INFORMAČNÍ BULLETIN

České statistické společnosti

Ročník 22, číslo 1, březen 2011

REPORT ON THE ACTIVITIES OF THE CZECH STATISTICAL SOCIETY IN 2010

ZPRÁVA O ČINNOSTI ČESKÉ STATISTICKÉ SPOLEČNOSTI V ROCE 2010

Gejza Dohnal

Adresa: ÚTM FS ČVUT v Praze, Karlovo náměstí 13, 121 35 Praha 2

E-mail: gejza.dohnal@fs.cvut.cz

Abstract: This report presents an overview of activities of the Czech Statistical Society organized under its auspices during the year 2010 together with the plan of activities of the Society for 2011. This report has been presented to the General Assembly of the Czech Statistical Society held on February 7, 2011, in Prague. The members of the Society are also informed about proposal of establishing position of Vice President of the Society ex-officio, which will be offered to the President of the Czech Statistical Office.

Keywords: Czech Statistical Society, annual report.

Abstrakt: Tato zpráva byla přednesena na Valném shromáždění České statistické společnosti, které se konalo dne 7. 2. 2011 v budově VŠE v Praze. Zpráva obsahuje základní údaje o společnosti z roku 2010, přehled činnosti výboru společnosti a organizovaných akcí. Ve zprávě je i přehled akcí plánovaných na rok 2011. V závěru zprávy je návrh na ustanovení funkce místopředsedy společnosti ex oficio, která bude nabídnuta vrcholnému představiteli Českého statistického úřadu.

Klíčová slova: Česká statistická společnost, zpráva o činnosti.

1. Základní údaje o společnosti

Uplynulý rok byl druhým rokem dvouletého funkčního období výboru České statistické společnosti (ČStS), který byl zvolen na valné hromadě dne 29. ledna 2009. Předsedou byl zvolen doc. RNDr. Gejza Dohnal, CSc. (FS ČVUT v Praze), funkci místopředsedy vykonávala prof. Ing. Hana Řezanková, CSc. (VŠE) a hospodářem Ing. Tomáš Löster (VŠE Praha). Zvolený místopředseda Ing. Jan Fischer, CSc., působil jako předseda vlády a od září působí jako viceprezident Evropské banky pro obnovu a rozvoj v Londýně.

K dnešnímu dni má ČStS 236 členů. Za uplynulý vstoupilo do společnosti 20 členů. V roce 2010 zemřel 1 člen společnosti, 2 ukončili členství na vlastní

Doručeno redakci: 15. 2. 2011, imprimatur: 18. 3. 2011.
MSC2010: 01-06, DOI: 10.5300/IB/2011-1/1

žádost. V jedenácti případech bylo členství ukončeno pro neplacení členských příspěvků. V zahraničí žije celkem 10 členů, z toho 6 na Slovensku (4 jsou studenti, kteří studují zde, ale mají trvalé bydliště na Slovensku).

Na základě údajů od 227 členů lze konstatovat, že průměrný věk ve společnosti se příliš nemění : průměrný věk se snížil z 50,3 na 50,1, mediánový věk se zvýšil z 51 na 52 let. Dva nejmladší členové společnosti mají 23 let a nejstarší (prof. František Fabián) 83 let.

2. Činnost výboru společnosti

V průběhu roku se konala tři zasedání výboru České statistické společnosti. O každém z nich byl pořízen zápis, který je všem zájemcům k dispozici. V mezidobí byli členové výboru v kontaktu prostřednictvím e-mailu a diskutovali všechny důležité záležitosti, zejména přípravu akcí a bulletinů. Kromě toho proběhla řada neformálních setkání a porad při jednotlivých akcích.

Rok 2010 byl jubilejním rokem ve kterém naše společnost oslavila 20 let svého trvání. Oficiálně jsme toto výročí oslavili na slavnostním zasedání, které se konalo 4. září při příležitosti Brněnských statistických dnů. Rektor VUT Brno nám pro tuto příležitost propůjčil slavnostní novobarokní aulu. Nechyběl koncert barokní hudby a zahraniční hosté. Na zasedání, jehož podtitul zněl „Quo vadis stochastica?“, vystoupili kolegové z různých oblastí statistiky: za matematickou statistiku pohovořila prof. Marie Hušková (MFF UK), za státní statistiku vrchní ředitel sekce obecné metodiky a registrů ČSÚ pan ing. František Konečný, za medicínskou a biostatistiku RNDr. Marek Malý, o aplikacích statistiky v průmyslu pohovořil doc. Gejza Dohnal. Za zahraniční hosty vystoupil doc. Josef Chajdiak, současný předseda Slovenské štatistické a demografické spoločnosti. Škoda jen, že řada předem ohlášených účastníků se na poslední chvíli omluvila z nejrůznějších důvodů. V rámci společenského večera v příjemném prostředí rekonstruovaného Starého pivovaru (v areálu FII VUT Brno) a za cimbálové muziky bratří Wimmerovců ze Slovenska, byly předány Pamětní listy hostům a zakládajícím členům společnosti.

V září 2010 se zástupci společnosti (Dohnal, Řezanková) zúčastnili také oslav 150. výročí založení francouzské statistické společnosti v Institutu Henri Poincaré v Paříži. Při této příležitosti se konala schůzka zástupců čtrnácti evropských statistických společností. Hlavními téma úloha národních statistických společností při vzdělávání na základních a středních školách, o potřebě certifikace statistiků, o podpoře mladých statistiků a v neposlední řadě i o nutnosti vytvořit jakousi zastřešující instituci pro evropské národní společnosti. Jednání bude pokračovat v tomto roce v Dublinu při příležitosti 58. světového statistického kongresu ISI.

Na podzim 2010 se konalo také další, v pořadí již šesté společné zasedání zástupců skupiny V6 (Maďarsko, Rakousko, Česko, Slovensko, Slovinsko a Rumunsko), které tentokrát svolala Rakouská statistická společnost do Vídně. Zasedání se zúčastnil předseda společnosti a první místopředsedkyně prof. Řezanková. Na zasedání se hovořilo o otázkách projednávaných v Paříži. Byla diskutována možnost založení společné webové stránky skupiny V6 a zástupci společností se zavázali k vydávání společného periodika.

3. Odborná aktivita společnosti

- Dne 28. 1. 2010 se konala v budově VŠE v Praze Valná hromada společnosti, na které byl zvolen předseda a výbor společnosti. Na Valné hromadě přednesl odbornou přednášku doc. RNDr. Jan Picek, CSc.
- 31. 1. – 5. 2. 2010 naše společnost organizovala spolu s KPMS MFF UK a JČMF již 16. konferenci ROBUST 2010. Tentokrát v „zimní“ verzi na Hoře matky Boží v Králíkách. Tak jako v minulosti, i tentokrát byl ROBUST věnován vybraným trendům matematické statistiky, teorie pravděpodobnosti a analýzy dat. Počet účastníků ze čtyř evropských zemí (ČR, SR, Švýcarska a Velké Britanie) překročil stovku. V tomto roce však poprvé nemá ROBUST vlastní sborník. Část příspěvků vyšla v časopise Acta Universitatis Carolinae a část v časopise Informační Bulletin, číslo 3, ročník 21, 2010.
- V prvním zářijovém týdnu se konaly v Brně dvě akce současně. Ve dnech 2. – 3. 9. se konala v Akademii Sting konference o průmyslové statistice REQUEST, kterou naše společnost už podruhé pořádala spolu s Centrem pro jakost a spolehlivost výroby CQR. Souběžně s touto akcí se 3. a 4. září konaly Brněnské statistické dny. 4. 9. odpoledne se v novobarokní aule VUT Brno konalo slavnostní zasedání ČStS ke 20. výročí založení společnosti.
- Třetí Mikuklášský statistický den ČStS zorganizovala dne 6. 12. 2010 na MFF UK v pražském Karlíně. Tohoto setkání se zúčastnila i nová předsedkyně ČSÚ Iva Ritschelová.
- Internetové stránky společnosti byly pravidelně udržovány a aktualizovány díky práci kolegy doc. Jiřího Žváčka. Bohužel, zatím nedošlo k původně plánované změně grafické úpravy těchto stránek.
- V roce 2010 vyšla čtyři čísla Informačního Bulletinu, z nichž čtvrté je vyrobeno až v lednu 2011 a je k dispozici na tomto valném shromáždění. Třetí číslo Informačního Bulletinu obsahuje část příspěvků z konference ROBUST 2010, která tentokrát nemá vlastní sborník.
- ČStS formálně spolupracovala na vydávání časopisu Statistika.

4. Plán aktivit pro rok 2011

- V září 2011 se bude konat společná konference se SŠDS STAKAN 2011.
- V srpnu se zúčastníme jednání v rámci 58. světového statistického kongresu ISI v Dublinu.
- V říjnu 2011 se připravuje další konference REQUEST 2011 v Praze.
- Na podzim se bude konat další setkání skupiny V6, tentokrát opět v Maďarsku.
- V rámci možností se budeme podílet na organizaci statistických konferencí u nás i v zahraničí.
- Mikuklášský den v prosinci v Praze.

5. Návrh na funkci místopředsedy ex oficio

Naše společnost od svého založení v roce 1990 má jako jeden z hlavních cílů propojení statistiků všech zaměření a oblastí aplikace. Především potom užší spolupráci pracovníků z oblasti statistiky aplikované v ekonomických, společenských, přírodních a technických vědách, z oblasti teorie matematické statistiky a z oblasti státní statistiky.

Státní (oficiální) statistika má v naší zemi, stejně jako v ostatních evropských zemích) svého nejvyššího představitele v osobě předsedy národního statistického úřadu. Vědomi si důležitosti státní statistiky pro rozvoj a propagaci statistiky jako celku a důležitosti funkce předsedy ČSÚ a současně i jeho pracovního zatížení, výbor navrhuje ustanovení funkce stálého místopředsedy České statistické společnosti ex oficio, která by byla nabídnuta předsedovi ČSÚ. Tento ji může nebo nemusí přijmout. Tato funkce není volenou funkcí, nenese s sebou žádné povinnosti odpovídající funkci voleného místopředsedy, nicméně poskytuje svému nositeli právo účastnit se jednání výboru a při zvláštních příležitostech, jako je například setkání zástupců národních společností, předávání cen ve studentských soutěžích a podobně, vystupovat vedle předsedy a voleného místopředsedy jako zástupce společnosti. V předchozích letech byl předseda ČSÚ zvoleným členem výboru naší společnosti, nicméně povinnosti, plynoucí z jeho úřadu mu nedovolovaly se plně zapojovat do práce výboru. Ustanovení navrhované čestné funkce by deklarovalo důležitost sepjetí ČStS se státní statistikou a současně by nezatěžovalo nositele této funkce povinnostmi člena výboru.

V Praze, dne 28. 1. 2010

Doc. RNDr. Gejza Dohnal, CSc.
předseda společnosti

INTRODUCTION TO RANDOM MATRICES
ÚVOD DO NÁHODNÝCH MATIC

Martin Veselý

Adresa: ČVUT, FJFI, KSE, Trojanova 13, 120 00 Praha 2

E-mail: veselm21@fjfi.cvut.cz

Abstract: The random matrix is matrix with elements from certain distribution. Since eigenvalues are functions of the matrix elements, they are random variables too. Assume the hermitian matrix, then the distance between ordered eigenvalues can be defined because of they are real. Distances of eigenvalues are also random variables. In this arcicle, the distribution of eigenvalues and eigenvalues distances are presented for certain ensembles of the hermitian random matrices with gaussian distributed elements.

