)',$$

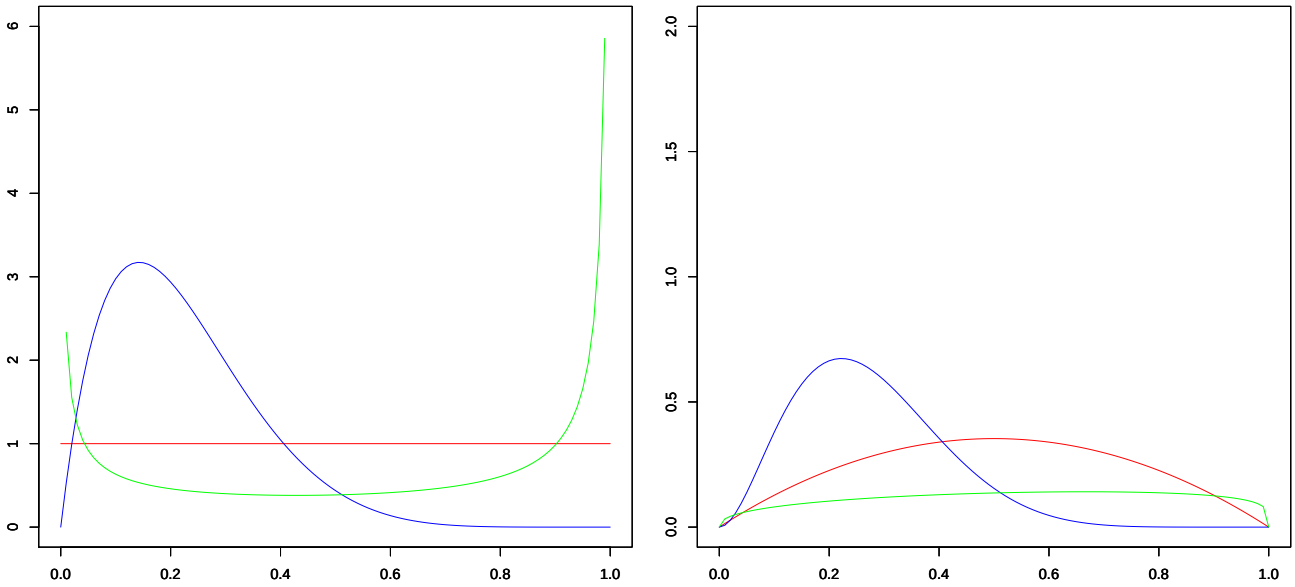
$$\mathrm{modus}_a(\mathbf{X}) = \left(\frac{\alpha_1}{\alpha_+}, \ldots, \frac{\alpha_D}{\alpha_+}\right)', \quad \mathrm{E}_a(\mathbf{X}) = \mathcal{C}\left(\mathrm{e}^{\psi(\alpha_1)}, \ldots, \mathrm{e}^{\psi(\alpha_D)}\right)',$$

kde $\psi(t) = \frac{\partial \ln \Gamma(t)}{\partial t}$ je digamma funkce a $\mathcal{C}$ je operátor uzávěru.

Z uvedených výrazů je patrné, že určení modu kompozice $\mathbf{X}$ je velmi podobné a výpočetně snadné vzhledem k míře Lebesgueově i Aitchisonově, zatímco výpočet střední hodnoty je mnohem komplikovanější. Střední hodnota vzhledem k Lebesgueově míře odpovídá modu vzhledem k míře λ_a. Nejjednodušším způsobem, jak určit očekávanou hodnotu kompozice $\mathbf{X}$ vzhledem k Aitchisonově míře spočívá ve vyjádření funkce hustoty Dirichletova rozdělení pomocí souřadnic vzhledem k ortonormální bázi a následné aplikaci standardní definice střední hodnoty na vektor $\mathrm{ilr}(\mathbf{x})$ [5]. Výsledkem jsou pak souřadnice kompozice $\mathrm{E}_a(\mathbf{X})$ vzhledem k dané ortonormální bázi.

Ke zjištění variability, např. k výpočtu metrického rozptylu, je nutné pracovat přímo na souřadnicích, tedy s $(D-1)$-složkovými vektory, jelikož metrický rozptyl není prvkem simplexu [7]. Jedná se pouze o numericky určenou hodnotu, která vyjadřuje míru celkové disperze kompozice. Metrický rozptyl kompozice $\mathbf{X} \sim \mathcal{D}^D(\boldsymbol{\alpha})$ definované na jednotkovém simplexu je dán ve tvaru

$$\mathrm{Mvar}(\mathbf{X}) = \frac{D-1}{D}(\psi'(\alpha_1) + \cdots + \psi'(\alpha_D)),$$



(a) (b)

Obrázek 1: Hustoty Dirichletova rozdělení vzhledem (a) k Lebesgueově míře λ a (b) k Aitchisonově míře λ_a s parametry $\boldsymbol{\alpha} = (1, 1)'$ (—), $\boldsymbol{\alpha} = (2, 7)'$ (—) a $\boldsymbol{\alpha} = (0.4, 0.2)'$ (—).

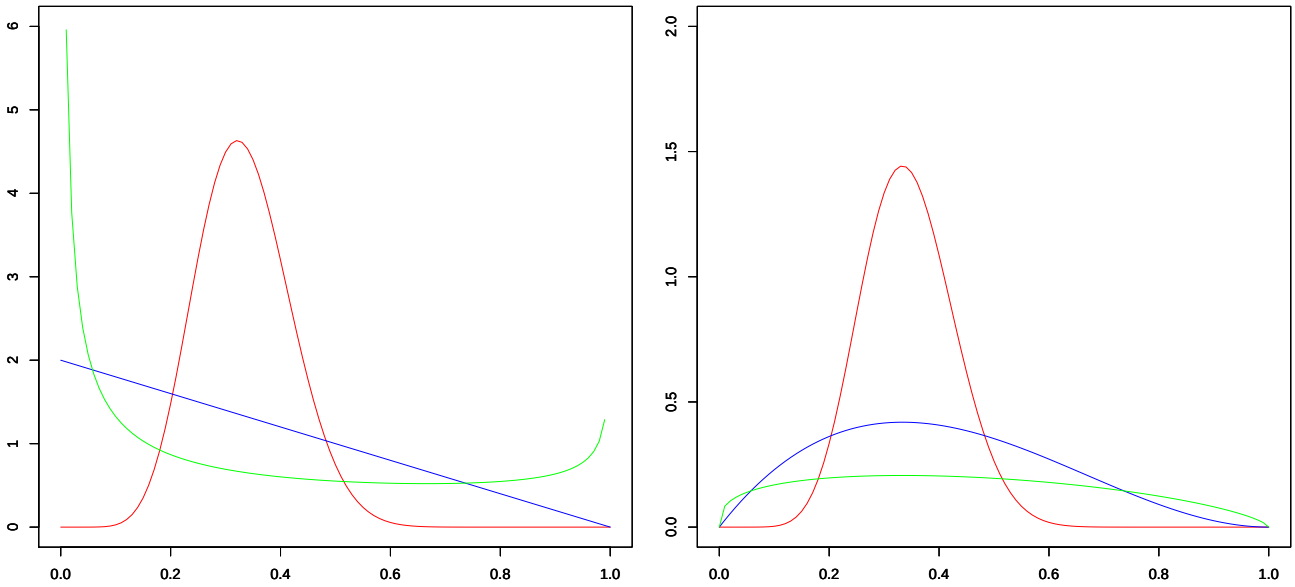
kde $\psi'(t) = \frac{\partial \psi(t)}{\partial t}$, $t > 0$, je trigamma funkce. Dosud nebyl zjištěn žádný způsob, jak získat explicitní vyjádření, aniž by bylo nutné volit ortonormální bázi pro reprezentaci kompozice.

Výše uvedené vlastnosti Dirichletova rozdělení jsou zjevné i na simulacích, které byly provedeny pro dvousložkové (Obrázky 1 a 2) a třísložkové kompozice (Obrázek 3).

Porovnání hustot Dirichletova rozdělení vzhledem k Lebesgueově míře λ a vzhledem k Aitchisonově míře λ_a pro dvousložkové kompozice při různé volbě parametru $\boldsymbol{\alpha}$ můžeme pozorovat na Obrázku 1 a 2. Pro Dirichletovo rozdělení vzhledem k Aitchisonově míře λ_a dostáváme ve všech situacích unimodální funkci. Pro hustotu vzhledem k Lebesgueově míře λ ovšem tato výhodná vlastnost neplatí. Unimodalita zde nastává pouze za předpokladu, že jsou všechny složky parametru $\boldsymbol{\alpha}$ větší než 1. V případě, že jsou všechny složky $\alpha_i < 1$, má funkce vertikální asymptoty v 0 a 1. Speciálně pak pro $\boldsymbol{\alpha} = (1, 1)'$ je hustota konstantní.

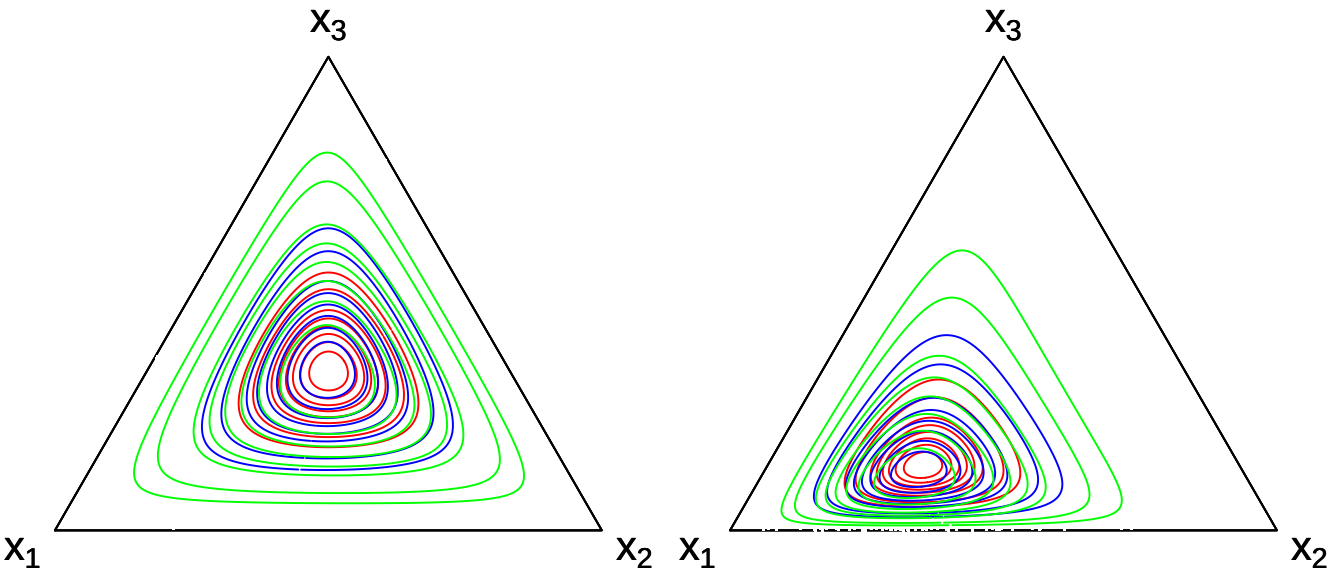
Analogické chování jako pro dvousložkové kompozice je obecně pozorovatelné i pro počet složek kompozice $D > 2$. Hustoty Dirichletova rozdělení vzhledem k Aitchisonově míře λ_a na simplexu jsou vždy unimodální.

Na provedených simulacích bylo zjištěno, že hustoty Dirichletova rozdělení se chovají značně specificky v případě, budeme-li volit parametr $\boldsymbol{\alpha}$ v rámci



(a) (b)

Obrázek 2: Hustoty Dirichletova rozdělení vzhledem (a) k Lebesgueově míře λ a (b) k Aitchisonově míře λ_a s parametry $\boldsymbol{\alpha} = (10, 20)'$ (—), $\boldsymbol{\alpha} = (1, 2)'$ (—) a $\boldsymbol{\alpha} = (1/3, 2/3)'$ (—).



(a) (b)

Obrázek 3: Hustoty Dirichletova rozdělení vzhledem k Aitchisonově míře λ_a na simplexu s parametry (a) $\boldsymbol{\alpha} = (10, 10, 10)'$ (—), $\boldsymbol{\alpha} = (5, 5, 5)'$ (—), $\boldsymbol{\alpha} = (2, 2, 2)'$ (—), (b) $\boldsymbol{\alpha} = (20, 10, 5)'$ (—), $\boldsymbol{\alpha} = (10, 5, 2.5)'$ (—), $\boldsymbol{\alpha} = (5, 2.5, 1.25)'$ (—).

třídy ekvivalentních kompozic, tj. zachováme-li poměry mezi jednotlivými složkami parametru $\boldsymbol{\alpha}$ (Obrázek 2 a 3). Vzhledem k Aitchisonově míře λ_a jsou hustoty nejen opět unimodální, jak již bylo obecně ukázáno, ale navíc mají vždy též stejný modus. Vzhledem k Lebesgueově míře ani jedno tvrzení neplatí. Unimodalita zde nastává pouze v případě, že jsou všechny složky kompozice $\boldsymbol{\alpha}$ větší než jedna, ale modus zde stejný není.

Při zkoumání variability v Dirichletově modelu vzhledem k Aitchisonově míře na simplexu obecně platí, že čím větší jsou hodnoty složek parametru $\boldsymbol{\alpha}$, tím menší je metrický rozptyl.

6. Závěr

Z provedených simulací je patrné, že Dirichletovo rozdělení vyjádřené vzhledem k Aitchisonově míře přináší výhodné vlastnosti při modelování dat jako je právě výše zmíněná unimodalita, která vzhledem k Lebesgueově míře nebývá dosaženo. Obecně můžeme říci, že použití Aitchisonovy míry nám umožňuje eliminovat nepříznivé vlastnosti Dirichletova rozdělení vzhledem k míře Lebesgueově.

Stále však v tomto případě vyvstává otázka interpretovatelnosti a použitelnosti Dirichletova rozdělení v reálných aplikacích, s výjimkou použití Dirichleta jako apriorního rozdělení v Bayesovských metodách. Tato problematika ovšem vyžaduje další studium.

Poděkování

Tato práce vznikla za podpory Operačního programu vzdělávání pro konkurenceschopnost – Evropský sociální fond (projekt CZ.1.07/2.3.00/20.0170 Ministerstva školství, mládeže a tělovýchovy České republiky).

Literatura

[1] Aitchison J.: *The Statistical Analysis of Compositional Data.* Chapman & Hall, London, 1986.

[2] Egozcue J. J., Pawlowsky-Glahn V., Mateu-Figueras G., Barceló-Vidal C.: Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, **35** (3), 279–300, 2003.

[3] Filzmoser P., Hron K.: Outlier detection for compositional data using robust methods, *Mathematical Geosciences*, **40** (3), 233–248, 2008.

[4] Hron K.: Elementy statistické analýzy kompozičních dat, In: Antoch J., Dohnal G. (eds.): *Sborník prací 16. letní školy JČMF Robust 2010*, Praha, 41–48, 2010.

[5] Mateu-Figueras K., Pawlowsky-Glahn V.: The Dirichlet distribution with respect to the Aitchison measure on the simplex – a first approach, In: Mateu-Figueras G., Barceló-Vidal C. (eds.): *Compositional Data Analysis Workshop – CoDaWork'05*, Proceedings, Universitat de Girona, 2005.