Keywords: random matrix, hermitian matrix, quaternion, Heaviside step function, Gaussian orthogonal/unitary/symplectic ensemble, band random matrix ensemble, Wigner's Semicircle Law.

Abstrakt: Náhodnou maticí rozumíme takovou matici, jejíž prvky jsou tvořeny náhodnými veličinami s určitým rozdělením. Jelikož vlastní čísla matice jsou funkcemi prvků, jsou také ona náhodná, a má smysl zkoumat jejich pravděpodobnostní rozdělení. Pokud je matice navíc hermitovská, její vlastní čísla jsou reálná, lze tudíž definovat vzdálenost uspořádaných vlastních čísel a opět zkoumat její pravděpodobnostní rozdělení. Tento příspěvek dává odpověď na otázku, jaký tvar mají tato rozdělení pro různé typy náhodných hermitovských matic s gaussovsky rozdělenými prvky.

Klíčová slova: náhodné matice, hermitovské matice, kvaternion, Heavisideova funkce, GOE, GUE, GSE, BRME, Wignerův polokruhový zákon.

1. Úvod

První zmínky o náhodných maticích se objevují ve 30. letech 20. století. Tehdy však zůstávájí na okraji zájmu, jelikož neexistovala výkonná výpočetní technika, která by umožňovala studium jejich vlastností. Většímu zájmu se náhodné matice těší od 50. let v souvislosti s pracemi Eugena Wignera v oblasti matematické fyziky. Ten využívá spekter náhodných matic pro aproximaci spekter hamiltoniánu (diferenciální operátor používaný v kvantové fyzice pro popis energetických stavů částic) jader těžkých prvků. V letech osmdesátých

Doručeno redakci: 7. 12. 2010, imprimatur: 12. 2. 2011.
MSC2010: 15B52, DOI: 10.5300/IB/2011-1/5

byla objevena vazba mezi náhodnými maticemi a teorií chaosu. Od této chvíle se náhodné matice používají jako podklad pro simulaci chaotických jevů ve fyzice, ekonomii, biologii, dopravě a mnoha další oborech.

Není bez zajímavosti, že existuje spojení mezi náhodnými maticemi a pravděpodobnostním rozdělením imaginárních částí netriviálních nul Riemannovy zeta funkce za předpokladu platnosti Riemannovy hypotézy.

Zmiňme se ještě o použití náhodných matic pro modelování dopravních proudů. Lze totiž ukázat, že rozdělení vzdáleností vozidel pohybujících se na jednoproudé silnici je velmi podobné rozdělení vzdáleností uspořádaných vlastních čísel hermitovských náhodných matic.

2. Základní pojmy

V této části zavedeme některé pojmy, se kterými budeme dále pracovat. Maticí *hermitovsky sdruženou* k matici $\mathbf{A}$ rozumíme matici transponovanou a zároveň komplexně sdruženou. Hermitovsky sdruženou matici značíme $\mathbf{A}^H$. O matici $\mathbf{A}$ říkáme, že je

- *symetrická* $\Leftrightarrow \mathbf{A} \in \mathbb{R}^{n,n}$, $\mathbf{A} = \mathbf{A}^T$
- *antisymetrická* $\Leftrightarrow \mathbf{A} \in \mathbb{R}^{n,n}$, $\mathbf{A} = -\mathbf{A}^T$
- *hermitovská* $\Leftrightarrow \mathbf{A} \in \mathbb{C}^{n,n}$, $\mathbf{A} = \mathbf{A}^H$
- *ortogonální* $\Leftrightarrow \mathbf{A} \in \mathbb{R}^{n,n}$, $\mathbf{A}\mathbf{A}^T = \mathbf{A}^T\mathbf{A} = \mathbf{I}$
- *unitární* $\Leftrightarrow \mathbf{A} \in \mathbb{C}^{n,n}$, $\mathbf{A}\mathbf{A}^H = \mathbf{A}^H\mathbf{A} = \mathbf{I}$
- *symplektická* vzhledem k matici $\mathbf{\Omega}$
  $\Leftrightarrow \mathbf{A} \in \mathbb{C}^{2n,2n}, \mathbf{\Omega} \in \mathbb{R}^{2n,2n}, \mathbf{\Omega} = -\mathbf{\Omega}^T, h(\mathbf{\Omega}) = 2n : \mathbf{A}^T\mathbf{\Omega}\mathbf{A} = \mathbf{\Omega}$

Poznamenejme, že v dalším textu $n \in \mathbb{N}$ značí řád matice.

Dalším pojmem, o kterém se zmíníme je *kvaternion*. Kvaternionem rozumíme hyperkomplexní číslo tvaru $q = r + x\mathtt{i} + y\mathtt{j} + z\mathtt{k}$, kde $r, x, y, z \in \mathbb{R}$. Kvaternion $q^* = r - x\mathtt{i} - y\mathtt{j} - z\mathtt{k}$ nazýváme *konjugovaný* ke kvaternionu q. Kvaternionovou matici, pro kterou platí $a_{ij} = a_{ji}^*$ nazveme hermitovskou. Vlastní čísla hermitovských matic jsou reálná[1].

[1] Úplně korektně bychom v případě kvaternionových matic měli říci pravá vlastní čísla, neboť díky faktu, že kvaternionové násobení není komutativní, má kvaternionová matice pravé a levé spektrum.

Na závěr tohoto oddílu ještě definujme *Heavisideovu funkci*

$$\theta(x) = \begin{cases} 0 & \text{pro } x \leq 0 \\ 1 & \text{pro } x > 0. \end{cases} \tag{1}$$

Tuto funkci s výhodou využijeme k zjednodušenému zápisu hustot pravděpodobnosti.

3. Skupiny náhodných matic

Nyní se podívejme na některé specifické skupiny hermitovských náhodných matic.

- **Matice GOE** – *Gaussian orthogonal ensemble*
  Jedná se o symetrické matice s prvky z normálního rozdělení, přičemž pro rozdělení prvků platí následující pravidla:
  - $a_{ij} \sim \mathcal{N}(\mu, \sigma^2)$ pro $i \neq j$
  - $a_{ij} \sim \mathcal{N}(\mu, 2\sigma^2)$ pro $i = j$

  Matice jsou dále invariantní vůči transformaci ortogonální maticí $\mathbf{U}$, tzn.
  $$\mathbf{A} \in \text{GOE} \Rightarrow \mathbf{U^T A U} \in \text{GOE}$$

- **Matice GUE** – *Gaussian unitary ensemble*
  Jde o skupinu komplexních hermitovských matic s prvky s normálním rozdělením. Pro parametry rozdělení reálných částí prvků platí stejná pravidla jako v případě matic GOE. Imaginární části diagonálních prvků jsou nulové díky hermiticitě matice. V případě mimodiagonálních prvků, mají jejich imaginární části stejné rozdělení jako reálné. Matice jsou dále invariantní vůči transformaci unitární maticí $\mathbf{U}$, tzn.
  $$\mathbf{A} \in \text{GUE} \Rightarrow \mathbf{U^H A U} \in \text{GUE}$$

- **Matice GSE** – *Gaussian symplectic ensemble*
  Jedná se o skupinu kvaternionových hermitovských matic. Pro rozdělení reálných částí prvků platí stejná pravidla jako v případě matic typu GOE. Imaginární části mimodiagonální prvků jsou rozděleny stejně jako jejich reálné části. Matice této skupiny jsou invariantní vůči podobnostní transformaci symplektickou maticí $\mathbf{U}$, tzn.
  $$\mathbf{A} \in \text{GSE} \Rightarrow \mathbf{U^{-1} A U} \in \text{GSE}$$

Jelikož existují symplektické matice pouze sudého řádu, také řád matic třídy GSE je vždy sudý.

- **Matice BRME** – *Band random matrix ensemble*
  Tato skupina je tvořena pásovými symetrickými maticemi. Každá z matic této skupiny je charakterizována tzv. *položířkou pásu* (*angl. band half–width*), ozn. b. Pro prvky matice platí $a_{ij} = 0 \Leftrightarrow |i-j| \geq b$, přičemž $1 \leq b \leq n$. Rozdělení nenulových prvků je shodné se skupinou GOE. Matice nevykazují neměnnost vůči podobnostním transformacím.

V případě všech tříd náhodných matic musí být diagonální prvky a prvky nad diagonálou statisticky nezávislé náhodné veličiny. Prvky pod diagonálou jsou naopak závislé na prvcích nad diagonálou, neboť se jedná o vzájemně konjugované hodnoty (resp. přímo kopie v případě reálných matic).

Poznamenáváme, že dále uvedené zákony platí pro matice, pro něž $\mu = 0$ a $\sigma^2 = 1$.

4. Rozdělení vlastních čísel

V této části popíšeme rozdělení vlastních čísel jednotlivých, výše jmenovaných, skupin náhodných matic. Z numerických experimentů vyplývá, že rozdělení je popsáno hustotou pravděpodobnosti

$$f(\lambda) = \theta(\rho - |\lambda|) C \sqrt{\rho^2 - \lambda^2}, \tag{2}$$

kde C představuje normalizační konstantu zajišťující, že výše uvedený vztah je hustotou pravděpodobnosti a ρ je *spektrální poloměr* matice (tj. nejvyšší absolutní hodnota vlastního čísla). Pro normalizační konstantu platí

$$C = \frac{2}{\pi \rho^2}. \tag{3}$$

Spektrální poloměr závisí na typu matice. Platí

$$\rho_{\mathrm{GOE}} = 2\sqrt{n} \tag{4}$$

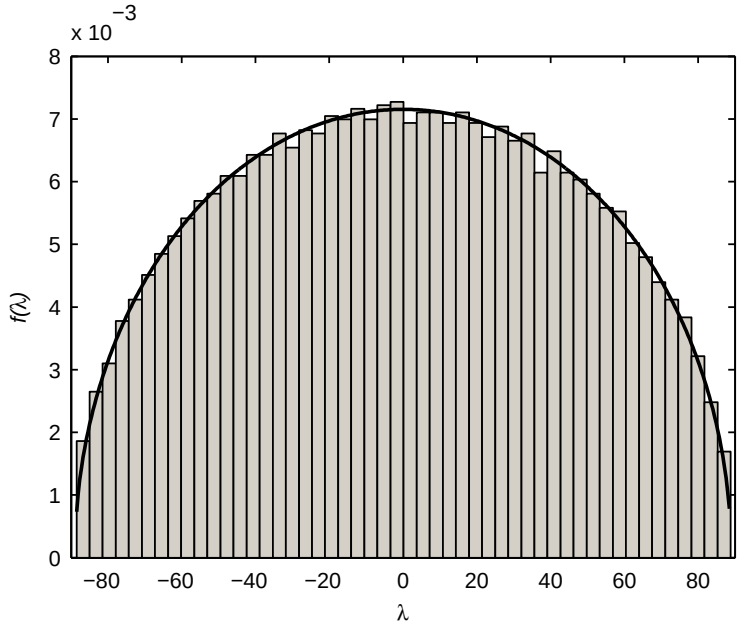
$$\rho_{\mathrm{GUE}} = 3\sqrt{n} \tag{5}$$

$$\rho_{\mathrm{GSE}} = 4\sqrt{n} \tag{6}$$

$$\rho_{\mathrm{BRME}} = \sqrt{\frac{4b}{n}(2n - b + 1)}. \tag{7}$$

Vztah pro matice třídy BRME a GOE byl převzat z [5]. Pro třídy GUE a GSE byl určen numericky pomocí metody Monte Carlo. Získaná data byla následně

zpracována pomocí nelineární regrese. Poznamenejme, že vztah (2) nazýváme *Wignerův polokruhový zákon*. Srovnání empirických dat s předpovědí Wignerova zákona pro matice třídy GSE řádu 500 shrnujeme na obrázku 1.



Obrázek 1: Wignerův polokruhový zákon pro matice třídy GSE řádu 500.

Jelikož již známe hustotu pravděpodobnosti vlastních čísel, můžeme určit další charekteristiky jejich rozdělení. Vzhledem k tomu, že (2) je sudá funkce, je zřejmé, že všechny liché obecné momenty rozdělení jsou nulové, neboť jsou reprezentovány antisymetrickým integrálem. Rozptyl je díky nulovosti střední hodnoty (první obecný moment) roven druhému obecnému momentu. Pro jednotlivé skupiny matic platí

$$\sigma^2_{\mathrm{GOE}} = n \tag{8}$$

$$\sigma^2_{\mathrm{GUE}} = 2n \tag{9}$$

$$\sigma^2_{\mathrm{GSE}} = 4n \tag{10}$$

$$\sigma^2_{\mathrm{BRME}} = \frac{b}{n}(2n - b + 1). \tag{11}$$

Poznamenejme, že poslední ze vztahů byl převzat z článku [4], jež se zabývá právě pásovými náhodnými maticemi.

5. Rozdělení vzdáleností vlastních čísel

Popišme nejprve tzv. *unfolding* spektra matice. Unfolding je proces, během něhož jsou hodnoty libovolné spojité náhodné veličiny přeškálovány tak, aby

hustota pravděpodobnosti nově vzniklé veličiny odpovídala rozdělení rovnoměrnému. Proces unfoldingu si lze lépe představit na histogramu. V tomto případě provedeme škálování změnou šířky jednotlivých intervalů histogramu tak, že každý z intervalů obsahuje stejné procento počtu pozorování. Histogram tak bude tvarově odpovídat rovnoměrně rozdělené veličině. V našem případě provádíme unfolding ořízntutím výběrového souboru shora a zdola tzn. vynecháme určité množství nejvyšších hodnot souboru a stejné množství hodnot nejnižší. Jelikož vlastní čísla náhodných matic jsou rozdělena podle Wignerova polokruhového zákona, způsobí výše popsaný ořez přechod k přibližně rovnoměrnému rozdělení, neboť část polokružnice blízko vertikální osy je málo zakřivena a svým tvarem se blíží úsečce. Získáváme tedy přibližně rovnoměrně rozdělenou veličinu.

Získaná unfoldovaná vlastní čísla uspořádáme dle velikosti a následně určíme rozdíly sousedních. Tyto rozdíly dále normujeme tak, aby střední hodnota vzdálenosti byla rovna jedné[2]. Tím získáme realizace náhodné veličiny zvané *vzdálenost uspořádaných vlastních čísel.* Podívejme se, jak vypadá rozdělení této veličiny[3]. Přesný tvar hustoty pravděpodobnosti není dosud znám. Existují však různé více či méně kvalitní aproximace. První z těchto aproximací je tzv. *Wignerova domněnka*

$$f(r) \approx \theta(r) A r^{\beta} \mathrm{e}^{-Br^2}, \tag{12}$$

kde A a B představují normalizační konstanty zajišťují, že vztah (12) je hustotou pravděpodobnosti a střední hodnota je rovna jedné. Parametr β charakterizuje strukturu matice. V případě matic GOE, GUE a GSE nabývá hodnot 1, 2 a 4 v uvedeném pořadí. Pro matice třídy BRME jej určíme podle přibližného vztahu (převzato z [5])

$$\beta_{\mathrm{BRME}} \approx \frac{1{,}4b^2}{1{,}4b^2 + n}. \tag{13}$$

Hodnoty parametrů A a B závisejí na β, a platí pro ně (viz literaturu [1])

$$A(\beta) = 2\frac{\left(\Gamma\left(\frac{\beta+2}{2}\right)\right)^{\beta+1}}{\left(\Gamma\left(\frac{\beta+1}{2}\right)\right)^{\beta+2}} \quad B(\beta) = \frac{\left(\Gamma\left(\frac{\beta+2}{2}\right)\right)^{2}}{\left(\Gamma\left(\frac{\beta+1}{2}\right)\right)^{2}}. \tag{14}$$

Poznamenejme, že Wignerova domněnka přesně popisuje rozdělení vzdáleností vlastních čísel matic řádu 2.

[2]Normalizaci provádíme kvůli srovnatelnosti vzdáleností vlastních čísel různých typů náhodných matic.

[3]Rozdělení je v anglicky psané literatuře známo pod označením *spacing distribution.*

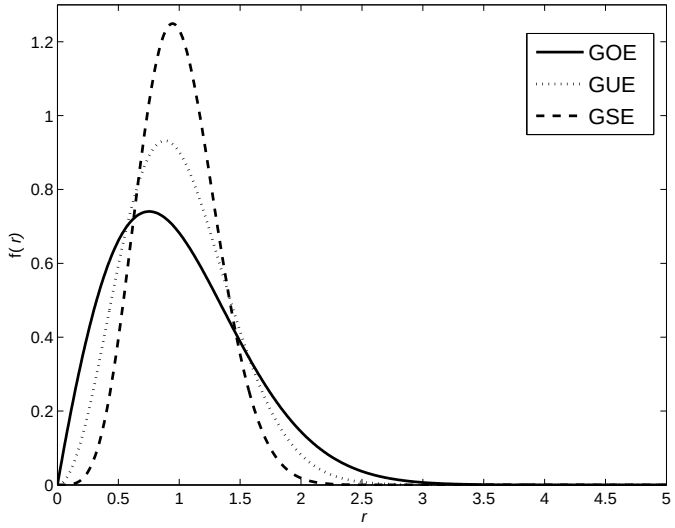
Podívejme se ještě na další, kvalitnější, odhad zmíněné hustoty pravděpodobnosti, totiž na tzv. *Izrailevovu formuli*

$$f(r) \approx \theta(r) A \left(\frac{\pi r}{2}\right)^{\beta} \mathrm{e}^{-\frac{\beta\pi^2}{16} r^2 - \left(B - \frac{\beta\pi}{4}\right) r}. \tag{15}$$

Význam parametrů je stejný jako v případě Wignerovy domněnky. Výpočet hodnot A a B je však nepoměrně složitější a vyžaduje užití numerických metod. V případě matic třídy GOE, GUE, resp. GSE parametr A nabývá hodnot 1,198, 1,369 resp. 1,551, dále B je rovno 1,183, 1,658 resp. 2,711[4].

Zaměřme se ještě na speciální případ, kdy $\beta = 0$ (pak $A = B = 1$). Ten odpovídá diagonální náhodné matici (tj. s položířkou pásu $b = 1$). Rozdělení vzdáleností vlastních čísel je pak exponenciální s hustotou pravděpodobnosti $f(r) = \mathrm{e}^{-r}$.

Na závěr tohoto oddílu uveďme průběhy popsané Izrailevovou formulí pro matice tříd GOE, GUE a GSE (obrázek 2) a výsledek numerické simulace pro matice třídy GSE řádu 1000 (obrázek 3).

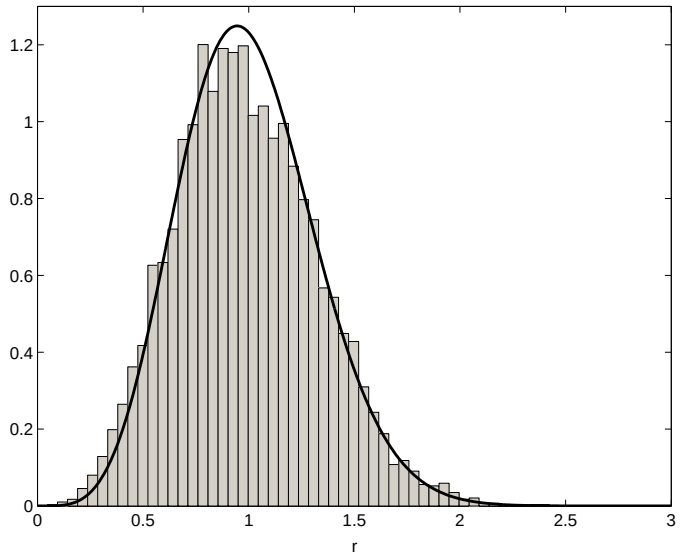


Obrázek 2: Průběhy hustoty pravděpodobnosti vzdálenosti uspořádaných vlastních čísel popsané Izrailevovou formulí.

6. Závěr

Tento článek představuje lehký nástin teorie náhodných matic. Zavedli jsme několik tříd hermitovských gaussovských náhodných matic a zkoumali některé statistické vlastnosti jejich spekter. Uvědomme si však, že se jedná pouze o úplné základy studia tohoto velmi zajímavého fenoménu. Existuje značné

[4]Zde prezentované hodnoty pocházejí z literatury [2].



Obrázek 3: Srovnání empirických dat s průběhem hustoty pravděpodobnosti popsaným Izrailevovou formulí pro matice třídy GSE řádu 1000.

množství další typů náhodných matic, které již nutně nemusejí být hermitovské, a dokonce nemusejí mít ani gaussovsky rozdělené prvky. Je zavedena také řada další statistických charakteristik spekter náhodných matic. Ty poznatky se v budoucnu budeme snažit přinést také na stránkách tohoto bulletinu.

Literatura

[1] NIEMINEN J. M.: *Eigenvalue spacing statistics of a four-matrix model of some four-by-four random matrices*, J. Phys. A: Math. and Theoretical, vol. 42, 2009. ISSN 1751-8113. doi: 10.1088/1751-8113/42/3/035001

[2] IZRAILEV F. M., SCHARF R.: *Dyson's Coulomb gas on a circle and intermediate eigenvalue statistics*, J. Phys. A: Math. and general, vol. 23, no. 6, pp. 963–977, 1990. ISSN 0305-4470.
doi: 10.1088/0305-4470/23/6/018

[3] MEHTA M. L.: *Random matrices*, 3. vyd., Amsterdam: Elsevier/Academic Press, 2004. ISBN 0-12-088409-7.

[4] CASATI G., IZRAILEV F. M. a MOLINARI L.: *Scaling properties of the eigenvalue spacing distribution for band random matrices*, J. Phys. A: Math. and General, vol. 24, no. 20, pp. 475–476, 1991.
ISSN 0305-4470. doi: 10.1088/0305-4470/24/20/011

[5] KRBÁLEK M.: *Traffic systems – particle gases in thermal equilibrium (Random Matrix Theory approach) (disertační práce)*, ČVUT v Praze, Fakulta jaderná a fyzikálně inženýrská, 2003.

ROBUST MULTIVARIATE STATISTICS IN GENETIC APPLICATIONS

ROBUSTNÍ MNOHOROZMĚRNÁ STATISTIKA V GENETICKÝCH APLIKACÍCH

Jan Kalina

Adresa: Centrum biomedicínské informatiky, Ústav informatiky AV ČR, Pod Vodárenskou věží 2, 182 07 Praha 8

E-mail: kalina@euromise.cz

Abstract: The paper describes possible applications of robust statistical methods in genetic research. Standard approaches for the analysis of images measured by the microarrays technology turn out to be too sensitive with respect to outliers in the data. Therefore in a genetic study at the Centre of Biomedical Informatics we apply robust statistical methods to image analysis and classification analysis of gene expression measurements. Further we describe the MWCD estimator of multivariate location and scatter, which is used to obtain a robust classification analysis method based on implicit weighting of individual observations.

Keywords: robust statistics, multivariate statistics, genetic research, MWCD method, Next-Gen Sequencing.