[6] Mateu-Figueras G., Pawlowsky-Glahn V., Egozcue J. J.: *The principle of working on coordinates.* In: Pawlowsky-Glahn V., Buccianti A. (eds.): *Compositional Data Analysis: Theory and Applications*, Wiley, Chichester, 31–42, 2011.

[7] Monti G. S., Mateu-Figueras G., Pawlowsky-Glahn V., Egozcue J. J.: The shifted-scaled Dirichlet distribution in the simplex, In: Egozcue J. J., Tolosana-Delgado R., Ortego M. I. (eds.): *Compositional Data Analysis Workshop – CoDaWork'11*, Proceedings, International Center for Numerical Methods in Engineering (CIMNE), Barcelona, 2011.

[8] Pawlowsky-Glahn V., Buccianti A. (eds.): *Compositional Data Analysis: Theory and Applications.* Wiley, Chichester, 2011.

ODHADY ZÁKLADNÍHO RIZIKA V REGRESNÍCH MODELECH OPRAV

ESTIMATION OF THE BASELINE HAZARD IN REGRESSION MODELS FOR REPAIRABLE SYSTEMS

Petr Novák[1,2]

Adresa: [1]MFF UK v Praze, KMPS, Sokolovská 83, 186 75 Praha 8
[2]ÚTIA AV ČR, Pod Vodárenskou věží 4, 182 08 Praha 8

E-mail: [1]novakp@karlin.mff.cuni.cz

Abstrakt: Pozorujeme nezávislá zařízení podléhající opotřebení a pomocí vhodných regresních modelů se snažíme popsat vliv jejich průběžných oprav a údržby na rozdělení doby do selhání. Nejčastěji používané modely, jako je Coxův model proporcionálního rizika nebo model zrychleného času, popisují vliv regresorů na určitou základní rizikovou funkci. Tu je potřeba buď vhodně parametrizovat, nebo odhadnout neparametricky. V této práci se zaměřujeme na metody porovnávání a testování hypotéz o tvaru základního rizika v modelech oprav a předvádíme jejich využití.

Klíčová slova: Analýza spolehlivosti, modely oprav, regrese, základní riziko.

Abstract: When observing independent devices which are subject to degradation, we want to describe the influence of repairs and preventive maintenance actions on the time to failure distribution with the help of suitable regression models. Commonly used models, as the Cox proportional hazards and the accelerated failure time model assume, that the covariates influence a certain baseline hazard function, which must be either parametrized or estimated nonparametrically. In this work we focus on methods how to estimate and test hypotheses about the shape of the baseline hazard and we show their applications.

Keywords: Reliability analysis, repair models, regression, baseline hazard.

1. Úvod – údržba a opravy

Zkoumáme data reprezentující životnost n nezávislých systémů podléhajících opotřebení. Když se systém porouchá, je nutné provést opravu. Selhání se také snažíme předejít preventivními údržbami. Označíme T_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, n_i$ seřazené časy oprav a údržeb i-tého zařízení a Δ_{ij} indikátory, zda na i-tém zařízení byla v j-tém čase provedena oprava ($\Delta_{ij} = 1$) nebo

preventivní údržba ($\Delta_{ij} = 0$). Zavedeme čítací procesy oprav a údržeb do času t:

$$N_{i\bullet}(t) = \sum_{j=1}^{n_i} I(T_{ij} \leq t, \Delta_{ij} = 1), \quad M_{i\bullet}(t) = \sum_{j=1}^{n_i} I(T_{ij} \leq t, \Delta_{ij} = 0).$$

Označíme rizikové funkce pro každé zařízení

$$\lambda_i(t) = \lim_{h \to 0} P(N_{i\bullet}(t+h) - N_{i\bullet}(t) \geq 1 | \mathcal{H}(t))/h,$$

kde $\mathcal{H}(t)$ značí historii událostí do času t. Pracujeme s kumulovanými rizikovými funkcemi $\Lambda_i(t) = \int_0^t \lambda_i(s)\mathrm{d}s$, příslušnými funkcemi přežití $S_i(t) = \exp(-\Lambda_i(t))$ a hustotami $f_i(t) = -\frac{\mathrm{d}}{\mathrm{d}t} S_i(t)$. Věrohodnost lze přepsat jako

$$L = \prod_{i=1}^{n} \prod_{j=1}^{n_i} \left(\frac{f_i(T_{ij}^-)}{S_i(T_{i(j-1)})} \right)^{\Delta_{ij}} \left(\frac{S_i(T_{ij})}{S_i(T_{i(j-1)})} \right)^{1-\Delta_{ij}} = \prod_{i=1}^{n} \prod_{j=1}^{n_i} \lambda_i(T_{ij}^-)^{\Delta_{ij}} \cdot S_i(T_{in_i})$$

a log-věrohodnost má pak tvar

$$l = \sum_{i=1}^{n} \sum_{j=1}^{n_i} \Delta_{ij} \log \lambda_i(T_{ij}^-) - \int_0^{T_{in_i}} \lambda_i(t)\mathrm{d}t.$$

Věrohodnost dat lze zapsat pomocí čítacích procesů. Zavedeme čítací procesy pro j-té selhání či opravu i-tého zařízení a příslušný indikátor rizika

$$N_{ij}(t) = \Delta_{ij} I(T_{ij} \leq t), \quad M_{ij}(t) = (1 - \Delta_{ij}) I(T_{ij} \leq t),$$

$$Y_{ij}(t) = I(T_{i,j-1} < t \leq T_{ij}).$$

Dostaneme

$$l = \sum_{ij} \int_0^{\infty} \left(\log \lambda_i(t^-)\mathrm{d}N_{ij}(t) - Y_{ij}(t)\lambda_i(t^-)\mathrm{d}t \right).$$

2. Regresní modely oprav

2.1. Coxův model proporcionálního rizika

Předpokládáme, že každá oprava či údržba multiplikativně sníží nebo zvýší riziko, vliv mohou mít i případné další regresory, ozn. $Z_i(t)$. Uvažujeme rizikovou funkci [1]

$$\lambda_i(t) = \lambda_0(t) \mathrm{e}^{M_{i\bullet}(t)\rho + N_{i\bullet}(t)\varphi + Z_i^T(t)\beta}.$$

Při parametrickém základním riziku lze dosadit do logaritmické věrohodnosti a maximalizovat. Považujme ale základní riziko za neznámé. Označíme $\boldsymbol{\beta} = (\rho, \varphi, \beta)^T$ a $\boldsymbol{X}_i^T(t) = (N_{i\bullet}(t), M_{i\bullet}(t), Z_i^T(t))$. Skóre, získané dosazením rizikové funkce do logaritmické věrohodnosti a derivováním podle parametrů, $U(\boldsymbol{\beta}) = \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\beta}} l$, závisí na neznámé $\Lambda_0(t)$, kterou nahradíme odhadem Nelson-Aalenova typu

$$\widehat{\Lambda}_0(t, \boldsymbol{\beta}) = \int_0^t \frac{\mathrm{d}N_{\bullet\bullet}(s)}{\sum_{ij} \mathrm{e}^{\boldsymbol{X}_i^T(s^-)\boldsymbol{\beta}} Y_{ij}(s)},$$

kde $\bullet$ značí součet přes příslušný index. Po dosazení získáme skóre ve tvaru

$$\widehat{U}(\boldsymbol{\beta}) = \sum_{ij} \int_0^\infty \left(\boldsymbol{X}_i(t^-) - \frac{\sum_{kl} \boldsymbol{X}_k(t^-) \mathrm{e}^{\boldsymbol{X}_k^T(t^-)\boldsymbol{\beta}} Y_{kl}(t)}{\sum_{kl} \mathrm{e}^{\boldsymbol{X}_k^T(t^-)\boldsymbol{\beta}} Y_{kl}(t)} \right) \mathrm{d}N_{ij}(t)$$

a pro nalezení odhadů parametrů řešíme rovnice $\widehat{U}(\boldsymbol{\beta}) = 0$.

2.2. Model zrychleného času

Můžeme také předpokládat, že každá oprava či údržba a regresory způsobí, že virtuální čas plyne pomaleji nebo rychleji (Accelerated Failure Time model, AFT). Využijeme transformaci času [2]:

$$t \to \int_0^t \mathrm{e}^{M_{i\bullet}(s)\rho + N_{i\bullet}(s)\varphi + Z_i^T(s)\beta} \mathrm{d}s =: h_i(t, \boldsymbol{\beta}).$$

Riziková fukce pak má tvar

$$\lambda_i(t) = \lambda_0(h_i(t, \boldsymbol{\beta})) \mathrm{e}^{M_{i\bullet}(t)\rho + N_{i\bullet}(t)\varphi + Z_i^T(t)\beta}.$$

Pokud základní riziková funkce bude konstantní, tedy odpovídající exponenciálnímu rozdělení, oba modely splývají. Zavedeme transformované procesy

$$N_{ij}^*(t, \boldsymbol{\beta}) = \Delta_{ij} I(h_i(T_{ij}, \boldsymbol{\beta}) \le t), \quad M_{ij}^*(t, \boldsymbol{\beta}) = (1 - \Delta_{ij}) I(h_i(T_{ij}, \boldsymbol{\beta}) \le t),$$
$$Y_{ij}^*(t, \boldsymbol{\beta}) = I(h_i(T_{i,j-1}, \boldsymbol{\beta}) < t \le h_i(T_{ij}, \boldsymbol{\beta})), \quad \boldsymbol{X}_i^*(t, \boldsymbol{\beta}) = \boldsymbol{X}_i(h_i^{-1}(t, \boldsymbol{\beta})).$$

Přesné skóre má složitější tvar, je ale možné jej nahradit přibližným [2] a opět dosadit Nelson-Aalenův odhad kumulovaného základního rizika

$$\widehat{\Lambda}_0(t, \boldsymbol{\beta}) = \int_0^t \frac{\mathrm{d}N_{\bullet\bullet}^*(s, \boldsymbol{\beta})}{\sum_{ij} Y_{ij}^*(t, \boldsymbol{\beta})}.$$

Získáme

$$\widetilde{U}(\boldsymbol{\beta}) = \sum_{ij} \int_0^\infty \left(\boldsymbol{X}_i^*(t^-, \boldsymbol{\beta}) - \frac{\sum_{kl} \boldsymbol{X}_k^*(t^-, \boldsymbol{\beta}) Y_{kl}^*(t, \boldsymbol{\beta})}{\sum_{kl} Y_{kl}^*(t, \boldsymbol{\beta})} \right) \mathrm{d}N_{ij}^*(t, \boldsymbol{\beta}).$$

Protože skóre není spojité v $\boldsymbol{\beta}$, najdeme odhady parametrů minimalizací $\|\widetilde{U}(\boldsymbol{\beta})\|$.

3. Vlastnosti odhadů Λ_0

Navážeme zde na [3], kde bylo hlavním cílem hledání a interpretace odhadů $\widehat{\boldsymbol{\beta}}$, a zaměříme se na studování vlastností odhadů kumulovaného základního rizika. Pro každý z modelů zvlášť uvažujme proces

$$W(t) = n^{1/2} \big(\widehat{\Lambda}_0(t, \widehat{\boldsymbol{\beta}}) - \Lambda_0(t) \big).$$

Pomocí funkcionální centrální limitní věty [4] se dá ukázat, že pro $n \to \infty$ konverguje $W(t)$ slabě ke Gaussovskému procesu s nulovou střední hodnotou a konečnou kovarianční funkcí. Tím je zajištěna konzistence Nelson-Aalenových odhadů. Kovarianční funkce je pro každý z modelů různá. V AFT modelu závisí na neznámých λ_0 a $\lambda_0' = \mathrm{d}\lambda_0/\mathrm{d}t$ a není možné ji snadno odhadnout přímo. Předvedeme postup pomocí simulační metody.

Resampling zákadního rizika

Nejprve uvažujme Coxův model. Generujme $G_1, \ldots, G_n$ (i.i.d.) z $N(0,1)$. Mějme

$$\widehat{U}_G(\boldsymbol{\beta}) = \sum_{ij} \int_0^\infty \left(\boldsymbol{X}_i(t^-) - \frac{\sum_{kl} \boldsymbol{X}_k(t^-) \mathrm{e}^{\boldsymbol{X}_k^T(t^-)\boldsymbol{\beta}} Y_{kl}(t)}{\sum_{kl} \mathrm{e}^{\boldsymbol{X}_{ik}^T(t^-)\boldsymbol{\beta}} Y_{kl}(t)} \right) G_i \mathrm{d}N_{ij}(t).$$

Najdeme $\widehat{\boldsymbol{\beta}}_G$ jako řešení rovnice $\widehat{U}(\widehat{\boldsymbol{\beta}}_G) = \widehat{U}_G(\widehat{\boldsymbol{\beta}})$ a položíme

$$\widehat{W}(t) = n^{1/2} \Big(\widehat{\Lambda}_0(t, \widehat{\boldsymbol{\beta}}) - \widehat{\Lambda}_0(t, \widehat{\boldsymbol{\beta}}_G) + \widehat{\Lambda}_{0G}(t, \widehat{\boldsymbol{\beta}}) \Big),$$

kde

$$\widehat{\Lambda}_{0G}(t, \boldsymbol{\beta}) = \sum_{ij} \int_0^t \frac{G_i \mathrm{d}N_{ij}(s)}{\sum_{kl} \mathrm{e}^{\boldsymbol{X}_k^T(s^-)\boldsymbol{\beta}} Y_{kl}(s)}.$$

Pomocí funkcionální CLV [4] se dá ukázat, že za platnosti Coxova modelu $\widehat{W}(t)$ konverguje slabě ke stejnému Gaussovskému procesu jako $W(t)$. Důkaz

vychází z postupu pro Coxův model s rekurentními událostmi [5], přičemž je potřeba zohlednit použití oprav a údržeb jako regresorů.

V modelu zrychleného času postupujeme obdobně, vyrobíme replikované přibližné skóre

$$\widetilde{U}_G(\boldsymbol{\beta}) = \sum_{ij} \int_0^\infty \left(\boldsymbol{X}_i^*(t^-) - \frac{\sum_{kl} \boldsymbol{X}_k^*(t^-) Y_{kl}^*(t)}{\sum_{kl} Y_{kl}^*(t, \boldsymbol{\beta})} \right) G_i \mathrm{d}N_{ij}^*(t),$$

najdeme $\widehat{\boldsymbol{\beta}}_G$ řešení rovnice $\widetilde{U}(\widehat{\boldsymbol{\beta}}_G) = \widetilde{U}_G(\widehat{\boldsymbol{\beta}})$ a položíme

$$\widehat{\Lambda}_{0G}(t, \boldsymbol{\beta}) = \sum_{ij} \int_0^t \frac{G_i \mathrm{d}N_{ij}^*(s, \boldsymbol{\beta})}{\sum_{kl} Y_{kl}^*(s, \boldsymbol{\beta})}.$$

Potom stejně sestavený replikovaný proces $\widehat{W}(t)$ má opět stejnou limitu jako $W(t)$ v AFT modelu. Postup je založen na funkcionální CLV [4] a inferenci pro AFT model s rekurentními událostmi [6].