Abstrakt: Článek popisuje možnosti použití robustních statistických metod v genetickém výzkumu. V obecné situaci jsme zjistili, že standardní postupy pro zpracování obrazové informace měřené technologií *microarrays* jsou příliš citlivé vůči přítomnosti odlehlých pozorování. V konkrétní studii, která probíhá v Centru biomedicínské informatiky, proto aplikujeme metody robustní statistiky na analýzu obrazu a na klasifikační analýzu pro zpracování naměřených genových expresí. Dále zde popíšeme odhad metodou MWCD pro střední hodnotu a varianční matici mnohorozměrných dat, s jehož pomocí získáme robustní metodu pro klasifikační analýzu založenou na implicitním vážení jednotlivých pozorování.

Klíčová slova: robustní metody, vícerozměrná statistika, genetický výzkum, metoda MWCD, sekvenace.

Doručeno redakci: 9. 11. 2010, imprimatur: 12. 1. 2011.
MSC2010: 62H35, DOI: 10.5300/IB/2011-1/13

1. Robustní analýza obrazu pro hodnocení genetických studií

V Centru biomedicínské informatiky (CBI) probíhá studie genových expresí u pacientů s kardiovaskulárními onemocněními. Projekt může přinést výsledky aplikovatelné do lékařské diagnostiky a terapie tím spíše, že kardiovaskulární onemocnění jsou hlavní příčinou úmrtí v České republice. V této kapitole se zaměříme na obecný postup zpracování dat získaných v molekulárně genetických studiích za pomoci *microarrays*. Typicky se naměřené genové exprese analyzují standardními automatickými procedurami, které jsou citlivé vůči přítomnosti odlehlých pozorování [3]. V současné době pracujeme na vývoji alternativních postupů, které jsou založený na robustních statistických metodách.

Genetický výzkum v rámci CBI má za cíl najít optimální sadu genů pro diagnostiku a prognózu kardiovaskulárních onemocnění. Používá celogenomové technologie *microarrays* (mikročipy) pro měření genových expresí. Městská nemocnice Čáslav odebírá vzorky periferní krve pacientům s ischemickým onemocněním (akutní infarkt myokardu nebo cévní mozková příhoda) a také kontrolním osobám, což jsou pacienti hospitalizovaní s jinou příčinou než s manifestovaným ischemickým onemocněním. Z krve se izoluje ribonukleová kyselina (RNA) a aplikuje se na mikročip typu *BeadArray*, který je založen na náhodném rozmístění mikroskopických kuliček odpovídajících různým lidským genům na povrchu mikročipu. Zatímco lidský genom obsahuje asi 23 000 genů, díky použití různých transkriptů k týmž genům se získají genové exprese pro celkový počet 48 701 druhů kuliček (různých transkriptů). Mikročipy jsou naskenovány a výsledkem měření je proto obrazová informace odpovídající genovým expresím. Nejde přitom o složení genetického kódu obsaženého v chromozómech, ale o monitorování biologických procesů, které odpovídají aktivitě (syntéza proteinů) nebo neaktivitě genů v daném okamžiku odběru krve.

Originálními daty získanými při měření genových expresí pomocí *microarrays* jsou naskenované obrazy, v nichž vyšší fluorescenční intenzita odpovídá genům s velkou expresí. Standardní postup pro zpracování obrazové informace [3] načte data pomocí speciálních funkcí pro čtení datového souboru ve dvoubytovém formátu v jazyce C++. V našem případě se jedná o obrazy o velikosti 2389×18309 pixelů. Obrazová informace se standardně zpracuje pomocí posloupnosti transformací, která zahrnuje lokální odhad intenzity pozadí v okolí každé kuličky, dále zaostření obrazu a jeho vyhlazení, odhad intenzity po odstranění vlivu pozadí, normalizaci dat a odstranění odlehlých hodnot. První kroky jsou však silně ovlivněny lokálním šumem v okolí

jednotlivých kuliček a výsledné vychýlené hodnoty jsou předány i do dalších fází analýzy. Odstranění odlehlých hodnot se provádí až na samém konci celé procedury. Proto je analýza diferenciálních expresí citlivá k náhodným nebo systematickým chybám v původních datech.

V Centru biomedicínské informatiky pracujeme na návrhu robustní alternativy ke zpracování naskenovaných obrazů s měřením genových expresí, která zahrnuje hledání systematických artefaktů v datech. V každém kroku procedury se odstraňují odlehlé hodnoty za pomoci metod založených na robustní statistice; zde jde zejména o aplikaci metody nejmenších vážených čtverců [8] aplikovanou na analýzu obrazu. Normalizace dat se provádí teprve po odstranění odlehlých hodnot. Zároveň lze tyto metody počítat rychlými algoritmy. Je žádoucí, aby se modifikoval i standardní software [3] pro analýzu genových expresí pomocí robustních metod.

Po očištění dat a jejich normalizaci uvažujeme lineární model, který obsahuje řádově desítky pacientů a 48 701 proměnných, což jsou průměrné exprese jednotlivých genů (transkriptů). Testování hypotéz ukáže, které geny mají významně odlišné diferenciální exprese u pacientů s akutním infarktem myokardu nebo cévní mozkovou příhodou v porovnání s kontrolními osobami. Klinická a biochemická data příslušná každému pacientovi přispívají k porozumění genetickým predispozicím pro kardiovaskulární onemocnění. Cílem studie genových expresí v Centru biomedicínské informatiky je patentovat optimální sadu genů, která umožňuje diagnostiku, prognózu a terapii příslušných kardiovaskulárních onemocnění. Tyto geny lze následně použít na oligonukleotidový čip; jde tedy o příspěvek k rozvoji personalizované a prediktivní lékařské péče v souladu s novým paradigmatem medicíny založené na důkazech [7].

Pro účely klasifikace nového pacienta do jedné ze skupin (pacient vs. kontrolní osoby; těžká vs. lehká forma onemocnění) je žádoucí používat robustní přístupy ke klasifikační analýze. Jednu takovou metodu dále popíšeme.

2. Robustní klasifikační analýza

V posledních letech bylo navrženo několik robustních metod pro klasifikační analýzu. [4] studuje robustní odhady střední hodnoty a varianční matice pro mnohorozměrná data a následně jimi nahradí výběrový průměr a varianční matici v předpisu pro lineární či kvadratickou diskriminanční analýzu. Jiným příkladem je [1], který používá lineární klasifikační analýzu při zpracování medicínské obrazové informace a navrhuje modifikovat běžné klasifikační postupy pomocí *shrinkage* přístupu.

V této kapitole navrhneme odhad MWCD (*minimum weighted covariance determinant*) jako váženou obdobu odhadu MCD (*minimum covariance determinant*) [6]. Myšlenka implicitního přiřazení vah se již osvědčila v metodě nejmenších vážených čtverců (*least weighted squares, LWS*) [8], což je metoda pro odhad parametrů v lineární regresi, který má velký bod selhání. Navíc je pravda, že implicitně vážené odhady netrpí lokální senzitivitou vůči malým změnám v centru dat, která je typická pro odhady založené na úplném ignorování odlehlých hodnot. Odhad MWCD následně použijeme pro robustifikaci klasifikační analýzy. Nakonec ilustrujeme použití MWCD odhadu při klasifikaci v kontextu analýzy obrazu obličeje.

Uvažujeme nejprve náhodný výběr $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^T$ z p-rozměrného normálního rozdělení. Pokud uvažujeme pevné váhy

$$w_1, \ldots, w_n, \quad \sum_{i=1}^{n} w_i = 1,$$

můžeme označit pomocí $\overline{\mathbf{X}}_w$ vážený průměr $\overline{\mathbf{X}}_w = \sum_{i=1}^{n} w_i \mathbf{X}_i$ a pomocí $\mathbf{S}_w$ váženou varianční matici

$$\mathbf{S}_w = \sum_{i=1}^{n} w_i (\mathbf{X}_i - \overline{\mathbf{X}}_w)(\mathbf{X}_i - \overline{\mathbf{X}}_w)^T. \tag{16}$$

MWCD odhad definujeme jako odhad mnohorozměrné polohy a měřítka s vysokým bodem selhání. Málo spolehlivým pozorováním budou přiřazeny malé váhy. Zde zvolíme jen velikosti vah

$$w_1 \geq w_2 \geq \cdots \geq w_n, \tag{17}$$

protože váhy budou přiřazeny jednotlivým pozorováním až implicitně v průběhu výpočtu MWCD odhadu. Jednou z možností je zvolit lineárně klesající váhy, což je oblíbená volba také pro LWS metodu v lineární regresi. Odhad metodou MWCD odhaduje zároveň parametr polohy i varianční matici, a to jako vážený průměr $\overline{\mathbf{X}}_{MWCD}$ a váženou varianční matici $\mathbf{S}_{MWCD}$ s takovými vahami, které vedou k minimálnímu determinantu vážené varianční matice přes všechny možné permutace uvažovaných vah (17) a přes všechny možné hodnoty odhadu parametru polohy, kterými se nahradí vážený průměr v (16). MWCD odhad lze vypočítat modifikací aproximativního algoritmu [6].

Odhad MCD [6] lze považovat za speciální případ MWCD odhadu zcela ignorující hodnoty, které považuje za odlehlé. Uvažuje tedy váhy rovné 1 nebo 0 s tím, že předem zvolený pevný počet h pozorování má váhu rovnou 1.

Popišme myšlenku robustní klasifikace založené na MWCD odhadu. Následujme myšlenku robustní klasifikační analýzy [4]. Uvažujme mnohorozměrná data v celkovém počtu J skupin. Označme tato data pomocí

$$\mathbf{X}_{1i}, \ldots, \mathbf{X}_{1n_1}, \mathbf{X}_{2i}, \ldots, \mathbf{X}_{2n_2}, \ldots, \mathbf{X}_{Ji}, \ldots, \mathbf{X}_{Jn_J}. \tag{18}$$

Robustní analogii lineární klasifikační analýzy založenou na MWCD založíme na robustních odhadech pro průměr v j-té skupině (pro $j = 1, \ldots, J$), který označme $\overline{\mathbf{X}}_{j,MWCD}$. Předpokládáme stejné varianční matice v jednotlivých skupinách a označíme robustní odhad varianční matice spočtené ze všech pozorování napříč skupinami jako $\mathbf{S}_{MWCD}$. Robustní lineární klasifikační analýzu definujme předpisem, který přiřadí nové pozorování $\mathbf{Z}$ do j-té skupiny, pokud robustní diskriminační skór

$$d_j = \overline{\mathbf{X}}_{j,MWCD}^T \mathbf{S}_{MWCD}^{-1} \mathbf{Z} - \frac{1}{2} \overline{\mathbf{X}}_{j,MWCD}^T \mathbf{S}_{MWCD}^{-1} \overline{\mathbf{X}}_{j,MWCD} + \log p_j$$

je roven $\max\{d_1, \ldots, d_J\}$. Zde vystupují i apriorní pravděpodobnosti p_j, že nové pozorování bude patřit do j-té skupiny.

Jde o lineární klasifikační pravidlo založené na robustifikaci Mahalanobisovy vzdálenosti každého pozorování od (robustního) odhadu střední hodnoty dat v každé skupině. Obdobně lze definovat i robustní kvadratickou klasifikaci založenou na MWCD odhadu. Ukazuje se, že odhad je velmi robustní pro vysoce kontaminované datové soubory a zároveň eficientní pro normální data bez kontaminace. Odhad také netrpí lokální senzitivitou, která sužuje odhady LTS a MCD. Jde o důsledek přiřazení malých ale kladných vah méně spolehlivým pozorováním [8], [5].

Použití odhadu metodou MWCD nyní ilustrujeme na úloze při zpracování obrazu obličeje. Zároveň jde o analýzu s genetickými aplikacemi, k jejímž cílům patří implementace systému pro podporu lékařské diagnostiky pro klasifikaci genetických pacientů s různými formami obličejové dysmorfie. Prvním úkolem při analýze je nicméně automatická lokalizace úst v databázi dvourozměrných černobílých obrazů obličejů, v níž každý obraz obsahuje právě jeden standardizovaný obraz obličeje.

Pracujeme zde s databází obrazů obličejů zdravých osob, která byla pořízena na Ústavu genetiky člověka Univerzity Duisburg-Essen (SRN) v rámci projektů DFG s kódy BO 1955/2-1 a WU 314/2-1. Databáze obsahuje 424 obrazů, z nichž 212 odpovídá ústům a 212 obsahuje jiné části obličeje nebo jiné objekty, kterým souhrnně říkejme neústa. Jde o matice velikosti 26×56 pixelů, které převedeme pro účely klasifikace na vektory délky $26 \times 56 = 1456$ pixelů. Naším cílem je diskriminovat mezi ústy a neústy.

Spočítali jsme 5 robustních hlavních komponent algoritmem *projection pursuit* [2] a aplikovali jsme MWCD-klasifikaci na databázi 424 obrazů. Výsledky ověříme tak, že klasifikujeme každý ze 424 obrazů v databázi. Tak jsme získali výsledky správné ve 100 % případů s použitím MWCD odhadu s lineárně klesajícími vahami. Ukazuje se, že pozorování s malými vahami se nacházejí na okraji úst nebo v oblasti pod ústy, směrem k bradě.

3. Závěr

Měření genových expresí prochází rapidním rozvojem. Nová dostupná technologie zvaná *Next-Gen Sequencing* (sekvenace) ze sebe chrlí obrovské datové soubory. Pro její rozvoj a použitelnost výsledků je rozhodující, zda budou k dispozici rychlé algoritmy pro výpočet robustních statistických metod. Navíc se výrobci *Next-Gen* technologií chystají implementovat systémy pro online zpracování měřených dat. Přitom považujeme za klíčové, aby se používaly metody ušité na míru pro jednotlivé aplikace a aby umožňovaly nastavit správnou úroveň jednotlivých parametrů.

Navržený odhad metodou MWCD jsme použili na reálná data v analýze obrazu obličeje. Ukazuje se, že klasifikace založená na MWCD odhadu je vhodná pro mnohorozměrná data s velkou dimenzí. Celkově můžeme vyslovit naději, že robustní metody najdou své uplatnění i v analýze dat s velkou dimenzí (např. při zpracování obrazové informace), a to při redukci dimenze, potlačení vlivu odlehlých pozorování a při klasifikační analýze.

Poděkování

Tato práce vznikla v rámci projektu 1M06014 Ministerstva školství, mládeže a tělovýchovy České republiky.

Literatura

[1] Aretusi G., Fontanella L., Ippoliti L., Merla A. (2010): Space-time texture analysis in thermal infrared imaging for classification of Raynaud's Phenomenon. In Mantovan P., Secchi P. (Eds.): *Complex data modeling and computationally intensive statistical methods (Contributions to Statistics)*. Springer, Milano, 1 – 12, 2010. ISBN 978-88-470-1385-8. doi: 10.1007/978-88-470-1386-5_1

[2] Croux C., Filzmoser P., Oliveira M. R. (2007): Algorithms for projection-pursuit robust principal component analysis. *Chemometrics and*

Intelligent Laboratory Systems, vol. 87, no. 2, pp. 218 – 225. ISSN 0169-7439. doi: 10.1016/j.chemolab.2007.01.004

[3] Dunning M. J., Barbosa-Morais N. L., Lynch A. G., Tavaré S., Ritchie M. E. (2008): Statistical issues in the analysis of Illumina data. *BMC Bioinformatics,* vol. 9, no. 85. ISSN 1471-2105. doi: 10.1186/1471-2105-9-85

[4] Hubert M., Rousseeuw P. J., van Aelst S. (2008): High-breakdown robust multivariate methods. In *Statistical Science,* vol. 23, no. 1, pp. 92 – 119. doi: 10.1214/088342307000000087

[5] Kalina J. (2010): Robust econometrics: diagnostic tools and multivariate methods. Zasláno do *Prague Economic Papers.*

[6] Rousseeuw P. J., Van Driessen K. (1999): A fast algorithm for the minimum covariance determinant estimator. In *Technometrics,* vol. 41, no. 3, pp. 212 – 223. ISSN 0040-1706. doi: 10.2307/1270566

[7] Tanaka H. (2010): Omics-based medicine and systems pathology. *Methods of Information in Medicine,* vol. 49, no. 2, pp. 173 – 185. ISSN 0026-1270. doi: 10.3414/ME9307

[8] Víšek J. Á. (2001): Regression with high breakdown point. In Antoch, J., Dohnal, G. (Eds.): *ROBUST 2000, Sborník prací 11. letní školy JČMF.* JČMF a Česká statistická společnost, Praha, 324 – 356.

CLUSTER ANALYSIS IN MATLAB
SHLUKOVÁ ANALÝZA V MATLABU

Martin Kovářík, Petr Klímek

Adresa: Tomas Bata University in Zlín, nám. T. G. Masaryka 5555, 760 01 Zlín, Czech Republic

E-mail: m1kovarik@fame.utb.cz, klimek@fame.utb.cz

Abstract: This paper focuses on methodical view of cluster analysis and detailed theoretical description of this conception, the purpose of which is to find similar properties and differences among objects and to cluster (group) them in groups (segments). The second practical part contains of practical application of cluster analysis to specific data using Matlab 2007b software.

Keywords: Matlab, Statistics Toolbox, Cluster Analysis, Clusters, Metric, Methods of Clustering, Dendrogram.

Abstrakt: Tento článek přináší metodický pohled na shlukovou analýzu a bližší teoretické seznámení s touto koncepcí, jejímž smyslem je nalezení podobných vlastností a rozdílů mezi objekty a jejich seskupování (shlukování) do skupin (segmentů). Závěrečnou, praktickou část tvoří konkrétní aplikace této analýzy na demonstračních datech za použití programového prostředí Matlab 2007b.

Klíčová slova: program Matlab, shluková analýza, shluky, metrika, metody shlukování, dendrogram.

1. Metodika

Podstata metody

Analýza shluků (Cluster analysis, CLU) patří mezi metody, které se zabývají vyšetřováním podobnosti vícerozměrných objektů tj. objektů, u nichž je změřeno větší množství znaků a následnou klasifikací objektů do shluků. Hodí se zejména tam, kde objekty projevují přirozenou tendenci se seskupovat. Podle způsobu shlukování se postupy dělí na hierarchické a nehierarchické. Hierarchické se dělí dále na aglomerativní a divizní.

Doručeno redakci: 8. 10. 2010, imprimatur: 6. 2. 2011.
MSC2010: 62H30, DOI: 10.5300/IB/2011-1/20

Hierarchické postupy

Jsou založeny na postupném spojování objektů a jejich shluků do dalších, větších shluků. Nejprve se vypočte základní matice vzdáleností mezi objekty. U aglomerativního shlukování se dva objekty, jejichž vzdálenost je nejmenší, spojí do prvního shluku a vypočte se nová matice vzdáleností, v níž jsou vynechány objekty z prvního shluku a naopak tento shluk je zařazen jako celek. Celý postup se opakuje tak dlouho, dokud všechny objekty netvoří jeden velký shluk nebo dokud nezůstane určitý, předem zadaný počet shluků. Divizní postup je obrácený. Vychází se z množiny všech objektů jako jediného shluku a jeho postupným dělením získáme systém shluků, až skončíme ve stadiu jednotlivých objektů. Výhodou hierarchických metod je nepotřebnost informace o optimálním počtu shluků v procesu shlukování; tento počet se určuje až dodatečně. Při shlukování vznikají dva základní problémy:

a) způsob měření vzdáleností mezi objekty. I když existuje celá řada měr vzdáleností (vícerozměrných metrik), nejčastěji se užívá eukleidovská metrika, která je přirozeným zobecněním běžného pojmu vzdálenosti;
b) volba vhodné shlukovací procedury dle zvoleného způsobu metriky. [1]

Metody shlukování podle typu metriky

a) Metoda průměrová (v programech je označena heslem Average): vzdálenost dvou shluků se počítá jako průměr z možných mezishlukových vzdáleností dvou objektů, kdy se mezishlukovou vzdáleností objektů rozumí vzdálenost dvou objektů, z nichž každý patří do jiného shluku. Nejbližší jsou shluky, které mají nejmenší průměrnou vzdálenost mezi všemi objekty jednoho a všemi objekty druhého shluku. Dendrogramy mají strukturu podobnou dendrogramům metody nejvzdálenějšího souseda, pouze spojení je provedeno při obvykle vyšších vzdálenostech.
b) Metoda centroidní (Centroid): vzdálenost shluků se počítá jako eukleidovská vzdálenost jejich těžišť. Nejbližší jsou ty shluky, které mají nejmenší vzdálenost mezi těžišti.
c) Metoda nejbližšího souseda (Single, Nearest Neighbour): kritériem pro vytváření shluků je minimum z možných mezishlukových vzdáleností objektů. Metoda tvoří nový shluk na základě nejkratší vzdálenosti mezi shluky (či objekty) a neumí proto rozlišit špatně separované shluky. Na druhé straně je to jedna z mála metod, která umí roztřídit a rozlišit i neeliptické shluky.
d) Metoda nejvzdálenějšího souseda (Complete, Furthest Neighbour): počítá vzdálenost dvou shluků jako maximum z možných mezishluko-

vých vzdáleností objektů. Probíhá podobně jako metoda Single s jednou důležitou výjimkou, že vzdálenost (či nepodobnost) mezi shluky je určována vzdáleností (či nepodobností) mezi dvěma nejvzdálenějšími objekty, každý přitom je z jiného shluku. Proto všechny objekty ve shluku jsou klasifikovány na základě maximální vzdálenosti či minimální podobnosti vůči objektům ve druhém shluku.

e) Metoda mediánová (Median): jde o jisté vylepšení centroidní metody, neboť se snaží odstranit rozdílné „váhy“, které centroidní metoda dává různě velkým shlukům.

f) Wardova metoda: je založena na minimalizaci ztráty informace při spojení dvou tříd. [2], [3]

Nehierarchické shlukovací metody

U metody typických bodů (v programech označené heslem Seeded) uživatel na základě svých věcných znalostí určí, které objekty mají být „typickými“ představiteli nově vytvořených shluků a systém rozdělí objekty do shluků podle jejich euklidovské vzdálenosti od těchto typických objektů. V nehierarchických shlukovacích metodách je počet shluků obvykle předem dán, i když se v průběhu výpočtu může změnit. Zůstává-li počet shluků zachován, hovoříme o nehierarchických metodách s konstantním počtem shluků, v opačném případě o nehierarchických metodách s optimalizovaným počtem shluků.

Nehierarchické metody zahrnují dvě základní varianty – optimalizační metody a analýzu módů, medoidů. Optimalizační nehierarchické metody hledají optimální rozklad přeřazováním objektů ze shluku do shluku s cílem minimalizovat nebo maximalizovat nějakou charakteristiku rozkladu. Metody, označované jako analýza módů (medoidů), představují hledání rozkladu do shluků, kde shluky jsou chápány jako místa se zvýšenou koncentrací objektů v m-rozměrném prostoru proměnných.