Když tedy zreplikujeme mnohokrát $\widehat{W}(t)$, můžeme empiricky odhadnout rozptyl $W(t)$ a spočítat bodové konfidenční intervaly kumulovaného rizika jako

$$\widehat{\Lambda}_0(t, \widehat{\boldsymbol{\beta}}) \pm u_{1-\alpha/2} n^{-1/2} \sqrt{\widehat{\mathrm{var}} \widehat{W}(t)},$$

nebo pomocí log-transformace jako $\widehat{\Lambda}_0(t, \widehat{\boldsymbol{\beta}}) \exp\left(\pm u_{1-\alpha/2} n^{-\frac{1}{2}} \frac{\sqrt{\widehat{\mathrm{var}} \widehat{W}(t)}}{\widehat{\Lambda}_0(t, \widehat{\boldsymbol{\beta}})} \right)$, kde $u_{1-\alpha/2}$ je příslušný kvantil $N(0,1)$.

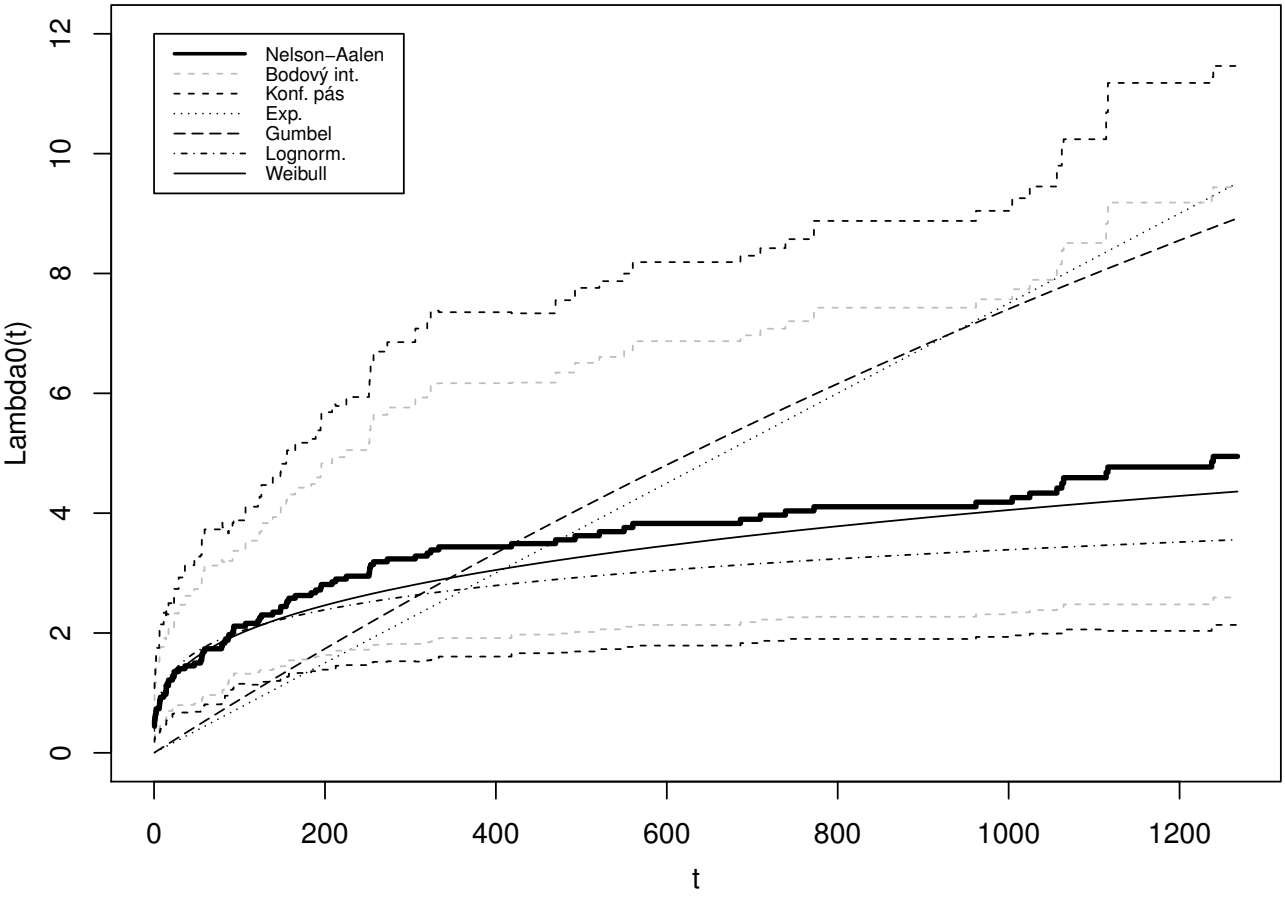
Testování hypotéz o tvaru základního rizika

Chceme-li testovat hypotézy o tvaru celého rizika, je potřeba najít příslušný konfidenční pás pro supremový test. Najdeme $q_{1-\alpha}$ vyběrový $1-\alpha$ kvantil generovaných hodnot $\sup_{[\tau_1, \tau_2]} \left| \frac{\widehat{W}(t)}{\sqrt{\widehat{\mathrm{var}} \widehat{W}(t)}} \right|$, kde $[\tau_1, \tau_2]$ pokrývá zkoumanou část časového intervalu a spočítáme konfidenční pás pomocí logaritmické transformace jako

$$\widehat{\Lambda}_0(t, \widehat{\boldsymbol{\beta}}) \exp\left(\pm q_{1-\alpha} n^{-1/2} \frac{\sqrt{\widehat{\mathrm{var}} \widehat{W}(t)}}{\widehat{\Lambda}_0(t, \widehat{\boldsymbol{\beta}})} \right).$$

Hypotézu zamítáme, pokud testovaná kumulovaná základní riziková funkce neleží v konfidenčním pásu. Na Obrázku 1 vidíme příklad Nelson-Aalenova odhadu (tučně černě) pro data o rozsahu $n = 20$ z Coxova modelu s $\varphi = 1/10$,

$\rho = -1/10$ a Weibullovým základním rozdělením s $a = 1/2$ a $\lambda = 1/10$. Dále jsou zobrazeny příslušné bodové intervaly spolehlivosti (čárkovaně šedě), konfidenční pás (čárkovaně černě) a parametrické odhady pro různá rozdělení (šedě). V tomto případě bychom jasně zamítali exponenciální rozdělení, kde kumulovaná riziková funkce tvoří přímku (tečkovaně), i Gumbelovo rozdělení (čárkovaně). Naopak parametrický odhad původního Weibullova rozdělení (plně) je Nelson-Aalenovu odhadu velmi blízko, nezamítali bychom patrně ani lognormální rozdělení (čerchovaně).



Obrázek 1: Porovnání Nelson-Aalenova odhadu, konfidenčních mezí a parametrických odhadů kumulované rizikové funkce.

4. Simulační studie

Generovali jsme data z Coxova i AFT modelu o velikosti $n = 20$ a $n = 50$ s různými základními rizikovými funkcemi a parametry. Každé zařízení bylo sledováno do desáté události, údržba byla prováděna náhodně se stejným základním rozdělením. Parametry jsme stanovili tak, aby oprava zvýšila riziko

Tabulka 1: Podíl zamítnutých hypotéz o tvaru základního rizika při generování dat z různých rozdělení.

Generované rozdělení					Testované rozdělení – podíl zamítnutí			
Model	λ_0	λ	a	n	Exp.	Weibull	Gumbel	LN
Cox	Weibull	1/10	5	20	0,908	0	0	0,216
				50	1	0	0	0,276
	Weibull	1/10	1/2	20	0,934	0	0,916	0,036
				50	1	0	1	0,134
	Gumbel	1/10	1,2	20	0,096	0,008	0	0,272
				50	0,642	0,020	0	0,640
	LN	μ=2	σ^2=4	20	0,992	0	1	0
				50	1	0,082	1	0
AFT	Weibull	1/10	5	20	0,912	0,006	0,008	0,066
				50	1	0	0	0,492
	Weibull	1/10	1/2	20	0,904	0,002	0,796	0,088
				50	1	0	1	0,644
	Gumbel	1/10	1,2	20	0,372	0,014	0,008	0,592
				50	0,990	0,050	0	0,994
	LN	μ=2	σ^2=4	20	0,996	0,142	0,878	0,092
				50	1	0,022	1	0

či zrychlila čas ($\varphi = 1/10$) a údržba naopak ($\rho = -1/10$), jiné kovariáty nebyly uvažovány.

Testovali jsme na hladině $\alpha = 0{,}05$, zda je základní rozdělení exponenciální, Weibullovo $\lambda(t) = a\lambda^a t^{a-1}$, useknuté Gumbelovo $\lambda(t) = \lambda a^t$ či lognormální (LN) s parametry odhadnutými metodou maximální věrohodnosti původního modelu považovanými za pevné a sledovali jsme podíl zamítnutých hypotéz. Každý případ byl simulován 500×, testovali jsme na intervalu mezi 5% a 95% kvantilem generovaných dat. $\widehat{W}(t)$ bylo počítáno ze 40 replikací.

Z tabulky výsledků 1 je patrné, že testy jsou s vyšším počtem pozorovaných zařízení přesnější, tj. nezamítají původní a zamítají ostatní základní rozdělení. U dat s Weibullovým základním rozdělením záleží, zda je Λ_0 konvexní či konkávní, podle toho je spíše zaměnitelné s Gumbelovým nebo log-

normálním rozdělením. Exponenciální rozdělení je zamítáno skoro vždy – zde se nabízí srovnání s parametrickým testem zda $a = 1$ ve Weibullově rozdělení.

5. Závěr

Zkoumali jsme metody pro testování hypotéz o tvaru základního rizika při modelování vlivu údržby a oprav na životnost sledovaného zařízení. Pro data z Coxova modelu i modelu zrychleného času jsme představili asymptotický test založený na resamplingu a na simulovaných datech zkoumali jeho vlastnosti v různých situacích. Dalším krokem může být zohlednění variability testovaných parametrických odhadů.

Poděkování

Tato práce byla podporována granty SVV 260105/2014 a GAUK 11122/2013.

Literatura

[1] Percy D. F., Alkali B. M.: Generalized proportional intensities models for repairable systems. *IMA Journal of Management Mathematics* **17**, 171–185, 2005.

[2] Lin D. Y., Ying Z.: Semiparametric inference for the accelerated life model with time-dependent covariates. *Journal od Statistical Planning and Inference* **44**, 47–63, 1995.

[3] Novák P.: Regrese v modelech oprav. *Informační bulletin České statistické společnosti* **24** (3–4), 83–88, 2013.

[4] Pollard D.: *Empirical Processes: Theory and Applications*. Hayward, California, 1990.

[5] Lin D. Y., Wei L. J., Ying Z.: Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika* **80**, 557–572, 1993.

[6] Lin D. Y., Wei L. J., Ying Z.: Accelerated failure time models for counting processes. *Biometrika* **85**, 605–618, 1998.

STOCHASTICKÉ MODELOVANIE VEĽKÝCH ŠKÔD V POISŤOVNÍCTVE

STOCHASTIC MODELLING LARGE CLAIMS IN INSURANCE

Zuzana Rošťáková

Adresa: Ústav merania SAV, Dúbravská cesta 4, 841 04 Bratislava 4

E-mail: zuzana.rostakova@gmail.com

Abstrakt: Veľké škody tvoria v neživotnom poistení len malú časť z celkového počtu poistných udalostí. Na druhej strane, ich príspevok v konečnej sume poistných plnení je pomerne vysoký. Vo všeobecnosti ich môžeme chápať ako extrémne pozorovania. V tomto príspevku sa zaoberáme stochastickým modelovaním škôd s vysokým poistným plnením. Vhodným nástrojom na modelovanie ťažkých chvostov je zovšeobecnené rozdelenie veľkých hodnôt (GEV distribution). Toto rozdelenie umožňuje definovať „prah“ ako hranicu medzi veľkými a zanedbateľnými škodami. V centre záujmu príspevku stojí metóda POT – Peaks Over Threshold, ktorá je založená práve na voľbe vhodného prahu pomocou zovšeobecneného Paretovho rozdelenia. Na odhady potrebných parametrov je použitá metóda maximálnej vierohodnosti a vážená momentová metóda. Vybudovaná teória je ilustrovaná na reálnych pozorovaniach.

Kľúčová slová: Rozdelenie s ťažkým chvostom, QQ-graf, zovšeobecnené rozdelenie veľkých hodnôt GEV, zovšeobecnené Paretovo rozdelenie GPD, metóda POT.

Abstract: Large claims in non-life insurance comprise only a small part of the overall number of claims. Nevertheless, they contribute a significant portion to the overall claim amounts. Generally, the large claims tend to be outliers in the experience. This article deals with stochastic modelling such claims with contributory negligence. The general extreme value (GEV) distribution is proposed as a suitable way for modelling the heavy tails. A threshold value to distinguish large and attritional claims is defined within the GEV framework and its connection to the generalized Pareto distribution is derived. The main aim of this article concerns method for the threshold selection—peaks over threshold (POT) method. Within this approach, the maximum likelihood method and the method of weighted moments are used for parameter estimation. Finally, real data examples are provided as an illustration of the potential benefits of the presented techniques.

Keywords: Heavy-tailed distributions, QQ-plot, generalized extreme value distribution, GEV, generalised Pareto distribution, GPD, POT method.

1. Úvod

Veľké škody tvoria v neživotnom poistení nemalú časť nákladov, ktoré musí poisťovňa vyplatiť svojim klientom ako odškodné pri poistných udalostiach. Môžeme ich teda chápať ako extrémne pozorovania. Z hľadiska poisťovne je dôležité vedieť (aspoň približne) odhadnúť ich budúcu výšku.

Pri modelovaní výšky poistných plnení sa často používajú najmä kladné rozdelenia s (pravým) ťažkým chvostom. Avšak pri odhade pravdepodobnosti výskytu extrémne vysokých poistných plnení tieto modely, ako aj klasické štatistické metódy, zlyhávajú, nakoľko sa môžu oprieť len o malý počet dát. Z tohto dôvodu bolo vytvorených viacero štatistických postupov, ktoré môžeme použiť pri modelovaní veľkých škôd. V tomto článku sme sa zamerali na metódu POT – Peaks Over Threshold.

Výhodou tejto metódy je, že funguje spoľahlivo aj pri menšom počte pozorovaní. Pomocou nej nielenže môžeme veľké škody modelovať, ale aj odhadnúť maximálnu výšku budúcich škôd.

2. Metóda POT

2.1. Funkcia priemerného prírastku

Ako už napovedá samotný názov metódy – *Peaks Over Threshold*, pri modelovaní veľkých škôd sme uvažovali len pozorovania prekračujúce určitú pevne zvolenú hranicu. Otázka znie, čo znamená „vhodne zvolená“. Hranica u nesmie byť príliš nízka, pretože potom by nemuseli fungovať zvolené metódy odhadu distribúcie veľkých škôd, príliš vysokú hodnotu by zas mohol prekročiť len malý počet pozorovaní.

Vhodným nástrojom na voľbu vhodného prahu u je ME-graf, založený na funkcii priemerného prírastku, a TC-grafy (z anglického *Threshold Choice*), viac v [8].

Definícia 2.1 (Funkcia priemerného prírastku)**.** *Nech X je náhodná premenná s pravým koncovým bodom x_F, nech $u \in \mathbb{R}^+$, $u < x_F$ je ľubovoľné pevné. Distribučná funkcia presahu F_u pre náhodnú premennú X je definovaná ako*

$$F_u(x) = P(X - u \leq x | X > u), \quad x \in \mathbb{R}^+. \tag{1}$$

Funkciu priemerného prírastku náhodnej premennej X potom definujeme ako podmienenú strednú hodnotu

$$e(u) = E(X - u|X > u), \quad 0 \leq u < x_F. \tag{2}$$

Uvažujme náhodný výber $X_1, X_2, \ldots, X_n$ rozsahu n z rozdelenia s distribučnou funkciou F. Nech $X_{(1),n} < X_{(2),n} < \cdots < X_{(n),n}$ je príslušný usporiadaný náhodný výber. Označme F_n zodpovedajúcu empirickú distribučnú funkciu a nech $\Delta_n(u) = \{i,\ i = 1, \ldots, n :\ X_i > u\}$, $u \in \mathbb{R}^+$ a $N_u = |\Delta_n(u)|$. Potom výberový ekvivalent funkcie priemerného prírastku má tvar:

$$e_n(u) = \frac{1}{1 - F_n(u)} \int_u^\infty [1 - F_n(y)]\mathrm{d}y = \frac{1}{N_u} \sum_{i \in \Delta_n(u)} (X_i - u), \tag{3}$$

Pomocou tejto funkcie je potom možné zostrojiť ME-graf:

$$\left\{ \left(X_{(k),n}, e_n \left(X_{(k),n} \right) \right) : k\ = 1, \ldots, n \right\}. \tag{4}$$

V praxi na os x nanášame usporiadané dáta, na os y príslušné hodnoty funkcie e_n.

2.2. GEV a GPD – základy metódy POT

Metóda POT je založená na dvoch rozdeleniach pravdepodobnosti. Ide o zovšeobecnené rozdelenie veľkých hodnôt (*GEV*) a zovšeobecnené Paretovo rozdelenie (*GPD*).

Nech H_ξ je distribučná funkcia náhodnej premennej X. Budeme hovoriť, že X má zovšeobecnené rozdelenie veľkých hodnôt alebo *generalized extreme value distribution* (GEV), ak jej distribučná funkcia je definovaná nasledovne:

$$H_\xi(x) = \begin{cases} \mathrm{e}^{-(1+\xi x)^{-\frac{1}{\xi}}}, & \xi \neq 0; \\ \mathrm{e}^{-\mathrm{e}^{-x}}, & \xi = 0; \end{cases} \tag{5}$$

pričom $1 + \xi x > 0$ a

$$\begin{aligned} x > -\xi^{-1}, &\quad \text{ak} \quad \xi > 0; \\ x < -\xi^{-1}, &\quad \text{ak} \quad \xi < 0; \\ x \in \mathbb{R}, &\quad \text{ak} \quad \xi = 0. \end{aligned}$$

S GEV rozdelením úzko súvisí pojem tzv. *maximum domain of attraction (MDA)*. Vlastnosti, ako aj ďalšie informácie o *MDA*, je možné nájsť v [1], str. 131–150.

Definícia 2.2. *Nech $X_1, X_2, \ldots, X_n$ je postupnosť i.i.d. náhodných premenných s distribučnou funkciou F, nech X je všeobecné označenie pre členy tejto postupnosti. Budeme hovoriť, že náhodná premenná X (jej distribučná funkcia F) patrí do maximum domain of attraction rozdelenia extrémnych hodnôt H, ak existujú konštanty $a_n > 0$, $b_n \in \mathbb{R}$ také, že*

$$a_n^{-1}(M_n - b_n) \xrightarrow{d} H,$$

pričom $M_n = \max\{X_1, \ldots, X_n\}$.

Druhým dôležitým rozdelením je zovšeobecnené Paretovo rozdelenie. Distribučná funkcia tohto rozdelenia závisí od parametrov ξ, β ($\xi \in \mathbb{R}$, $\beta > 0$) a možno ju zapísať v tvare

$$G_{\xi,\beta}(x) = \begin{cases} 1 - \left(1 + \xi\dfrac{x}{\beta}\right)^{-\frac{1}{\xi}}, & \xi \neq 0; \\ 1 - \mathrm{e}^{-x}, & \xi = 0; \end{cases} \tag{6}$$

pričom

$$\begin{aligned} x \geq 0, \quad &\text{ak} \quad \xi \geq 0; \\ 0 \leq x \leq -\frac{\beta}{\xi}, \quad &\text{ak} \quad \xi < 0. \end{aligned}$$

V nasledujúcej vete sme uviedli ekvivalentnú podmienku toho, že $X \in \mathrm{MDA}(H_\xi)$. Náčrt dôkazu je možné nájsť v [1], str. 165.

Tvrdenie 2.1 (Niektoré vlastnosti GPD)**.** *Nech X je náhodná premenná s distribučnou funkciou F.*

a) *Pre všetky $\xi \in \mathbb{R}$ platí nasledujúca ekvivalencia:*

$$F \in \mathrm{MDA}(H_\xi) \iff \lim_{u \uparrow x_F} \sup_{0 < x < x_F - u} |F_u(x) - G_{\xi,\beta(u)}(x)| = 0, \tag{7}$$

$\beta(x)$ je kladná funkcia.

b) *Nech sú náhodné premenné $N, X_1, \ldots, X_n, \ldots$ nezávislé, $N \sim \mathrm{Poi}(\lambda)$ a $X_i \sim G_{\xi,\beta}$. Označme $M_N = \max\{X_1, \ldots, X_N\}$. Potom*

$$P(M_N < x) = \mathrm{e}^{-\lambda\left(1+\xi\frac{x}{\beta}\right)^{-\frac{1}{\xi}}} = H_\xi\left(\frac{x - \beta\xi^{-1}\left(\lambda^\xi - 1\right)}{\beta\lambda^\xi}\right).$$

Na základe Tvrdenie 2.1 možno pre dostatočne vysokú hodnotu prahu u aproximovať F_u pomocou $G_{\xi,\beta(u)}$.

Vráťme sa ale späť k funkcii priemerného prírastku. Pomerne jednoduchými výpočtami sme odvodili jej tvar pre GPD s parametrami ξ, β:

1. $\xi = 0$:

$$e(u) = \frac{\int_u^\infty \mathrm{e}^{-x}\mathrm{d}x}{\mathrm{e}^{-u}} = -\frac{0 - \mathrm{e}^{-u}}{\mathrm{e}^{-u}} = 1.$$

2. $\xi \neq 0$:

$$e(u) = \frac{\int_u^\infty \left(1+\xi\frac{x}{\beta}\right)^{-\frac{1}{\xi}}\mathrm{d}x}{\left(1+\xi\frac{u}{\beta}\right)^{-\frac{1}{\xi}}} = \frac{\frac{\beta}{\xi-1}\left[\lim_{x\to\infty}\left(1+\xi\frac{x}{\beta}\right)^{1-\frac{1}{\xi}} - \left(1+\xi\frac{u}{\beta}\right)^{1-\frac{1}{\xi}}\right]}{\left(1+\xi\frac{u}{\beta}\right)^{-\frac{1}{\xi}}}.$$

Uvedená limita sa dá zapísať do tvaru:

$$\lim_{x\to\infty}\left[\left(1+\xi\tfrac{x}{\beta}\right)^{-\frac{1}{\xi}}\left(1+\xi\tfrac{x}{\beta}\right)\right] = \lim_{x\to\infty}\left(1+\xi\tfrac{x}{\beta}\right)^{-\frac{1}{\xi}} + \tfrac{\xi}{\beta}\lim_{x\to\infty}\left[x\left(1+\xi\tfrac{x}{\beta}\right)^{-\frac{1}{\xi}}\right]. \tag{8}$$

Využívajúc základné vlastnosti distribučných funkcií

$$\lim_{x\to\infty}\left[1-F(x)\right] = 0,$$
$$\lim_{x\to\infty} x\left[1-F(x)\right] = 0,$$

sme odvodili, že obe limity uvedené v 8 sú rovné 0. Po menších úpravách sme potom dostali finálny tvar funkcie priemerného prírastku pre GPD.

$$e(u) = \frac{\frac{\beta}{\xi-1}\left(1+\xi\frac{u}{\beta}\right)^{1-\frac{1}{\xi}}}{\left(1+\xi\frac{u}{\beta}\right)^{-\frac{1}{\xi}}} = \frac{\beta}{1-\xi} + \frac{\xi}{1-\xi}u. \tag{9}$$

Po záverečných úpravách je vidieť, že funkcia priemerného prírastku pre GPD s parametrami ξ, β je vždy lineárna funkcia. Tento poznatok je základom pre grafický odhad dostatočne veľkého prahu u. Uvažujme vzorku $X_1, \ldots, X_n$. Zostrojíme empirical mean excess function pre uvedené dáta a budeme hľadať takú hodnotu $u \in \mathbb{R}^+$, pre ktorú je $e_n(x)$ pre $x > u$ približne lineárna funkcia. Pre túto hodnotu u je potom na základe bodu a) v Tvrdenie 2.1 prirodzené vziať GPD ako odhad pre F_u (Definícia 2.1).

2.3. Metóda POT

V tejto časti sme sa zamerali na samotnú metódu POT – nielen na jej matematickú podstatu ako je uvedená v [1] alebo v [2], ale snažili sme sa popísať aj postup jej aplikácie.

Uvažujme postupnosť nezávislých, rovnako rozdelených náhodných premenných $X_1, \ldots, X_n$ s distribučnou funkciou F. Nech pre nejaké $\xi \in \mathbb{R}$ je $F \in \mathrm{MDA}(H_\xi)$ a nech u je pevne zvolený, dostatočne vysoký prah. Nech rovnako ako v časti 2.1. označuje N_u počet presahov cez prah u.

Našou úlohou bolo nájsť odhad pre pravý chvost $1-F(u+x)$ distribučnej funkcie F, $x \geq 0$. Využili sme pri tom vyjadrenie $1-F(u+x)$ pomocou $F_u(x)$.

$$1 - F(u + x) = [1 - F(u)]\,[1 - F_u(x)] \tag{10}$$

Vďaka tomuto vyjadreniu sme mohli odhadnúť chvost $1 - F(u)$ tak, že sme postupne odhadli $1 - F_u(x)$ a $1 - F(u)$. Ako odhad pre $F(u)$ sme použili empirickú distribučnú funkciu $F_n(u)$.

$$1 - F(u) \approx 1 - F_n(u) = \frac{1}{n} \sum_{i=1}^{n} I_{\{X_i > u\}} = \frac{N_u}{n} \tag{11}$$

Na začiatku tejto časti sme predpokladali, že $F \in \mathrm{MDA}(H_\xi)$, čo je podľa Tvrdenie 2.1 ekvivalentné s tým, že

$$\lim_{u \uparrow x_F} \sup_{0 < x < x_F - u} |F_u(x) - G_{\xi,\beta(u)}(x)| = 0,$$

pričom $\beta(u)$ je kladná funkcia. Teda funkciu $F_u(x)$ sme mohli pre dostatočne veľkú hodnotu u aproximovať distribučnou funkciou zovšeobecneného Paretovho rozdelenia

$$F_u(x) \approx G_{\xi,\beta(u)}(x). \tag{12}$$

V praxi hodnoty neznámych parametrov nahradíme ich odhadmi $\widehat{\xi}$ a $\widehat{\beta} = \widehat{\beta}(u)$, kde u je pevne zvolený prah. Teda neodhadujeme celú funkciu $\beta(x)$, ale len odhad pre jej hodnotu v bode u. Vhodnými nástrojmi odhadu sú metóda maximálnej vierohodnosti, alebo zovšeobecnená momentová metóda. Odvodenie odhadov parametrov ξ, $\beta(u)$ pomocou oboch metód možno nájsť v [1], str. 356–358.

Chvost $1 - F(u + x)$ sme teda mohli na základe vzťahu 10 aproximovať v tvare

$$1 - F(u + x) \approx \frac{N_u}{n} \left(1 + \widehat{\xi}\frac{x}{\widehat{\beta}}\right)^{-\frac{1}{\widehat{\xi}}}. \tag{13}$$

2.4. Odhad maximálnej výšky budúcich škôd

Veľký význam v poisťovníctve má aj odhad výšky maximálnej škody v budúcnosti. Ide o tzv. *Probable Maximum Loss* alebo *PML*. Existujú rôzne definície pre PML, niektoré z nich sú uvedené v [2], str. 27. V tomto článku sme využili výpočet PML založený na metóde POT a GPD.

PML môžeme získať vyriešením rovnice

$$P\left(M_n \leq \mathrm{PML}_\varepsilon\right) = 1 - \varepsilon, \tag{14}$$

čo je ekvivalentné vyjadreniu

$$\mathrm{PML}_\varepsilon = F^{\leftarrow}_{M_n}(1-\varepsilon), \tag{15}$$

kde M_n je maximum z analyzovaných dát za určitý časový úsek a F_{M_n} je distribučná funkcia náhodnej premennej M_n. $F^{\leftarrow}_{M_n}(1-\varepsilon)$ je $(1-\varepsilon)$-kvantil distribučnej funkcie náhodnej premennej M_n, pričom $\varepsilon > 0$ je pevne zvolená konštanta.

Naším cieľom však bolo odhadnúť budúce maximálne škody a teda M_n, resp. jeho rozdelenie, sme nepoznali. Potrebovali sme teda vytvoriť odhad pre M_n. Podľa [2], str. 27, ak distribučnú funkciu presahov nad prahom u môžeme aproximovať pomocou GPD rozdelenia s parametrami ξ, β, pričom $\xi \neq 0$, tak náhodná premenná N, charakterizujúca počet presahov nad prahom u, má Poissonovo rozdelenie s parametrom λ. Dôkaz je možné nájsť v [2], str. 34. Potom na základe Tvrdenie 2.1, b) rozdelenie náhodnej premennej M_N môžeme aproximovať pomocou zovšeobecneného rozdelenia veľkých hodnôt GEV, teda:

$$P(M_N < x) = H_\xi\left(\frac{x - \beta\xi^{-1}\left(\lambda^\xi - 1\right)}{\beta\lambda^\xi}\right). \tag{16}$$

Pomocou inverznej distribučnej funkcie k funkcii 16 je už jednoduché odvodiť zápis na výpočet PML_ε:

$$\mathrm{PML}_\varepsilon = \left[\left(-\frac{\lambda}{\ln(1-\varepsilon)}\right)^\xi - 1\right]\frac{\beta}{\xi} + u. \tag{17}$$

3. Aplikácia metódy POT

Pri aplikácii vyššie uvedenej metódy sme využívali dáta zo *SOA Group Medical Insurance Large Claims Database* z rokov 1997 až 1999. Výpočty prebiehali v prostredí softwaru R.

Každá z analyzovaných databáz obsahovala údaje o viac než 1 200 000 poistných udalostiach. Pri jednotlivých pozorovaniach boli okrem iného uvedené údaje o poistencovi, napríklad rok narodenia, pohlavie, diagnóza a mnohé iné. Pre nás však boli najdôležitejšie informácie o celkovej výške jednotlivých škôd. Ich výška sa pohybovala od veľmi nízkych čiastok po miliónové sumy. Ako však napovedá už názov, pre nás boli dôležité „veľké škody“ a preto sme v ďalších analýzach uvažovali len sumy vyššie ako $ 10 000.