Klíčovým problémem všech nehierarchických procedur zůstává volba shlukových zárodků. Při volbě sekvenčního prahu například závisí počáteční a konečný shluk na pořadí objektů v datové matici. Proto se provádí náhodné přeuspořádání objektů. Určením počátečních shlukových zárodků, jako je tomu v sekvenčním prahovém postupu, lze tento problém redukovat. I když se vyberou zárodky shluků náhodně, bude každý zárodek poskytovat jiné výsledky. Uživatel proto musí být velmi opatrný při zadávání shlukových zárodků, protože jimi může hodně ovlivnit konečné výsledky. [4]

2. Praktická část

Popis algoritmu v Matlabu

Pro hierarchické shlukování použijeme v Matlabu *Statistics Toolbox Functions,* budeme postupovat podle následujících kroků:

1) Nalezni podobnost nebo nepodobnost mezi každým párem objektů v souboru dat.
2) Seskup objekty do binárního hierarchického shlukovacího stromu.
3) Urči, kde je potřeba rozdělit hierarchický strom do shluků. [5]

Míry podobnosti

Funkce *pdist* vypočte vzdálenost mezi každým párem objektů v souboru dat. Pro každých m objektů lze sestavit $m(m-1)/2$ párů v souboru dat. Výsledky těchto výpočtů shrnujeme do matice vzdáleností resp. matice nepodobností. Existuje mnoho způsobů, jak tuto vzdálenost vypočítat. Funkce *pdist* je v Matlabu nastavena na eukleidovskou vzdálenost mezi objekty. Jdou však manuálně nastavit i jiné typy výpočtů vzdáleností.

Praktický příklad na shlukovou analýzu

Uvažujme například soubor dat X, který je tvořen pěti objekty, z nichž každý má následující souřadnice x a y.

```
Objekt 1: 1; 2
Objekt 2: 2,5; 4,5
Objekt 3: 2; 2
Objekt 4: 4; 1,5
Objekt 5: 4; 2,5
```

Vzdálenosti

Funkce *pdist* vrátí tyto vzdálenosti pomocí vektoru Y, kde každá jeho složka obsahuje vzdálenost mezi párem objektů (eukleidovskou).

```
Y = pdist(X)
Y =
  Columns 1 through 5
    2.9155    1.0000    3.0414    2.5495
  Columns 6 through 10
    3.3541    2.5000    2.0616    1.0000
```

Pro lepší orientaci ve výsledcích je lepší převést tento vektor Y na matici pomocí další funkce Matlabu *squareform*. V této matici pak každý prvek i, j odpovídá vzdálenosti mezi jednotlivými objekty i a j. V následujícím příkladu prvek matice 1,1 reprezentuje vzdálenost objektu 1 a objektu 1 (ta je pochopitelně nulová). Prvek 1,2 reprezentuje dále vzdálenost objektu 1 a objektu 2 (2,9155) atd. Matice je symetrická kolem hlavní diagonály, na které jsou nuly.

```
squareform(Y)
ans =
         0    2.9155    1.0000    3.0414    3.0414
    2.9155         0    2.5495    3.3541    2.5000
    1.0000    2.5495         0    2.0616    2.0616
    3.0414    3.3541    2.0616         0    1.0000
    3.0414    2.5000    2.0616    1.0000         0
```

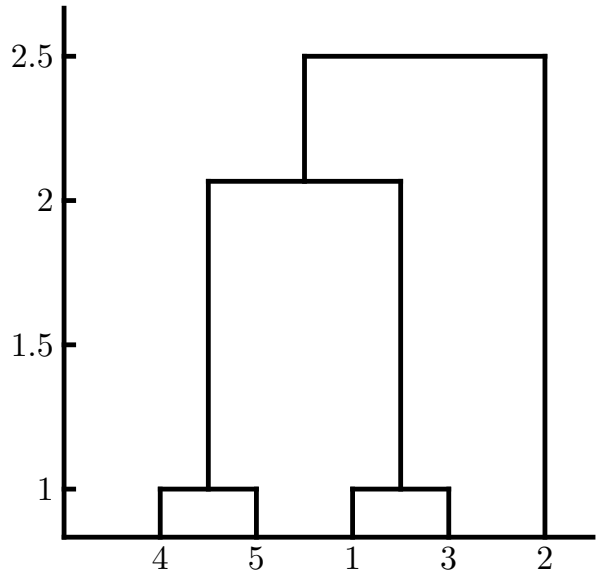
Propojení

Jakmile je vzdálenost mezi objekty vypočítána, můžeme dále určit, jak by měly být objekty v souboru dat rozděleny do shluků. To provedeme pomocí funkce *linkage*. Tato funkce bere vzdálenost vypočtenou pomocí funkce *pdist* a spojí páry blízkých objektů do binárních shluků (tj. shluků, které jsou tvořeny dvěma objekty). Funkce *linkage* potom spojí tyto nově vytvořené shluky navzájem a také s dalšími objekty, takže se vytvoří větší shluky. To se děje až do té doby, než jsou spojeny všechny objekty datového souboru do hierarchického stromu. Například pro daný vektor Y z předchozího odstavce (byl vypočten pomocí funkce *pdist* z daných dat o souřadnicích x a y), funkce *linkage* vygeneruje hierarchický strom, který je vyjádřen pomocí matice Z.

```
Z = linkage(Y)
Z =
    4.0000    5.0000    1.0000
    1.0000    3.0000    1.0000
    6.0000    7.0000    2.0616
    2.0000    8.0000    2.5000
```

Hierarchický binární strom vytvořený pomocí funkce *linkage* je lépe znázornit graficky. *Statistics Toolbox* v Matlabu obsahuje funkci dendrogram, která vykreslí tento strom do grafu na obrázku 1.

Na obrázku 1 jsou na ose x čísla původních objektů souboru dat. Spoje mezi objekty mají tvar převráceného písmene U. Výška tohoto písmene znamená vzdálenost mezi těmito jednotlivými objekty. Například spoj zahrnující



Obrázek 1: Hierarchický strom získaný `dendrogram(Z)`.

objekt 1 a 3 má výšku 1. Spoj reprezentující shluk objektu 2 spolu s objekty 1, 3, 4 a 5 (objekt 8) má výšku 2,5. Tato výška je vzdáleností mezi objekty 2 a 8.

Tvorba shluků

Po vytvoření hierarchického stromu můžeme tento strom prořezat pomocí shlukovací funkce. To slouží k rozdělení dat do jednotlivých shluků. Díky shlukovací funkci můžeme tvořit shluky dvěma následujícími postupy:

1. *nalézt přirozené rozdělení v datech;*
2. *specifikovat počet shluků.*

Například jestliže použijeme funkci cluster pro shlukování našich dat a parametr *cutoff* nastavíme na hodnotu 1,2, tato funkce rozdělí objekty pouze do jednoho shluku.

```
T = cluster(Z,'cutoff',1.2)
T' = 1 1 1 1 1
```

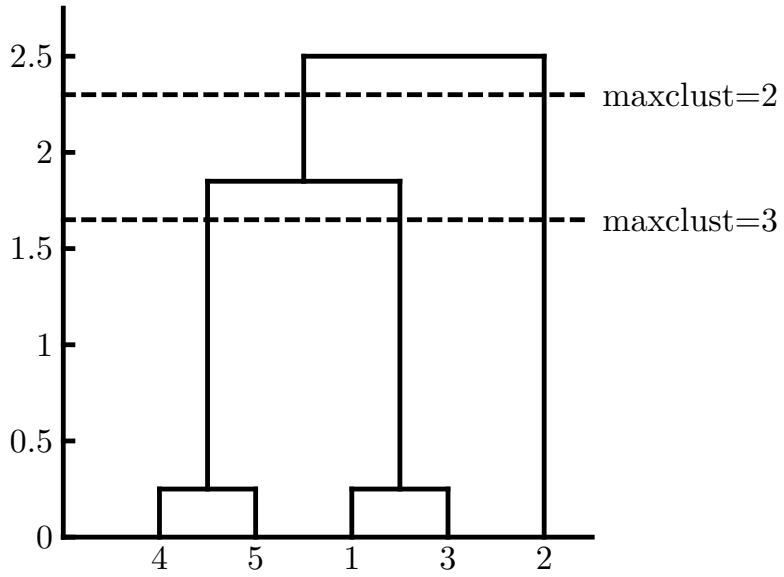
Funkce *cluster* vypočte vektor T, který má stejnou délku jako je původní soubor dat (v našem příkladu 5). Každý prvek vektoru znamená číslo shluku, do kterého příslušný objekt souboru dat patří. V tomto případě bude vhodné nastavit hodnotu *cutoff* menší než 1,2. Jestliže snížíme prahovou hodnotu koeficientu nekonzistence (cutoff) na 0,8, funkce *cluster* rozdělí dané objekty do tří oddělených shluků.

```
T = cluster(Z,'cutoff',0.8)
T' = 1 3 1 2 2
```

Tento výstup znamená, že objekty 1 a 3 byly umístěny do shluku 1, objekty 4 a 5 do shluku 2 a konečně objekt 2 do shluku 3.

Vizualizace shluků

Na obrázku 2 vidíme vizualizaci shlukování pomocí dendrogramu. Zde budeme ručně zadávat počet shluků pomocí funkce *maxclust*. Horizontální přerušovaná čára kříží dvě čáry dendrogramu při nastavení funkce *maxclust* na hodnotu 2. Tyto dvě čáry rozdělují objekty do dvou shluků: objekty pod levou čarou jmenovitě 1, 3, 4 a 5 patří do jednoho shluku, zatímco objekt pod pravou čarou (jmenovitě objekt 2) patří do druhého shluku (viz obrázek 2).



Obrázek 2: Dendrogram `T=cluster(Z,'maxclust',2)`, resp. 3.

Na druhé straně, jestliže nastavíme *maxclust* na hodnotu 3, shlukovací funkce seskupí objekty 4 a 5 do jednoho shluku, objekty 1 a 3 do druhého a objekt 2 do třetího shluku. Tentokrát je ale shlukovací funkce na nižší úrovni a přetne tedy tři spoje dendrogramu, jak můžeme vidět na obrázku 2.

3. Závěr

Tento příspěvek se zabýval v prvé části shlukovou analýzou jak po teoretické a metodické stránce. Byly zde zmíněny hierarchické i nehierarchické metody shlukování. Následující druhá část příspěvku obsahovala praktickou aplikaci jednoduchého příkladu za použití softwaru Matlab 2007b.

Poděkování

Tento příspěvek vznikl za podpory Interní grantové agentury UTB, projekt č. IGA/73/FaME/10/D, pod názvem Rozvoj využívání matematicko-statistických metod v řízení kvality.

Literatura

[1] Hogg R. V., Ledolter J. *Engineering Statistics.* MacMillan, 1987. 1. vyd. 442 s. ISBN 978-0023557903.

[2] Meloun M., Militký J. *Kompendium statistického zpracování dat.* 2. vyd. Praha: Academia, nakladatelství Akademie věd České republiky, 2006. 982 s. ISBN 80-200-1396-2.

[3] Meloun M., Militký J., Hill M. *Počítačová analýza vícerozměrných dat v příkladech.* 1. vyd. Praha: Academia, nakladatelství věd České republiky, 2005. 450 s. ISBN 80-200-1335-0.

[4] Meloun M., Militký J. *Statistická analýza experimentálních dat.* 2. vyd. Praha: Academia, nakladatelství Akademie věd České republiky, 2004. 953 s. ISBN 80-200-1254-0.

[5] Matlab 2007b – Help dokumentace programu.