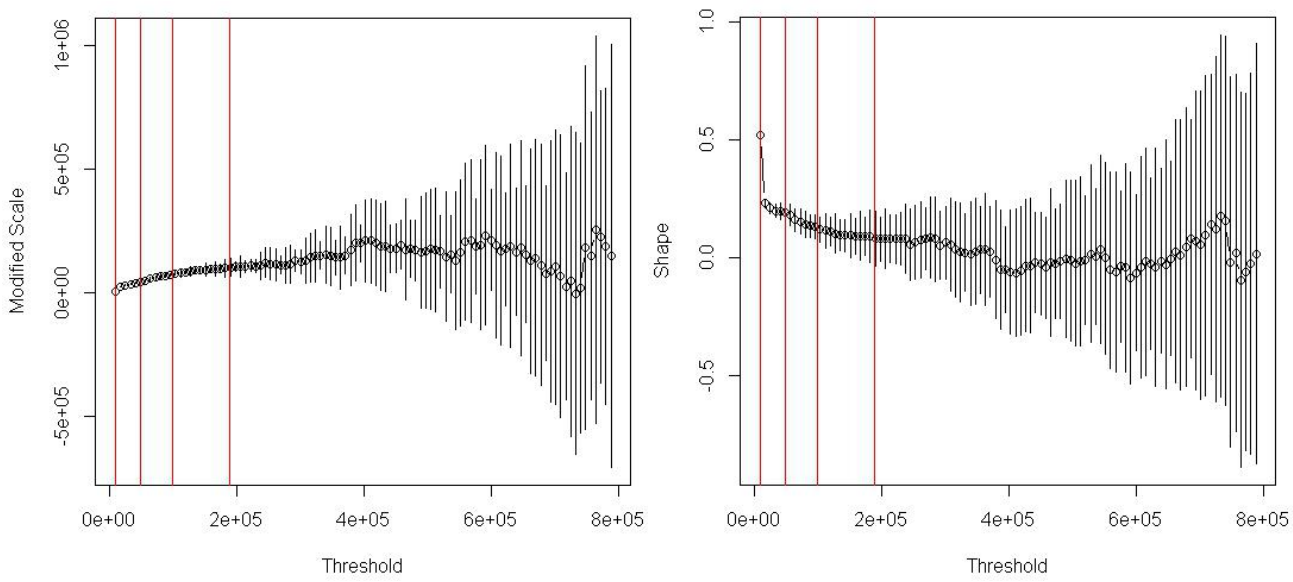
V roku 1997 dosiahla celková výška škôd hodnotu $ 2 003 162 218 pri celkom 1 241 438 poistných udalostiach, pričom hranicu $ 10000 prekročilo 33 325 pozorovaní. Maximálna škoda presiahla výšku $ 1 200 000. Oveľa zaujímavejšie boli hodnoty aritmetického priemeru a tretieho kvartilu. Priemer pozorovaní prevyšujúcich $ 10 000 sa rovnal $ 27 799,64, zatiaľ čo 75 % pozorovaní bolo nižších ako $ 27 397,69. Tento fakt nás viedol k predpokladu, že pozorovania by mohli pochádzať z rozdelenia s ťažkým chvostom. Potvrdila nám to aj kladná hodnota koeficientu šikmosti. Podrobnejšie údaje sú uvedené v Tabuľke 1.

Tabuľka 1: Základné štatistiky pre škody nad $ 10 000 (1997).

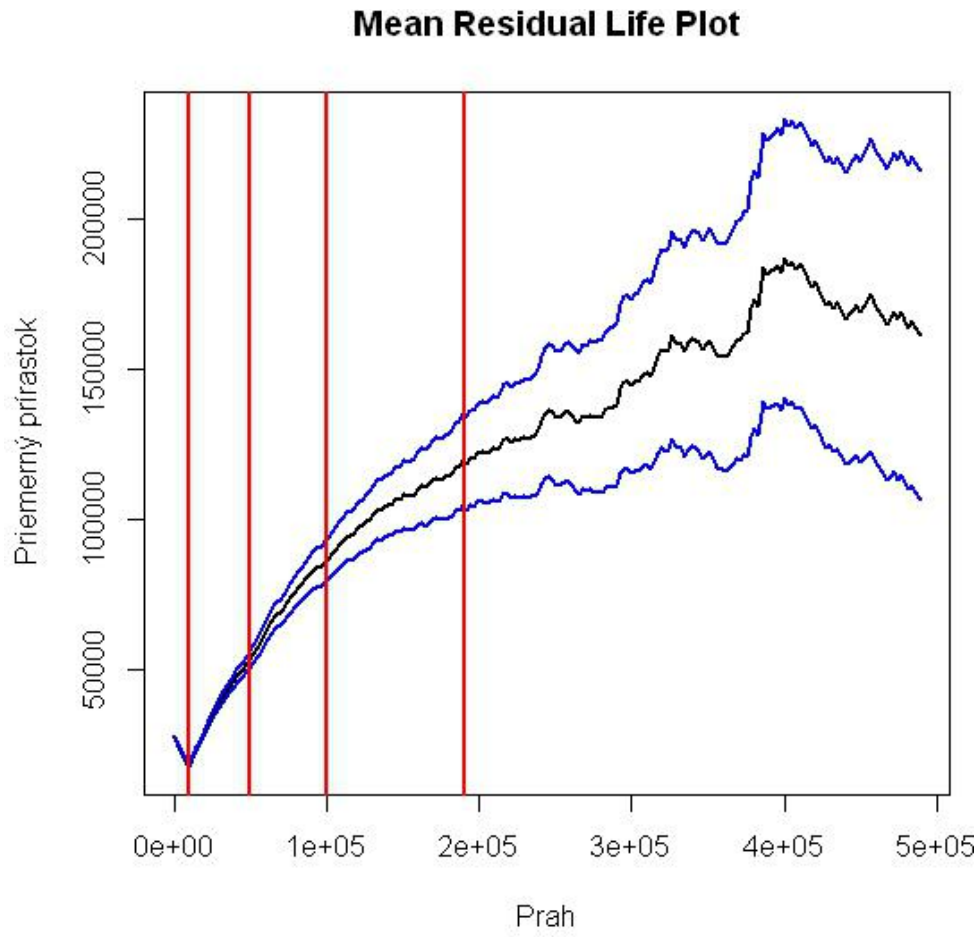
Celková výška škôd	$ 926 422 888,000
Maximum	$ 1 225 908,300
Minimum	$ 10 000,190
Výberový 25 %-ný kvantil	$ 12 295,840
Medián	$ 16 618,000
Výberový 75 %-ný kvantil	$ 27 397,690
Priemer	$ 27 799,640
IQR	$ 15 101,850
Šikmosť	8,350
Špicatosť	124,020
Smerodajná odchýlka	$ 39 111,620

Podľa postupu uvedeného v predchádzajúcej časti sme sa snažili pomocou ME-grafu a TC-grafov nájsť vhodnú hodnotu prahu u. Na základe Obrázku 1 sme za naše prahy určili hranice $ 10 000, $ 50 000, $ 100 000 a $ 190 000. Voľby týchto prahov nám odobril aj ME-graf.

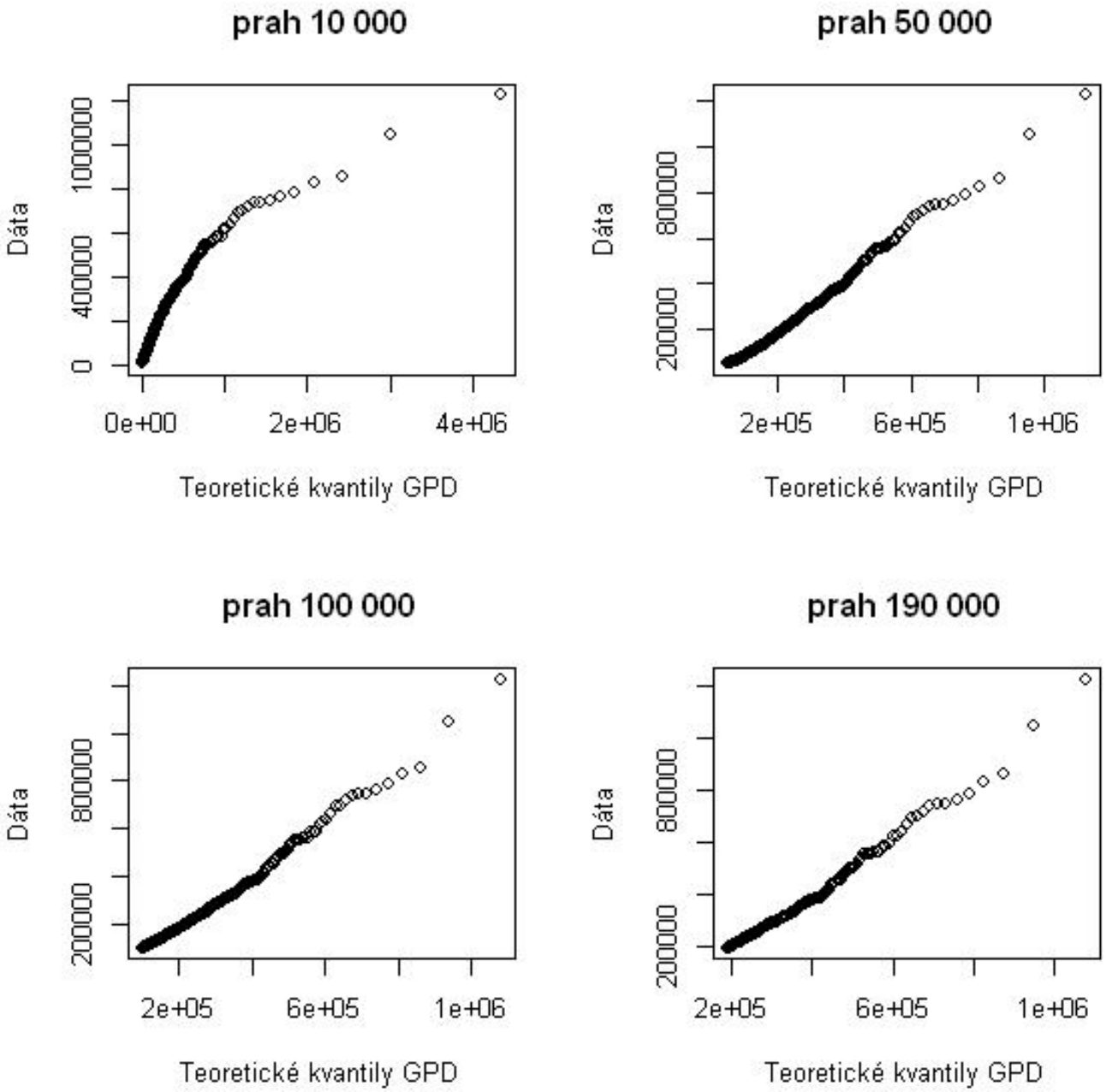
Pre vyššie uvedené prahy sme vypočítali odhady neznámych parametrov ξ a β v zovšeobecnenom Paretovom rozdelení pomocou metódy maximálnej vierohodnosti.



Obr. 1: TC-grafy.



Obr. 2: ME-graf.



Obr. 3: QQ-grafy pre rôzne hodnoty prahov u.

Mali sme teda vytvorené štyri modely pre naše pozorovania. Najskôr sme ich kvalitu otestovali pomocou QQ-grafov. Na os x sme nanášali teoretické kvantily príslušného GPD rozdelenia, na os y zas usporiadané dáta prevyšujúce prah u.

Z Obrázku 3 je zrejmé, že model pre pozorovania prevyšujúce prah \$ 10 000 nevyhovuje, nakoľko je očividný konkávny tvar grafu. Ostatné grafy majú približne lineárny priebeh. Medzi týmito modelmi sme sa rozhodli na základe Kolmogorovho-Smirnovho testu. Testovali sme hypotézu, že dáta prevyšu-

júce prah u pochádzajú z GPD s parametrami ξ, β. p-value testu prekročila hodnotu 5 % len v prípade modelu pre prah \$ 190 000.

Samozrejme, tvrdenie, že ide o model, ktorý najlepšie popisuje naše dáta, nie je namieste. Voľba prahu sa totižto zakladá len na grafických odhadoch a subjektívnom názore pozorovateľa, čo podľa neho znamená pojem „približne lineárny tvar“. Iný pozorovateľ by si na základe TC-grafov a ME-grafu mohol zvoliť iné východiskové hodnoty prahov a testoval by tie. My sme ďalej pokračovali so zvolenou hodnotou \$ 190 000.

Na základe vytvoreného modelu a vzťahu 17, sme chceli odhadnúť maximálne škody v roku 1998 a tento výsledok porovnať so skutočnými hodnotami získanými v tomto roku. Za ε sme si zvolili hodnoty 5 % a 1 %. Odhady neznámych parametrov pre distribučnú funkciu príslušného GPD mali hodnotu $\widehat{\xi} = 8{,}016 \cdot 10^{-2}$ a $\widehat{\beta} = 1{,}199 \cdot 10^{5}$. Ako odhad pre hodnotu λ sme použili počet presahov nad prahom u, t.j. 332.

$$\mathrm{PML}_{0,05} = 1\,716\,680 \tag{18}$$

$$\mathrm{PML}_{0,01} = 2\,138\,543 \tag{19}$$

S pravdepodobnosťou 95 % teda výška poistných udalostí v roku 1998 neprekročí \$ 1 716 680 a s pravdepodobnosťou 99 % budú škody nižšie ako \$ 2 138 543. Žiadna z poistných udalostí v roku 1998, maximum nevynímajúc, nepresiahla naše odhady $\mathrm{PML}_{0,05}$ a $\mathrm{PML}_{0,01}$. Maximum z dát v roku 1999 dosiahlo výšku \$ 2 568 512. Toto číslo je vyššie než oba naše odhady. Na druhej strane, táto škoda ako jediná presiahla hranicu $\mathrm{PML}_{0,01}$. Odhad $\mathrm{PML}_{0,05}$ bol prekročený len v troch prípadoch, konkrétne pri pozorovaniach s hodnotou \$ 1 763 385, \$ 1 838 290 a \$ 2 568 512. Po následnej analýze sme zistili, že škody v roku 1999 vykazovali ťažšie chvosty než v roku 1997.

4. Záver

V príspevku sme sa snažili o vytvorenie modelu popisujúceho správanie sa veľkých škôd. Využili sme pri tom metódu POT založenú na zovšeobecnenom rozdelení veľkých hodnôt a zovšeobecnenom Paretovom rozdelení. Teóriu sme následne aplikovali na sadu reálnych pozorovaní z roku 1997. Na základe vytvoreného modelu sme dohadli výšku budúcich škôd a odhady sme porovnali so skutočnými hodnotami. Aj napriek tomu, že odhady budúcich škôd v roku 1999 nie celkom vyšli, mohli sme vytvorený model považovať za funkčný a schopný pracovať s veľmi veľkými hodnotami.

Literatúra

[1] Embrechts P., Klüppelberg C., Mikosch T.: *Modelling Extremal Events for Insurance and Finance*. Springer-Verlag, 1997.

[2] Cebrián A. C., Denuit M., Lambert P.: Generalized Pareto Fit to the Society of Actuaries' Large Claims Database. *North American Actuarial Journal*, **3**, 2003.

[3] Coles S.: *An Introduction of Statistical Modeling of Extreme Values*. Springer-Verlag, 2001.

[4] Michalíková J.: *Rozdelenia s ťažkými chvostami v neživotnom poistení*. Bratislava: Univerzita Komenského. Fakulta matematiky, fyziky a informatiky, 2009. Diplomová práca. Vedúci diplomovej práce: doc. RNDr. Rastislav Potocký, CSc.

[5] Janková K., Pázman A.: *Pravdepodobnosť a štatistika*. 1. vyd. Bratislava: Vydavateľstvo UK, 2011.

[6] Lamoš F., Potocký R.: *Pravdepodobnosť a matematická štatistika: Štatistické analýzy*. 2. vyd. Bratislava: Vydavateľstvo UK, 1998.

[7] Society of Actuaries: *Medical Large Claims Experience Study.* [online] Publikované: september 2004. [Citované 15. 3. 2013] URL: `http://www.soa.org/Research/Experience-Study/Group-Health/research-medical-large-claims-experience-study.aspx`.

[8] Ribatet M. A.: *A User's Guide to the POT Package.* [online] Verzia 1.4 (2011). Publikované: august 2006. [Citované 20. 3. 2013] URL: `http://cran.r-project.org/web/packages/POT/vignettes/POT.pdf`.

KLASIFIKACE NA ZÁKLADĚ HLOUBKY DAT – GLOBÁLNÍ A LOKÁLNÍ PŘÍSTUPY

CLASSIFICATION BASED ON DATA DEPTH – GLOBAL AND LOCAL APPROACHES

Ondřej Vencálek

Adresa: Katedra matematické analýzy a aplikací matematiky, PřF, Univerzita Palackého v Olomouci, 17. listopadu 12, 771 46 Olomouc

E-mail: ondrej.vencalek@upol.cz

Abstrakt: Uspořádání bodů ve vícerozměrném prostoru (vzhledem k nějakému rozdělení) pomocí hloubky dat může být základem pro řešení mnoha statistických úloh. Jednou z nich je i úloha klasifikace – tvorby pravidla pro zařazení nového pozorování do jedné z několika skupin, jejichž reprezentanty pozorujeme. Během posledních přibližně deseti let bylo navrženo několik klasifikátorů využívajících hloubku dat. Cílem příspěvku je přehledně shrnout práci v oblasti klasifikace na základě hloubky dat a ukázat nové trendy v této oblasti.

Klíčová slova: Globální, hloubka dat, klasifikace, lokální.

Abstract: Ordering of points in multidimensional space according to their depth with respect to some probability distribution can be the basis for solving many statistical problems. It can be used for classification – the formation of a rule for the assessment of new observations into one of several groups whose representatives are observed. Several depth-based classifiers have been proposed during the last ten years. The aim of the this contribution is to summarize depth-based classifiers and to present new trends in this area.

Keywords: Global, data depth, classification, local.

1. Úvod

Hloubku dat jakožto prostředek k zobecnění pojmu uspořádání pro účely mnohorozměrné analýzy popularizuje v českém prostředí se svými studenty Daniel Hlubinka. Připomeňme zde jeho článek *Výpravy do hlubin dat* z roku 2009 [8], který mimo jiné mapuje různé používané hloubkové funkce, jako je poloprostorová hloubka, simplexová hloubka, zonoidová hloubka či L_1-hloubka.

V článku, který právě čtete, se snažíme shrnout výhody i nevýhody hloubky při řešení úlohy klasifikace. Jsme tedy v situaci, kdy máme (pro jednoduchost) dvě neznámá rozdělení P_1 a P_2 na d-rozměrném reálném prostoru.

Tato rozdělení nechť jsou spojitá. K dispozici máme náhodný výběr z P_1: $X_{1,j}$, $j = 1, \ldots, n_1$ a (nezávislý) náhodný výběr z P_2: $X_{2,j}$, $j = 1, \ldots, n_2$. Tyto náhodné výběry dohromady tvoří tzv. tréningovou množinu. Empirická rozdělení získaná na základě těchto náhodných výběrů označujeme $\widehat{P}_1$ a $\widehat{P}_2$. Hloubku bodu $\boldsymbol{x} \in \mathbb{R}^d$ (ať jde o jakoukoliv z výše uvedených hloubkových funkcí) vůči pravděpodobnostnímu rozdělení P, budeme značit $D(\boldsymbol{x}; P)$.

2. Klasifikace podle maximální hloubky – první nápad, jak využít hloubku pro klasifikaci

První klasifikátor využívající hloubky dat byl založen na jednoduché myšlence klasifikovat nové pozorování do té třídy, vůči níž má maximální hloubku. Přesněji:

$$d(\boldsymbol{x}) = \arg\max_{i=1,2} D(\boldsymbol{x}; \widehat{P}_i) \tag{1}$$

Tato myšlenka má poměrně jednoduché opodstatnění. Představme si například dvě dvourozměrná normální rozdělení s rovnými apriorními pravděpodobnostmi lišící se pouze posunutím. Body blízké středu symetrie jednoho rozdělení mají velkou hloubku vůči tomuto rozdělení a menší hloubku vůči druhému (posunutému) rozdělení. Je tedy „přirozené“ přiřadit je k tomu rozdělení, vůči němuž mají větší hloubku (jsou blíže jeho centru).

Klasifikaci na základě maximální hloubky využili ve svých pracích např. Jörnsten [12], Hartikainen a Oja [7] (použili L_1-hloubku), Ghosh a Chaudhuri [6] (použili poloprostorovou hloubku), Mosler a Hoberg [16] (použili kombinaci zonoidové a Mahalanobisovy hloubky) nebo Kosiorowski [13], Hubert a Van der Veeken [11] a Dutta a Ghosh [5] (použili projekční hloubku).

Již Ghosh and Chaudhuri [6] však dokázali, že klasifikátor založený na maximální hloubce je vhodný jen ve velmi specifické situaci. Ukázali, že klasifikátor založený na maximální hloubce je (asymptoticky) bayesovsky optimální za předpokladu, že použitá hloubka je afinně invariantní a uvažovaná rozdělení P_1, P_2 splňují (všechny) následující podmínky:

- jsou elipticky symetrická s hustotou klesající ze středu symetrie,
- liší se pouze parametrem polohy,
- mají stejnou apriorní pravděpodobnost $\pi_1 = \pi_2 = 1/2$.

Jak ukazuje následující jednoduchý příklad, problém nastane už v situaci, kdy se rozdělení liší v disperzi. Uvažujme například dvě jednorozměrná normální rozdělení s nulovou střední hodnotou a různými rozptyly $0 < \sigma_1^2 < \sigma_2^2$:

$P_1 = N(0, \sigma_1^2)$, $P_2 = N(0, \sigma_2^2)$. Budeme uvažovat poloprostorovou hloubku. Pro bod $x = 0$ platí $D(x, P_1) = D(x, P_2)$, pro všechny ostatní body je však $D(x, P_1) < D(x, P_2)$, což plyne z korespondence poloprostorové hloubky a kvantilu rozdělení v jednorozměrném případě. Získáme tedy klasifikátor, který s pravděpodobností 1 klasifikuje nové pozorování do druhé třídy (k rozdělení s větším rozptylem).

3. Další klasifikátory založené na hloubce

Poté, co v roce 2005 Ghosh a Chaudhuri odhalili podstatná omezení použitelnosti klasifikátoru založeného na maximální hloubce, začaly se hledat poněkud sofistikovanější postupy jak využít hloubku dat ke klasifikaci.

Typický postup nalezení klasifikátoru využívajícího hloubku dat je dvoukrokový. Prvním krokem je výpočet hloubek původních pozorování vůči jednotlivým skupinám bodů tréningové množiny. Jde tedy o zobrazení $\mathbb{R}^d \to [0, 1]^2$. Jelikož typicky (i když ne nutně) platí, že $d > 2$, můžeme tento krok chápat jako redukci dimenze úlohy. Navíc budeme nadále pracovat s kompaktní množinou $[0, 1]^2$. Druhým krokem je nalezení vhodného klasifikátoru na prostoru $[0, 1]^2$.

Různé klasifikátory se liší v tom, jak realizují dva výše uvedené kroky. Rozdílnost v prvním kroku plyne z různosti hloubek v tomto kroku použitých. Muže být použita např. poloprostorová, projekční, zonoidová či L_1 hloubka. Kterákoliv z výše uvedených hloubek určitého bodu je globální charakteristikou tohoto bodu udávající míru jeho centrality vůči nějakému rozdělení či skupině bodů. Je však možno také použít některou z lokálních hloubek. Rovněž druhý krok muže být realizován různě. Některé procedury se dají označit jako globální, jiné jako lokální. Za globální označíme ty procedury, kde k zařazení nového pozorování porovnáváme jeho hloubky s hloubkami všech bodů tréningové množiny. Naopak lokální procedury jsou typicky založené na porovnávání s blízkými sousedy (z hlediska hloubek).

4. První krok – hloubka dat

Hloubka dat, jakožto míra centrality bodu vůči nějakému pravděpodobnostnímu rozdělení, je ve své podstatě globální charakteristikou tohoto bodu, neboť popisuje jeho polohu vůči (celému) rozdělení – v případě empirické hloubky vůči náhodně vybraným bodům z tohoto rozdělení. V posledních několika letech se však množí pokusy o lokalizaci hloubky. Ty většinou využívají některé ze známých hloubkových funkcí při použití váhové funkce odlišující „důležitost“ jednotlivých bodů podle jejich vzdálenosti od bodu,

jehož hloubku počítáme. Byla navržena například vážená poloprostorová hloubka [9] nebo vážená simplexová hloubka [1].

Ukazuje se, že hloubka pojatá globálně, může vést k dobrým výsledkům pouze tehdy, mají-li uvažovaná rozdělení aspoň některé globální vlastnosti jako je symetrie či unimodalita. Pokud jsou však uvažovaná rozdělení nesymetrická, multimodální či pokud mají například nekonvexní úrovňové množiny hustoty, je použití hloubky přinejmenším problematické. Zmiňme v této souvislosti výsledek dokázaný Duttou a jeho spoluautory [4] poukazující na nesoulad úrovňových množin poloprostorové hloubky s úrovňovými množinami hustoty:

Mějme rozdělení na $\mathbb{R}^d$ *s hustotou* f *ve tvaru* $f(\boldsymbol{x}) = \phi(\|\boldsymbol{x}\|_p)$, *kde* ϕ *je nějaká klesající funkce. Úrovňové množiny poloprostorové hloubky odpovídají úrovňovým množinám hustoty právě tehdy, když* $p = 2$.

Jelikož je bayesovský optimální klasifikátor založený na hustotě uvažovaných rozdělení, je celkem pochopitelná snaha pomocí lokalizace hloubky řešit problém neshody úrovňových množin hloubky s úrovňovými množinami hustoty. Klasifikátory využívající lokální hloubku byly navrženy např. v článcích [3] či [10].

5. Druhý krok – klasifikace na prostoru hloubek

Klasifikaci na prostoru hloubek byla a stále je věnována značná pozornost. Pokusíme se nyní přiblížit čtenáři některé navržené postupy, přičemž opět budeme rozlišovat přístupy globální a lokální.

5.1. Globální klasifikátory na prostoru hloubek

1. Jednoduché vylepšení klasifikátoru založeného na maximální hloubce navrhla v roce 2008 skupina kolem Nedret Billor. Autoři článku [2] upozornili na skutečnost, že hloubka libovolného bodu vůči jednomu rozdělení může nabývat hodnot z širšího intervalu než hloubka téhož bodu vůči jinému rozdělení. Například nejhlubší bod symetrického rozdělení má poloprostorovou hloubku 1/2, avšak nejhlubší bod nesymetrického rozdělení má poloprostorovou hloubku striktně menší než 1/2. Toto jednoduché pozorování je vedlo k přesvědčení, že místo samotné hloubky by bylo vhodnější pracovat s kvantily hloubky. Navrhli proto následující klasifikátor:

$$d(\boldsymbol{x}) = \arg\max_{i=1,2} \frac{1}{n_i} \sum_{j=1}^{n_i} I\left(D(\boldsymbol{X}_{i,j}, \widehat{P}_i) \leq D(\boldsymbol{x}; \widehat{P}_i)\right).$$

Takto navržený klasifikátor jeho autoři nazvali „depth transvariation classifier“, stejně dobře bychom však mohli mluvit o klasifikaci na základě kvantilu hloubky. Ukázalo se však, že toto „vylepšení“ neřeší problém různých disperzí či různých apriorních pravděpodobností.

2. Podstatné vylepšení přinesl klasifikátor založený na DD-grafu (DD-plot classifier) navržený v roce 2012 (první verze článku [15] byla přitom elektronicky dostupná již v roce 2010). DD-grafem rozumíme dvourozměrný graf, kde na x-ové ose je vynášena hloubka bodu vůči jednomu rozdělení (resp. vůči jedné skupině bodů) a na y-ové ose hloubka vůči druhému rozdělení (druhé skupině bodů). Anglický název DD-plot vznikl zkrácením slov „depth versus depth“ plot. Klasifikátor založený na maximální hloubce je v tomto grafu znázorněn přímkou procházející počátkem a půlící první kvadrant. Prvotní ideou bylo hledat přímku procházející počátkem s takovou směrnicí, aby byl minimalizován počet chybných zařazení v tréningové množině. Nemusíme se však omezovat jen na přímky. Můžeme použít i polynomy vyššího stupně.

3. Pro praktickou analýzu dat se jeví jako velmi zajímavá možnost využití tzv. DDα-klasifikátoru navrženého v článku [14]. Tento klasifikátor místo dvojice $[D(\boldsymbol{x}, \widehat{P}_1), D(\boldsymbol{x}, \widehat{P}_2)]$, používané v klasifikaci pomocí DD-grafu, pracuje s šířeji pojatým vektorem

$$\boldsymbol{z} := \big[D(\boldsymbol{x}, \widehat{P}_1), D(\boldsymbol{x}, \widehat{P}_2), D(\boldsymbol{x}, \widehat{P}_1) \cdot D(\boldsymbol{x}, \widehat{P}_2), D(\boldsymbol{x}, \widehat{P}_1)^2, D(\boldsymbol{x}, \widehat{P}_2)^2\big].$$

V článku je navržen heuristický postup pro nalezení vhodných parametrů oddělující nadroviny $aD(\boldsymbol{x}, \widehat{P}_1) + bD(\boldsymbol{x}, \widehat{P}_2) + cD(\boldsymbol{x}, \widehat{P}_1)D(\boldsymbol{x}, \widehat{P}_2) + dD(\boldsymbol{x}, \widehat{P}_1)^2 + eD(\boldsymbol{x}, \widehat{P}_2)^2 = 0$.

5.2. Lokální klasifikátory na prostoru hloubek

1. *Klasifikátor využívající jádrového odhadu hustoty.* V článku z roku 2005 Ghosh a Chaudhuri [6] poukázali na skutečnost, že vzhledem ke korespondenci úrovňových množin hloubky a úrovňových množin hustoty u elipticky symetrických rozdělení muže být bayesovský optimální klasifikátor v tomto případě vyjádřen v podobě

$$d(\boldsymbol{x}) = \arg\max_{i=1,2} \pi_i \theta_i(D(\boldsymbol{x}; \widehat{P}_i)),$$

kde θ_i, $i = 1, 2$ jsou nějaké neznámé reálné funkce. Stačí tedy použít jádrový odhad, abychom z dat v podobě bodů $D(\boldsymbol{X}_{i,j}; \widehat{P}_i)$, $j = 1, \ldots, n_i$,

$i = 1, 2$, odhadli průběh funkcí θ_1, θ_2. Tento postup je vhodný pro eliptícky symetrická rozdělení, v jiných případech však jeho použití bylo zkoumáno jen empiricky.