MIXING AND SUBSTITUTING R, C, AND FORTRAN
R + C + FORTRAN

David Kraus

Adresa: ÚTIA AV ČR, Pod Vodárenskou věží 4, 182 08 Praha 8

E-mail: david.kraus@matfyz.cz

Abstract: This contribution shows how the R statistical software environment can be linked with computational routines written in R or Fortran in order to speed up the computation.

Keywords: R, C, Fortran, compilation, computing.

Abstrakt: V tomto příspěvku je ukázáno, jak v zájmu zrychlení výpočtu propojit statistické softwarové prostředí R s výpočetními rutinami napsanými v C nebo Fortranu.

Klíčová slova: R, C, Fortran, kompilace, výpočet.

Poznámka

Autor článku i redakce Informačního Bulletinu jsou si vědomi, že články tohoto typu ve světě informatiky a statistického výpočetního prostředí jazyka R obzvlášť rychle zastarávají, ale jde o důležitou ideu, na kterou cítíme potřebu naše čtenáře upozornit. Text byl napsán v roce 2006 a v současné době již technické detaily pravděpodobně neplatí, což ale autor nemůže ověřit, protože v současnosti nedisponuje ani jedním z operačních systémů v článku zmíněných (Microsoft Windows, GNU Linux).

1. Úvod

Programy v R bývají často neúnosně pomalé. Přepsáním výpočetně nejnáročnějších částí programu z R do C, Fortranu 77 a případně Fortranu 90 lze dosáhnout mnohonásobného zrychlení výpočtu. Tyto jazyky můžeme kombinovat a vše dohromady zkompilovat do jedné dynamické knihovny.

V příspěvku je nejprve popsáno, jak provádět kompilaci: jak překládat C a Fortran dohromady, jak k tomu přidat Fortran 90, jak to vše zprovoznit ve Windows a v Linuxu. Dále následuje několik poznámek o tom, na co si dát pozor, jak používat numerické knihovny (BLAS, LAPACK), jak volt vnitřní funkce R z Fortranu nebo C a jak volat funkce napsané v C z Fortranu a naopak.

Doručeno redakci: 5. 11. 2006, imprimatur: 8. 9. 2010.
MSC2010: 68N20, DOI: 10.5300/IB/2011-1/28

2. Motivační příklad: jak to vypadá, když to funguje

Představme si, že jsme si v C napsali funkci `soucin` na vynásobení dvou čísel a umístili ji do souboru `prog1.c`. Dále máme v souboru `prog2.f` ve Fortranu 77 napsanou subroutinu `soucet` sčítající dvě čísla. Soubory vypadají takto:

`prog1.c`

```
void soucin(double *x, double *y, double *z)
{
    *z=(*x)*(*y);
}
```

`prog2.f`

```
subroutine soucet(x,y,z)
double precision :: x,y,z
z=x+y
end subroutine soucet
```

Cílem je přeložit tyto dva soubory do dynamické knihovny (pod Linuxem soubor s příponu `.so`, pod Windows `.dll`) tak, abychom si ji mohli načíst do R a spustit v ní obsažené funkce `soucin` a `soucet`.

Při překladu C a/nebo Fortranu 77 v Linuxu se s problémem pravděpodobně nesetkáme. Ihned po instalaci R můžeme zadat příkaz

```
R CMD SHLIB prog1.c prog2.f -o knihovna.so
```

a dočkáme se kýženého výsledku.

Ve Windows nám toto (s příponou `.dll`) hned fungovat nebude, ale nic není ztraceno. Jen je potřeba nejprve hodně věcí nainstalovat. Návod čtenář najde v další sekci.

Hotovou knihovnu do R natáhneme pod Linuxem příkazem

```
dyn.load("knihovna.so")
```

Pod Windows zaměníme `.so` za `.dll`. Abychom měli kód nezávislý na platformě, použijeme raději

```
dyn.load(paste("knihovna",.Platform$dynlib.ext,sep=""))
```

Funkce v knihovnách se volají příkazy `.C` a `.Fortran`, které vracejí list s položkami odpovídajími argumentům volané funkce. Použití je následující:

```
a=2
b=3
vystup.c=.C("soucin",x=as.double(a),y=as.double(b),
             z=double(1),PACKAGE="knihovna")
vystup.c$z
vystup.f=.Fortran("soucet",x=as.double(a),y=as.double(b),
                  z=double(1),PACKAGE="knihovna")
vystup.f$z
```

Nataženou knihovnu uvolníme příkazem `dyn.unload` (volá se se stejným argumentem jako `dyn.load`). Pod Windows je knihovnu před novou kompilací nezbytné uvolnit z R, jinak na soubor `.dll` nelze zapisovat. V Linuxu toto není třeba. Provedené změny v programech se projeví po opětovném načtení pomocí `dyn.load`.

3. Zprovoznění `R CMD` ve Windows

Než si vyložíme, jak pod Windows uvést nástroj `R CMD` v život, poznamenejme, že překlad programů můžeme provádět 'ručně' bez použití tohoto nástroje. Avšak při použití `R CMD` bude vzniklá knihovna bez naší námahy nalinkována proti knihovnám R (knihovny vnitřních funkcí R, numerické knihovny, ...) a můžeme tedy v našich programech používat funkce z těchto knihoven.

Následující postup jsem úspěšně použil s R 2.2.0 pod MS Windows 2000 Professional SP2 a pod MS Windows 98. Postup nám kromě pohodlného používání `R CMD` také umožní zkompilovat R ze zdrojových kódů v (dle mých zkušeností zbytečné) naději, že bude rychlejší. Podrobnosti postupu (včetně vynechaných kroků) se najdou v Appendixu F manuálu R-admin (R Installation and Administration) a na `http://www.murdoch-sutherland.com/Rtools/`. Postup instalace také přehledně popisují Ligges & Murdoch [1].

1. **Perl.** Stáhněte ActivePerl z

   `http://www.activestate.com/`

   (soubor `.msi`). K instalaci je potřeba MSI (Microsoft Installer). Pokud jej nemáte (starší verse Windows), najdete ho rovněž na stránkách ActiveState. Pokud se vám MSI nepodaří nainstalovat (velmi staré verse Windows), použijte ActivePerl v souboru `.zip`. Perl instalujte do `c:\perl`.

2. **Unixové nástroje.** Je potřeba obohatit Windows o základní unixové příkazy (`ls`, `pwd`, `make`, `sed`, ...). Na adrese

   `http://www.murdoch-sutherland.com/Rtools/`

   najdete sadu těchto nástrojů a další poučné čtení. Rozbalte `tools.zip` do adresáře, jehož cesta neobsahuje mezery, například `c:\apps` (čili výsledkem bude adresář `c:\apps\tools\bin` s několika desítkami souborů).

3. **Kompilátor.** Nainstalujte kompilátor MinGW (neboli Minimalist GNU for Windows, `http://www.mingw.org/`), což je port kompilátoru GCC pro Windows. Nejjednodušší je stáhnout z adresy

`http://prdownloads.sf.net/mingw/MinGW-5.0.3.exe`

instalátor, který po spuštění stáhne a nainstaluje jednotlivé součásti prostředí MinGW (v instalátoru zvolte verzi 'Candidate' a kromě základního balíčku vyberte ještě g77 a g++). Instalujte například do `c:\apps\MinGW`.

Alternativně můžete ze SourceForge jednotlivé součásti MinGW (soubory `.tar.gz`) stáhnout 'ručně' a rozbalit je do `c:\apps\MinGW`. Soubory jsou vyjmenovány v R-admin v Appendixu F.3, odkazy na ně (či jejich novější verse) jsou na

`http://www.mingw.org/download.shtml#hdr2`

v části 'Candidate'.

Místo GCC byste (alespoň teoreticky) mohli použít jiný kompilátor (například od výrobce operačního systému nebo procesoru). Zřejmě by ale bylo nutno vynaložit podstatně vyšší úsilí na správnou konfiguraci jak R (cesty ke kompilátoru a linkeru) tak kompilátoru (floating-point aritmetika kompatibilní s R, exportování funkcí do knihoven pod správnými názvy atd.); Appendix C v R-admin pojednává o použití jiných kompilátorů než GCC (pod mnoha systémy, nikoli Windows). Výchozí nastavení R je připraveno na použití s GCC.

4. **Nastavení systémových proměnných.** Systémové proměnné se nastavují naklikáním v 'Ovládacích panelech' pod ikonou 'Systém' (musíte mít administrátorská práva). (Ve Windows 9x/Me se proměnné nastavují v souboru `autoexec.bat`.) Takto provedené změny se projeví až po restartu. Chcete-li si správné nastavení vyzkoušet bez nutnosti restartu, můžete navíc proměnné nastavit v konsoli pomocí příkazu `set`. Toto nastavení ale platí jen v konsoli, v níž bylo provedeno; po jejím uzavření se ztratí.

   Vytvořte proměnnou `LIBRARY_PATH` a nastavte její hodnotu na:
   `c:\apps\MinGW\lib`

   Proměnnou `R_HOME` nastavte na adresář, v němž je nainstalováno R. Předpokládejme, že R se nachází v adresáři `c:\Program␣Files\R\R-2.2.0`. Pak proměnnou `R_HOME` asi nastavíte na `C:\PROGRA~1\R\R-22~1.0`. Toto je bezmezerový formát cesty. Cesty se musejí uvádět v této podobě, která se zjistí v konsoli příkazem `dir /x`.

   Proměnnou `PATH` je nutno nastavit tak, aby začínala

   ```
   .;c:\apps\tools\bin;c:\perl\bin;
     c:\apps\MinGW\bin;%R_HOME%\bin;
   ```

Pak následuje to, co tam už bylo. Důležité je, aby `c:\apps\tools\bin` byl v proměnné co nejdříve. Musí tam být dříve než adresáře, které by mohly obsahovat soubory stejných jmen jako unixové programy. Takovým nebezpečným adresářem, který se nesmí v `PATH` vyskytovat dříve než `c:\apps\tools\bin`, může být adresář CygWinu.

V konsoli tyto proměnné můžete nastavit příkazy

```
set LIBRARY_PATH=c:\apps\MinGW\lib
set R_HOME=C:\PROGRA~1\R\R-22~1.0
set PATH=.;c:\apps\tools\bin;c:\perl\bin;
          c:\apps\MinGW\bin;%R_HOME%\bin;%PATH%
```

4. Fortran 90 pod Windows a Linuxem

V této části popíšu, jak kromě Fortranu 77 a C umožnit i kompilaci Fortranu 90. Výhoda Fortranu 90 spočívá v pohodlné práci s vektory, maticemi a poli. S těmito strukturami lze ve Fortranu 90 zacházet podobně jako v R nebo MATLABu.

4.1. Linux

Zatímco kompilace C a Fortranu 77 pomocí `R CMD` funguje pod Linuxem okamžitě po instalaci R, nastavení pro práci s Fortranem 90 už vyžaduje určité úsilí. Zde uvedený postup jsem úspěšně použil v operačním systému SUSE Linux 10.0 OSS s GCC 4.0.2 a R 2.2.1.

Uvedená verse GCC obsahuje kompilátor gfortran, který umí překládat kromě Fortranu 77 i Fortran 90 a 95. Problém tedy je na straně R, které neumí se soubory `.f90` zacházet. Zde je návod, jak ho to naučit.

1. V souboru `/usr/lib/R/bin/SHLIB` je mimo jiné napsáno, co se má dělat se vstupními soubory v závislosti na příponě. Chybějí zde instrukce pro soubory `.f90`, takže tyto jsou skriptem ignorovány. Přidáme je tak, že nahradíme řetězec `*.f)` řetězcem `*.f|*.f90)`. Ve versi R 2.2.1 se tato změna odehraje na řádku 55.