2. *Metoda k nejbližších sousedů na prostoru hloubek.* Jednoduchá myšlenka použít neparametrickou metodu k nejbližších sousedů na prostoru hloubek byla použita například v článku [19]. Otázkou je, jaká metrika je v tomto případě vhodná, resp. zda lze využít nějaké nestandardní metriky k zlepšení schopnosti klasifikace. Existují i jiné postupy využívající myšlenku k nejbližších sousedů, viz například [18].

3. *Klasifikátor využívající symetrizace.* Lokální klasifikátor, který však poněkud vybočuje z dvoukrokového schématu klasifikace s využitím hloubky, navrhli v roce 2012 Paindaveine a Van Bever [17]. Jejich klasifikátor využívá skutečnosti, že je-li rozdělení symetrické (v nějakém smyslu), je bodem s největší hloubkou právě bod, kolem kterého je rozdělení symetrické. Označíme-li body tréningové množiny $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, bude bod $\boldsymbol{x}$, který máme klasifikovat, (asymptoticky) nejhlubším bodem vůči bodům $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n, 2\boldsymbol{x} - \boldsymbol{X}_1, \ldots, 2\boldsymbol{x} - \boldsymbol{X}_n$ (k původním datům jsme přidali jejich obrazy při symetrickém zobrazení se středem v $\boldsymbol{x}$). Body z tréningové množiny tak můžeme uspořádat podle jejich hloubky vzhledem k takto doplněné (symetrizované) tréningové množině a vzít k nejvíce centrálních. Ty považujeme za k nejbližších sousedů nového pozorování $\boldsymbol{x}$ a použijeme většinový princip, tedy pozorování $\boldsymbol{x}$ přiřadíme do skupiny s (nej)větším počtem zástupců mezi k nejbližšími sousedy. Paindaveine a Van Bever [17] ukázali, že tento klasifikátor je za velmi obecných předpokladů konzistentní. S nutností provádět symetrizaci a výpočet hloubek s každým novým pozorováním je však spojena vysoká výpočetní náročnost této metody.

6. Závěr

V oblasti klasifikace založené na hloubce dat jsme za posledních deset let zaznamenali velký metodologický pokrok. Hloubka dat, jakožto nástroj pro uspořádání bodů, je dnes vnímána jako možný základ neparametrických metod mnohorozměrných dat. Očekávalo by se tedy, že klasifikátory založené na hloubce budou konzistentní za velmi obecných podmínek. Zdaleka to však neplatí. Mnoho navržených postupů je konzistentních (optimálních) jen pro úzkou třídu problémů. Zdrojem tohoto zklamání je samotná podstata hloubky, která je ve své podstatě globální charakteristikou polohy bodu vůči rozdělení. Hloubku lze tedy účelně využít jen tehdy, když uvažovaná roz-

dělení mají některé důležité globální vlastnosti (unimodalitu, symetrii ap.). V obecnějším případě se jako nutná jeví lokalizace hloubky či alespoň spojení hloubky s klasifikační procedurou, jejíž povaha je lokální (jako je například metoda k nejbližších sousedů či klasifikace založená na odhadu hustoty pomocí jádrových odhadů).

Poděkování

Článek byl napsán s podporou Operačního programu Vzdělávání pro konkurenceschopnost (projekt CZ.1.07/2.3.00/20.0170).

Literatura

[1] Agostinelli C., Romanazzi M.: Local depth. *Journal of Statistical Planning and Inference*, **141**, 817–830, 2011.

[2] Billor N. et al.: Classification based on depth transvariations. *Journal of Classification*, **25**, 249–260, 2008.

[3] Dutta S., Chaudhuri P., Ghosh A. K.: Some intriguing properties of Tukey's half-space depth. *Bernoulli*, **17**, 1420–1434, 2011.

[4] Dutta S., Chaudhuri P., Ghosh A. K.: Classification using Localized Spatial Depth with Multiple Localization. *Communicated for publication*, 2012.

[5] Dutta S., Ghosh A. K.: On robust classification using projection depth. *Annals of the Institute of Statistical Mathematics*, **64**, 657–676, 2012.

[6] Ghosh A. K., Chaudhuri P.: On maximum depth and related classifiers. *Scandinavian Journal of Statistics*, **32**, 327–350, 2005.

[7] Hartikainen A., Oja H.: On some parametric, nonparametric and semiparametric discrimination rules. In: *Data depth: robust multivariate analysis, computational geometry and applications,* Liu R. Y., Serfling R., Souvaine D. L. (eds.). 1. vyd. New York: American Mathematical Society, 61–70, 2006.

[8] Hlubinka D.: Výpravy do hlubin dat. In *Robust 2008*, Antoch J., Dohnal G. (eds.). Praha: Jednota českých matematiků a fyziků, 2009.

[9] Hlubinka D., Kotík L., Vencálek O.: Weighted data depth. *Kybernetika*, **46** (1), 125–148, 2010.

[10] Hlubinka D., Vencálek O.: Depth-Based Classification for Distributions with Nonconvex Support. *Journal of Probability and Statistics*, 2013.

[11] Hubert M., Van der Veeken S.: Robust classification for skewed data. *Advances in Data Analysis and Classification*, **4** (4), 239–254, 2010.

[12] Jörnsten R.: Clustering and classification based on the L_1 data depth. *Journal of Multivariate Analysis*, **90**, 67–89, 2004.

[13] Kosiorowski D.: Robust classification and clustering based on the projection depth function. In *COMPSTAT 2008: proceedings in computational statistics.* 18th symposium held in Porto, Portugal, Brito P. (ed.). [CD-ROM]. Heidelberg: Physica, 209–216, 2008.

[14] Lange T., Mosler K., Mozharovskyi P.: Fast nonparametric classification based on data depth. *Statistical Papers*, 1–21, 2012.

[15] Li J., Cuesta-Albertos J. A., Liu R.: DD-classifier: nonparametric classification procedure based on DD-plot. *JASA*, **107** (498), 737–753, 2012.

[16] Mosler K., Hoberg R.: Data analysis and classification with the zonoid depth. In *Data depth: robust multivariate analysis, computational geometry and applications.* Liu R. Y., Serfling R., Souvaine D. L. (eds.). 1. vyd. New York: American Mathematical Society, 49–59, 2006.

[17] Paindaveine D., Van Bever G.: Nonparametrically consistent depth-based classifiers. *Preprint arXiv:1204.2996*, 2012.

[18] Vencálek O.: Weighted data depth and depth based classification. PhD thesis, 2011. URL: `http://artax.karlin.mff.cuni.cz/~venco2am/DataDepth.html`.

[19] Vencálek O.: New depth-based modification of the k-nearest neighbour method. *SOP Transactions on Statistics and Analysis*, 2014.

PROBLÉM KONKURUJÍCÍCH SI RIZIK A JEJICH IDENTIFIKACE

ON PROBLEM OF COMPETING RISKS AND THEIR IDENTIFICATION

Petr Volf

Adresa: Institute of Information Theory and Automation AS CR, Pod Vodárenskou věží 4, 182 08, Prague 8

E-mail: volf@utia.cas.cz

Abstrakt: Příspěvek je věnován problému konkurujících si rizik ve statistické analýze přežití. Komplikace v analýze těchto případů je způsobena tím, že příslušné náhodné veličiny mohou být závislé. Nejprve ukážeme, jak je možné konzistentně odhadnout t.zv. incidenční funkce. Dále se zabýváme vztahy mezi marginálními, simultánním a incidenčními distribucemi v případě, že je simultánní rozdělení vyjádřeno pomocí kopuly.

Klíčová slova: Analýza přežití, konkurující si rizika, incidence, kopula.

Abstract: The contribution deals with the problem of competing risks (of competing events) in the statistical survival analysis. The case is complicated by the fact that the potential occurrence of both events may be dependent. We recall the notion of incidence function and the methods of statistical incidence analysis. Then we study the relationship between marginal, joint and incidence distributions of events when the joint distribution is modeled via a copula.

Keywords: Survival analysis, competing risks, incidence, copula.

1. Competing risks problem

Let us consider random times to certain competing (two or more) events, for instance a failure of a device caused by one of several possible causes. An underlying model assumes that there are K possibly dependent random variables T_j, $j = 1, \ldots, K$. Typically, we also have to add a censoring variable C, independent of all T_j's. It is further assumed that observation of the object (device) ends with the first occurring event (or by censoring). Hence, we observe $Z = \min(T_1, \ldots, T_K, C)$ and we know also what was the cause, so that we observe an indicator $\delta = 1, \ldots, K, 0$ if $Z = T_1, \ldots, T_K, C$, respectively.

It is known (e.g. Tsiatis, 1975) that, in general, from such observations (N i.i.d. couples (Z_i, δ_i), $i = 1, \ldots, N$) it is not possible to identify neither the joint distribution of (T_j) nor their marginal distributions. On the other hand,

the data allow to estimate consistently joint distributions of $(T_j, \delta = j)$ and corresponding sub-distribution functions called the (cumulative) incidence functions.

The situation can be better when our information is richer thanks to dependence of data on observed covariates. We shall recall briefly an identifiability results of Heckman and Honoré [2]. Part 3 then deals with a copula model applied to the competing risks case. Finally, an example shows the use of Gauss copula and estimation of incidence functions.

2. Incidence function

The structure of data observed in the competing risks setting enables us to estimate, consistently, the following characteristics: First, the distribution of $Z = \min(T_1, \ldots, T_K)$, namely $S(t) = P(Z > t) = P(T_1 > t, \ldots, T_K > t) = \overline{F}_K(t, \ldots, t)$, where by $\overline{F}_K(t_1, \ldots, t_k)$ we denote the joint survival function of $T_1, \ldots, T_K$. Further, we can estimate the **incidence densities**

$$f_j^*(t) = P'(Z = t, \delta = j) = -\frac{\partial \overline{F}_K(t_1, \ldots, t_K)}{\partial t_j}\bigg|(t_1 = \ldots = t_K = t),$$

and their integrals, **cumulative incidence functions** $F_j^*(t) = \int_0^t f_j^*(s)\,\mathrm{d}s = P(Z \leq t, \delta = j)$. Notice that $\lim F_j^*(t) = P(\delta = j) < 1$ if $t \to \infty$ and $S(t) = 1 - \sum_{j=1}^K F_j^*(t)$.

Another (equivalent, however more practical for estimation) definition of the cumulative incidence function is based on the cause-specific hazard functions for events $j = 1, 2, \ldots, K$,

$$h_j^*(t) = \lim_{d \to 0} \frac{P(t \leq Z < t + d,\, \delta = j \,|\, Z \geq t)}{d}.$$

Overall hazard rate for $Z = \min(T_1, \ldots, T_K)$ is then

$$h(t) = \lim_{d \to 0} \frac{P(t \leq Z < t + d \,|\, Z \geq t)}{d} = \sum_{j=1}^K h_j^*(t),$$

corresponding integrals are cumulated hazard rates $H_j^*(t)$, $H(t)$. Finally, the overall survival function $S(t) = P(Z > t) = \exp(-H(t))$. Then $f_j^*(t) = h_j^*(t) \cdot S(t)$ and cumulative incidence functions can be written as

$$F_j^*(t) = P(Z \leq t, \delta = j) = \int_0^t S(s) \cdot h_j^*(s)\,\mathrm{d}s.$$

2.1. Estimation method

Let us here recall some standard notation. $N_{ij}(t)$ is the counting process with value 0 at $t = 0$ and with step $+1$ at the moment when event of type j is observed on object i. Further, let $Y(t)$ denote the number of objects in the risk set at (just before) time t, i.e. of objects without any event and not censored before t. All cumulative hazard rates can be estimated standardly by the Nelson-Aalen estimator, namely

$$\widehat{H}_j^*(t) = \int_0^t \sum_{i=1}^n \frac{\mathrm{d}N_{ij}(s)}{Y(s)}, \qquad \widehat{H}(t) = \sum_{j=1}^K \widehat{H}_j^*(t). \tag{1}$$

Overall survival function can then be estimated by the Kaplan Meier "Product Limit" (PL) estimator, or by $\widehat{S}(t) = \exp(-\widehat{H}(t))$. Asymptotic properties of estimates of incidence functions

$$\widehat{F}_j^*(t) = \int_0^t \widehat{S}(s)\,\mathrm{d}\widehat{H}_j^*(s) \tag{2}$$

follow from good asymptotic properties of $\widehat{S}$ and $\widehat{H}_j^*$ and are derived for instance in Lin [3]. In general, limit distribution of $\sqrt{n}(\widehat{F}_j^*(t) - F_j^*(t))$ is that of Gauss random process, with estimable covariance structure. As it is not a martingale, further inference (e.g. statistical tests) is not easy. Notice, however, that in the simplest case without censoring $F_j^*(t)$ and $S(t)$ correspond, at each fixed t, to probabilities in a multinomial distribution, the estimates correspond to relative occurrence, so that their properties simplify. In general, confidence regions for statistical testing are obtained by a Monte Carlo random generation.

2.2. Non-identifiability

A. Tsiatis [5] has shown that for arbitrary joint model we can find a model with independent components having the same incidences, i.e. we cannot distinguish the models. Namely, this "independent" model is given by cause-specific hazard functions $h_j^*(t)$. In a parametric setting it also means that even if the MLE yields consistent estimates, we don't know parameters of which multivariate model are estimated.

On the other hand, Heckman and Honoré [2], and then others, have proved, under suitable conditions, that in the case of regression models (they considered Cox or AFT models), when our information is enriched due to knowledge of covariate values, the competing risk data suffices for full identification of the model.

3. Competing risk and copula

In the sequel we shall consider just a couple of competing events, $K = 2$, represented by random variables S, T. From the above it follows that without some knowledge about mutual dependence of S, T we are not able, in general, to estimate their distribution. The copulas offer a possibility how to model two (or multi-) dimensional distributions. Let us recall that Sklar's theorem ensures that to each 2-dimensional distribution function (of a continuous-type distribution) there exists an unique function $C(u, v)$, a distribution function on $(0, 1)^2$, such that

$$F_2(s, t) = C(F_S(s),\, F_T(t)), \tag{3}$$

where F_S, F_T are marginal distribution functions of variables S, T. The marginals of $C(u, v)$ correspond to random variables $U = F_S(S)$, $V = F_T(T)$ and have uniform distribution on $(0, 1)$. There are several classes of copulas analyzed theoretically or used practically (cf. Cherubini et al. [1]). Zheng and Klein [6] showed that in the competing risks setting, when the copula function is given (assumed), marginal distributions of S, T, and then also joint distribution from (3), are estimable. They also proposed a procedure of the non-parametric estimation, proved asymptotic results and showed that their estimator reduces to the Kaplan-Meier PL estimator if S, T are independent.

3.1. Use of Gauss copula

Let X, Y be standard normal random variables $N(0, 1)$ tied with (Pearson) correlation $\rho = \rho(X, Y)$. We denote ϕ, φ univariate standard normal distribution function and density and by $\phi_2(x, y)$, $\varphi_2(x, y)$ corresponding 2-dimensional functions. Then

$$\varphi_2(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\Big\{ -\frac{1}{2}\boldsymbol{x}'\Sigma^{-1}\boldsymbol{x}\Big\} \tag{4}$$

with $\boldsymbol{x} = (x, y)'$ and Σ the 2×2 covariance matrix with rows $(1, \rho)$ and $(\rho, 1)$. If we define $U = \phi(X)$, $V = \phi(Y)$, we obtain a 2-dimensional distribution on $(0, 1)^2$ with the copula

$$C(u, v) = \phi_2\big(\phi^{-1}(u),\, \phi^{-1}(v)\big). \tag{5}$$

Naturally, $\rho(U, V) \neq \rho(X, Y)$ (though they are rather close, as a rule), while Spearman's correlations coincide, namely $\rho_{\mathrm{SP}}(X, Y) = \rho_{\mathrm{SP}}(U, V) = \rho(U, V)$.

We are, however, primarily interested in the model for dependence of competing variables S, T. Let us assume that their joint distribution function fulfils (3), where $C(u, v)$ is the copula (5). It further follows that

$$F_2(s,t) = \phi_2\big(\phi^{-1}(F_S(s)),\, \phi^{-1}(F_T(t))\big), \tag{6}$$

and, inversely, $S = F_S^{-1}(\phi(X))$, $T = F_T^{-1}(\phi(Y))$. Hence, again $\rho_{\mathrm{SP}}(S,T) = \rho_{\mathrm{SP}}(U,V)$, and "initial" $\rho = \rho(X,Y)$ is the only parameter describing the dependence of S and T. It, naturally, differs from $\rho(S,T)$, however, all values $\rho(S,T)$ (at least from $(-1,1]$) can be achieved by convenient choice of $\rho(X,Y)$. Such a flexibility is not common to many other copula types. Let us remark here that the real dependence among S,T can be much more complicated, nevertheless the use of Gauss copula model offers rather simple and sufficiently flexible (as regards the correlation) set of distributions.

3.2. Estimation in model with Gauss copula

When parameter ρ is known, copula (5) is fully defined and from Zheng, Klein [6] it follows that the distribution of (S,T) can be estimated, in parametric and even non-parametric setting. On the other hand, when marginal distributions F_S, F_T are known and both (3) and (5) hold with the same copula, then $\rho = \rho(X,Y)$ is estimable, and then also is the joint distribution $F_2(s,t)$. The estimation procedure is based on the maximum likelihood method. The data are $(Z_i, \delta_i),\ i = 1, \ldots, N$. The likelihood function then has the form

$$L = \prod_{i=1}^{N} \left\{ -\frac{\partial}{\partial s} \overline{F}_2(s,t) \right\}^{I[\delta_i=1]} \times \left\{ -\frac{\partial}{\partial t} \overline{F}_2(s,t) \right\}^{I[\delta_i=2]} \times \overline{F}_2(s,t)^{I[\delta_i=0]},$$

evaluated at $s = t = Z_i$, with $\overline{F}_2(s,t) = P(S > s,\, T > t) = 1 - F_S(s) - F_T(t) + F_2(s,t)$. It is due transformation (3) and (5) that $F_2(s,t) = \phi_2(x,y)$ with $x = \phi^{-1}(F_S(s))$, $y = \phi^{-1}(F_T(t))$. Hence, when we put $X_i = \phi^{-1}(F_S(Z_i))$, $Y_i = \phi^{-1}(F_T(Z_i))$, we obtain after some computation – integration of 2-dimensional Gauss density $\varphi_2(x,y)$, that

$$\begin{aligned} L \;=\; & \prod_{i=1}^{N} \left\{ f_S(Z_i) \left[1 - \phi_1(Y_i; \rho X_i, 1-\rho^2) \right] \right\}^{I[\delta_i=1]} \times \\ & \times \left\{ f_T(Z_i) \left[1 - \phi_1(X_i; \rho Y_i, 1-\rho^2) \right] \right\}^{I[\delta_i=2]} \times \\ & \times \left\{ 1 - F_S(Z_i) - F_T(Z_i) + \phi_2(X_i, Y_i \right\}^{I[\delta_i=0]}, \end{aligned}$$

where $\phi_1(x;\mu,\sigma^2)$ denotes the distribution function of normal distribution $N(\mu,\sigma^2)$, evaluated at x. It is seen that the problem of maximization has to be solved by a convenient search procedure. Parameter ρ is hidden in ϕ_1 and in ϕ_2. Distributions of S and T are present both explicitly and also implicitly, in transformed X_i, Y_i. Nevertheless, experience suggests that solution of both problems (estimate of F_S, F_T for given ρ, estimate of ρ for given F_S, F_T) are solvable and have unique solution.

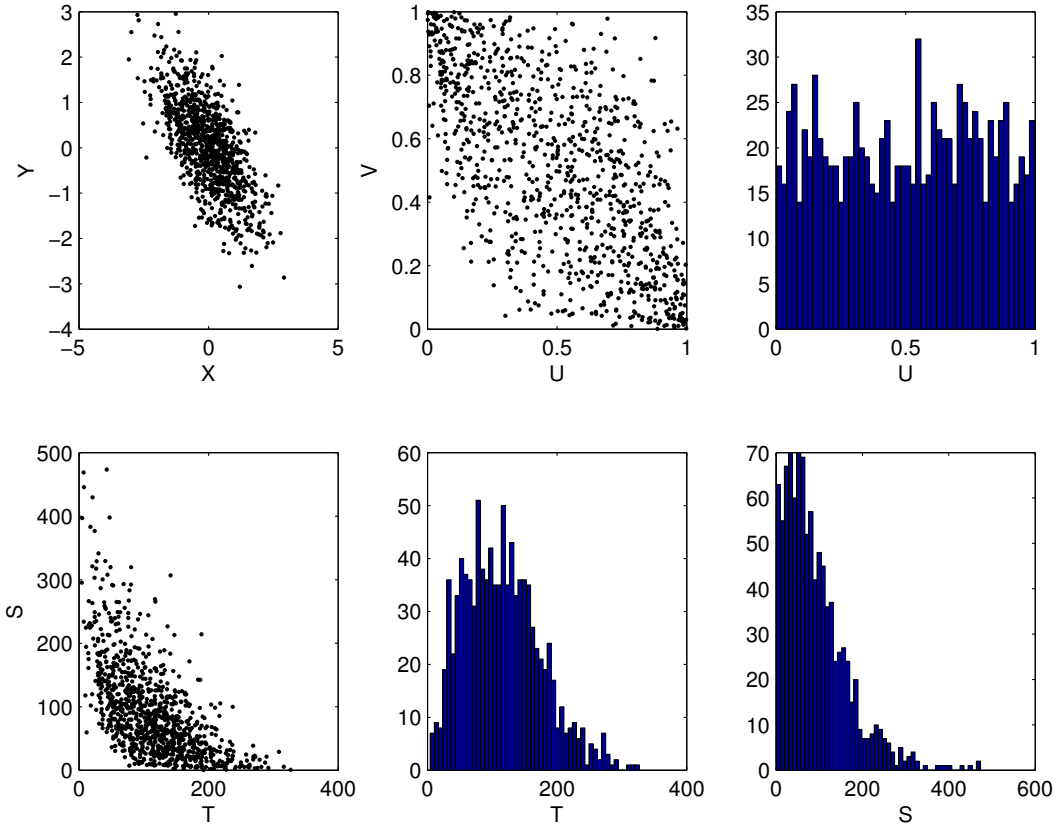


Figure 1: Scatter-plots and histograms of generated representation of X, Y, then transformed to U, V and S, T, the case with $\rho = -0.7$, N=1000.

4. Example using Gauss copula

We fixed both competing risks distributions, namely $S \sim$ Weibull ($a_s = 100$, $b_s = 1.2$), $T \sim$ Weibull ($a_t = 130$, $b_t = 3$), and censoring variable $C \sim |\text{Normal}(\mu = 150,\, \sigma = 50)|$. The rate of censoring was among $10-20\,\%$. Weibull distribution function was taken in form $F(s) = 1-\exp\left(-\left(\frac{s}{a}\right)^b\right),\ s>0$. The analysis was done for two values of ρ, namely for $\rho = 0.5$ and $\rho = -0.7$.

First, we show how the data (X, Y) generated from ϕ_2 are transformed to (U, V) by (5) and then to (S, T) by (3). Figure 1 shows the scatter-plots and

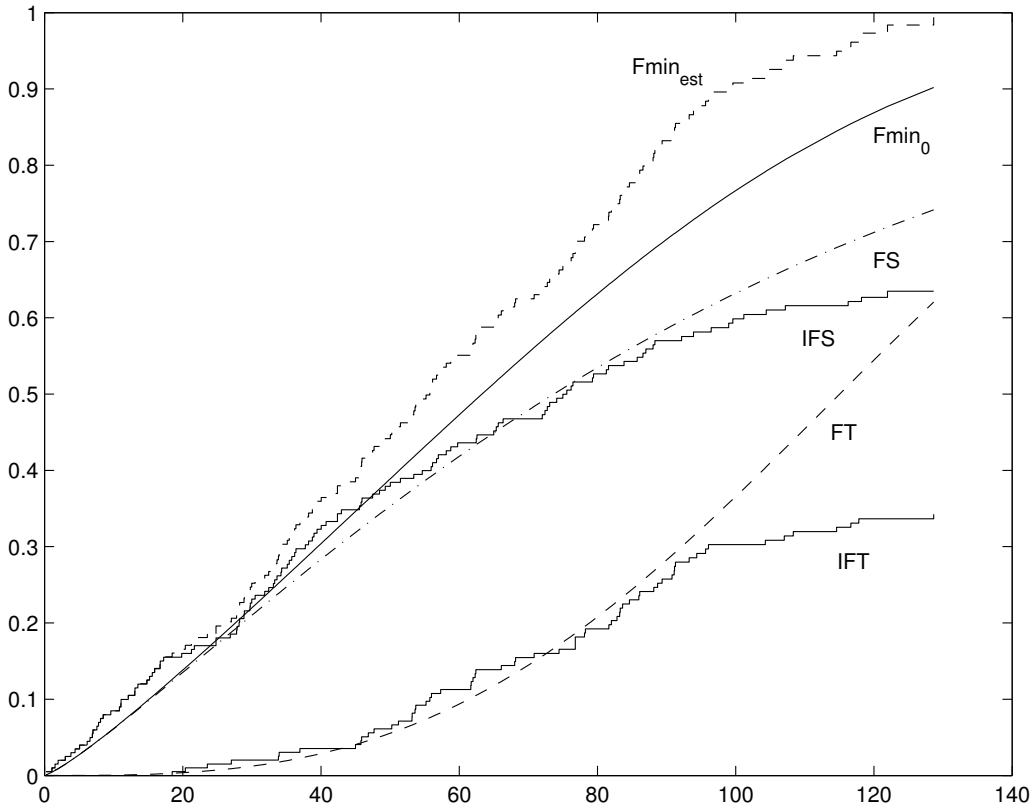


Figure 2: "True" distribution functions F_S, F_T, $F\min_0$ under independence hypothesis, estimated $F\min_{\text{est}}$, estimated cumulative incidence functions IF_S, IF_T. Case of $\rho = -0.7$, $N = 200$.

histograms of generated values ($N = 1000$), in the case $\rho = -0.7$. We also computed distributions numerically. Numerically computed correlations yield $\rho(U, V) = 0.432$, $\rho(S, T) = 0.376$ in the case with $\rho = 0.5$, and $\rho(U, V) = -0.685$, $\rho(S, T) = -0.625$ in the case with $\rho(X, Y) = -0.7$.

Further, we show an example of estimates of cumulative incidence functions, following the approach described in part 2. The same type of data as in previous example was generated. We display here just the case of $\rho = -0.7$, $N = 200$. Figure 2 shows both underlying "true" distribution functions F_S and F_T, and also $F\min_0$ of $\min(S, T)$ under hypothesis of independence. Dashed step-wise curve is the PL-estimate of true distribution $F\min_{\text{est}}$ of $\min(S, T)$. It differs from $F\min_0$, it could be taken as an evidence that independence hypothesis does not hold. Finally, two full step-wise curves are estimated cumulative incidence functions IF_S, IF_T of S, T, respectively. Notice that they summarize to $F\min_{\text{est}}$. Easy generation of artificial data is another advantage of Gauss copula.

In a particular case when marginal distribution are known, the hypothesis of independence (i.e. that $F\min = F\min_0$) can be tested easily with the

aid of asymptotic properties of the PLE $F\min_{\text{est}}$. If, moreover, the type of copula is assumed (as in our case here), parameter ρ can be estimated by the ML method and test of hypothesis on its value can be based on asymptotic normality of the MLE.

True cumulative incidence functions can be obtained by integration of expressions corresponding to the first and second part of the likelihood function. Namely, we used numerical integration of

$$dIF_S(t) = f_S(t)\left[1 - \phi_1(y; \rho x, 1-\rho^2)\right],$$
$$dIF_T(t) = f_T(t)\left[1 - \phi_1(x; \rho y, 1-\rho^2)\right],$$

where again $x = \phi^{-1}(F_S(t))$, $y = \phi^{-1}(F_T(t))$.

5. Conclusion

The problem of competing risks has been studied and the difference between marginal distributions and observed incidence of events has been analyzed. The main goal was to describe the procedure of estimation of incidence functions and, further, to study the use of Gauss copula in modeling and random generation of competing risks data.

Acknowledgement

This work was supported by the GA CR grant No. 13-14445S.

References

[1] Cherubini U., Luciano E., Vecchiato W.: *Copula Methods in Finance.* Wiley, Chichester, 2004.

[2] Heckman J. J., Honoré B. E.: The identifiability of the competing risks model, *Biometrika* **76**, 325–330, 1989.

[3] Lin D. Y.: Non-parametric inference for cumulative incidence fumctions in competing risks studies, *Statistics in Medicine* **16**, 901–910, 1997.

[4] Scheike T. H., Zhang M.: Flexible competing risks regression modelling and goodness-of-fit, *Lifetime Data Analysis* **14**, 464–483, 2008.

[5] Tsiatis A.: A nonidentifiability aspects of the problem of competing risks, *Proc. Nat. Acad. Sci. USA* **72**, 20–22, 1975.

[6] Zheng M., Klein J. P.: Estimates of marginal survival for dependent competing risks based on an assumed copula, *Biometrika* **82**, 127–138, 1995.

Obsah

Vědecké a odborné statě

Informační bulletin České statistické společnosti vychází čtyřikrát do roka v českém vydání. Příležitostně i mimořádné české a anglické číslo. Vydavatelem je Česká statistická společnost, IČ 00550795, adresa společnosti je Na padesátém 81, 100 82 Praha 10. Evidenční číslo registrace vedené Ministerstvem kultury ČR dle zákona č. 46/2000 Sb. je E 21214. Časopis je na Seznamu recenzovaných neimpaktovaných periodik vydávaných v ČR, více viz server `http://www.vyzkum.cz/`.

The Information Bulletin of the Czech Statistical Society is published quarterly.
The contributions in bulletin are published in English, Czech and Slovak languages.

Předsedkyně společnosti: prof. Ing. Hana Řezanková, CSc., KSTP FIS VŠE v Praze, nám. W. Churchilla 4, 130 67 Praha 3, e-mail: `hana.rezankova@vse.cz`.

DOI: 10.5300/IB, http://dx.doi.org/10.5300/IB
ISSN 1210–8022 (Print), ISSN 1804–8617 (Online)