   Podobně můžeme upravit soubor `/usr/lib/R/bin/COMPILE`, který prostřednictvím příkazu `R CMD COMPILE` slouží ke kompilaci bez vytváření dynamické knihovny. V tomto souboru na řádku 48 změníme:
   `*.[cfC]|*.cc|*.cpp)` na `*.[cfC]|*.cc|*.cpp|*.f90)`.

2. Musíme upravit soubor `/usr/lib/R/etc/Makeconf`, což je konfigurační soubor pro make, který říká co se jak a čím kompiluje nebo linkuje. (Soubor

obsahuje tabelátory, proto musíme být opatrní, abychom nepoužili editor, který tabelátory změní na mezery.) V souboru nejprve vytvoříme proměnné určující kompilátor a přepínače pro Fortran 90. Vzhledem k tomu, že používáme stejný kompilátor (jako pro Fortran 77), nastavíme proměnné na stejné hodnoty. Dále make seznámíme s příponou `.f90` (přidáním této přípony na řádek začínající `.SUFFIXES`) a na konci tohoto souboru určíme pravidla pro práci se soubory `.f90`, a to zcela analogicky tomu, co už tam je pro Fortran 77.

Upravené verse `/usr/lib/R/bin/{SHLIB,COMPILE}` a `/usr/lib/R/etc/Makeconf` jsou k disposici na stránkách společnosti.

4.2. Windows

Ve Windows musíme kromě konfigurace R nejprve vyřešit otázku kompilátoru. Vzhledem k tomu, že MinGW portuje GCC 3.4, které obsahuje jen kompilátor Fortranu 77 (g77), musíme si nějaký kompilátor Fortranu 90 pořídit jinak. Nabízejí se dva kompilátory: g95 (`http://www.g95.org/`) a gfortran (`http://gcc.gnu.org/fortran/`). Oba dva si s GCC rozumějí, jejich instalace jsou připravené na použití s MinGW. Vyzkoušený mám g95.

1. Ze serveru `http://www.g95.org/` stáhneme instalaci g95 (soubor `.exe` pro MinGW, nikoli `.tgz` pro Cygwin). Nainstalujeme do `c:\apps\MinGW`, čímž dojde k 'přimíchání' g95 do MinGW.

2. Majíce na paměti poznámku o tabelátorech z bodu 2 v předchozí části, upravíme soubor `%R_HOME%\src\gnuwin32\MkRules`, který má podobnou úlohu jako v Linuxu `Makeconf`.

   Proměnnou `F90` nastavíme na `g95 -mno-cygwin`. Přidáme pravidlo pro soubory `.f90`, dále upravíme pravidlo pro `.f` tak, aby se i Fortran 77 kompiloval pomocí g95, a jako linker nastavíme g95. (Pokud ponecháme původní nastavení, v němž se Fortran 77 kompiluje pomocí g77, nebudeme schopni výstupy z g77 a g95 linkovat.) Upravený soubor `MkRules` je na stránkách společnosti.

Nevyzkoušenou alternativou ke g95 by mohla být kompilace GCC 4 ze zdrojových kódů pomocí nainstalovaného GCC 3.4 (MinGW).

5. Poznámky k používání

Tato sekce obsahuje pár upozornění na pasti, do nichž by se mohl uživatel chytit. Dále stručně zmiňuje, co všechno se dá s R, C a Fortranem dělat.

Podrobnosti zájemce nalezne v manuálu R-exts (Writing R Extensions). Z `http://www.davidkraus.net/past/` si lze stáhnout několik příkladů.

5.1. Jména souborů

Jména vstupních souborů se musejí lišit nejen příponou. Nelze použít něco jako

```
R CMD SHLIB program.c program.f -o knihovna.so
```

V tomto případě se totiž `program.c` zkompiluje na soubor `program.o`, jehož obsah je ovšem vzápětí přepsán výstupem kompilace souboru `program.f`.

5.2. Typy proměnných

Jak už bylo vidět v příkladu v sekci 2, je potřeba zajistit, aby byly v souladu datové typy proměnných v R a jim odpovídajících proměnných v C nebo Fortranu. Pokud toto není v pořádku, program v lepším případě spadne nebo vrátí očividně nesmyslný výsledek, v horším případě vrátí špatný výsledek, na němž to ale nepoznáme. Dobré tedy je před předáním dynamické knihovně proměnné správně explicitně přetypovat (pomocí funkcí `as.double`, `as.integer`, ... nebo `storage.mode`). Spoléhat se na intuici je zrádné. Posuďte sami:

```
n=5
storage.mode(n)    # "double"
v=1:n
storage.mode(v)    # "integer"
v[1]=1
storage.mode(v)    # "double"
m=length(v)
storage.mode(m)    # "integer"
n==m               # TRUE
a=integer(1)
storage.mode(a)    # "integer"
a=2
storage.mode(a)    # "double"
```

5.3. Předávání proměnných, práce s poli

Při předání matice z R do Fortranu se nic pozoruhodného neděje. Pole ve Fortranu najdeme v takové podobě, v jaké bylo v R (stejný tvar, rozměry,

uspořádání prvků). Jen je třeba při předání pole předat i jeho rozměry (ve Fortanu 77 je to samozřejmé; ve Fortranu 90 to znamená, že nemůžeme při přechodu z R do Fortranu používat pole předpokládaného tvaru, byť uvnitř fortranského programu je používat můžeme).

V céčkovských funkcích, které mají být volány z R, se argumenty (pole i skaláry) deklarují jako pointery na typ (čili u každé proměnné musí být právě jedna hvězdička). Je-li z R předána matice, v C se z ní stane vektor (pointer na jeho začátek), který je vyplněn prvky matice, a to *po sloupcích* (což je podobné Fortranu, ale odlišné od C, které s maticemi pracuje po řádcích).

Například máme-li v R matici reálných čísel o rozměrech `n` × `p`, v C musí být v hlavičce funkce odpovídající proměnná typu pointer na double (součástí hlavičky tedy bude např. `double *x, int *n, int *p`), nikoli pointer na pointer na double (tj. `double **x, int *n, int *p`), jak by asi bylo v C obvyklejší. K prvku na místě (`i`,`j`) přistoupíme pomocí `x[i+*n*j]` (popřípadě `*(x+i+*n*j)`).

Samozřejmě si pak můžeme vytvořit přívětivější 'maticovou' representaci typu pointer na pointer na double (`double **x_mat`), abychom mohli psát `x_mat[i][j]`. Vyžaduje to ovšem předávat z R matice transponované, aby řádky tvořily v paměti souvislé úseky. Takto se s maticemi pracuje například v balíčku `survival`, výrobu maticové representace tam zajišťuje funkce `dmatrix`.

Dynamická alokace v C se nejpohodlněji provede pomocí funkce `R_alloc`. O dealokaci takto alokované paměti se nemusíme starat. Funkce je deklarována v hlavičkovém souboru `R_ext/Memory.h` (jenž se automaticky natáhne při natažení `R.h`). Více se nalezne v kapitole 5.1 v R-exts.

Výstupy z přeložených knihoven se do R vracejí v argumentech podprogramů. Jinými slovy podprogram, který chceme volat z R, musí ve Fortanu být subroutinou (nikoli funkcí), a v C musí být funkcí typu `void`.

5.4. Volání C z Fortranu a naopak

Máme-li část programu v C a část ve Fortranu, může se stát, že bychom chtěli v C zavolat nějakou fortranskou subroutinu, nebo naopak ve Fortranu nějakou funkci napsanou v C. Slouží k tomu funkce `F77_NAME`, `F77_CALL` a `F77_SUB`. Jejich užití je podrobně vysvětleno v sekci 5.6 v R-exts.

Všechny se používají v programech v C. Abychom je mohli použít, musíme do programu v C přidat

```
#include <R.h>
```

Zde vidíme užitečnost použití `R CMD`. Bez něj bychom museli kompilátoru říci, kde soubor `R.h` najde. Navzdory svým názvům tyto funkce fungují i s Fortranem 90.

5.5. Numerické knihovny BLAS a LAPACK

R tyto knihovny obsahuje a můžeme je tedy ve svých programech směle používat. Jediné co je třeba udělat, je zajistit, abychom tyto knihovny měli nalinkovány. K tomu stačí vytvořit si v kompilačním adresáři soubor `Makevars` a napsat do něj

```
PKG_LIBS = $(LAPACK_LIBS) $(BLAS_LIBS) $(FLIBS)
```

(V Linuxu můžeme `Makevars` také umístit do `~/.R/`.) Numerické knihovny BLAS a LAPACK jsou napsány ve Fortranu, proto je můžeme ze svých fortranských subroutin přímo volat. Volání z céčkovských funkcí se provádí prostřednictvím `F77_NAME` a `F77_CALL` zmíněných v předešlé části. Deklarování pomocí `F77_NAME` si ušetříme, natáhneme-li si příslušný hlavičkový soubor:

```
#include <R_ext/Lapack.h>
```

Pokud nechceme použít LAPACK dodávaný s R (nebo je to komplikované při nefunkčním `R CMD`), můžeme si z Netlibu (`http://www.netlib.org/`) stáhnout příslušnou subroutinu (Netlib ji nachystá včetně všech závislostí) a přidat ji do kompilace.

5.6. Vnitřní funkce R

Chceme-li používat céčkovské funkce, pomocí nichž počítají funkce v R, natáhneme hlavičkový soubor `Rmath.h` (`#include <Rmath.h>`). Pak máme k disposici to, na co jsme zvyklí z prostředí R: generátory náhodných čísel, distribuční funkce, hustoty a kvantily, dále matematické funkce (gamma, ...), sortování a další. Dále jsou zde zejména při ladění užitečné funkce a subroutiny zajišťující tisk do konsole R a nástroje umožňující přerušit z prostředí R výpočet probíhající v námi vytvořené knihovně. Další užitečné funkce jsou v `R_ext/Applic.h`. O tom všem pojednává kapitola 5 v R-exts.

Literatura

[1] Uwe Ligges, Duncan Murdoch. R Help Desk. Časopis *R News,* vol. 5, no. 2, 2005, s. 27–28. ISSN 1609-3631. Dostupné z URL: `http://cran.r-project.org/doc/Rnews/Rnews_2005-2.pdf`

CZECHOSLOVAK STATISTICAL SOCIETY BEFORE THE SECOND WORLD WAR

ČESKOSLOVENSKÁ STATISTICKÁ SPOLEČNOST PŘED 2. SVĚTOVOU VÁLKOU

Jaroslav Češka

Adresa: Pavel Stříž (redakce), ÚSKM FaME UTB, Mostní 5139, 760 01 Zlín

E-mail: striz@fame.utb.cz

Abstract: Czechoslovak Statistical Society was founded in January 1929. Its members were both experts in the field of statistics and specialists who applied statistics in their field of interest. This contribution summarizes several most important activities held under the auspices of the society before the Second World War. It contains also a reflection about the role of statistics in contemporary society.

Keywords: Statistics, History, Czechoslovak Statistical Society.

Abstrakt: Československá statistická společnost byla založena v lednu 1929. Mezi její členy patřila řada nejenom statistických expertů, ale i odborníků, kteří statistiku aplikovali ve své odborné disciplíně. Příspěvek shrnuje některé nejdůležitější aktivity společnosti, jež se uskutečníly před druhou světovou válkou. Dále se zamýšlí nad rolí statistiky v současné společnosti.

Klíčová slova: statistika, historie, Český statistický úřad.

V souvislosti s dvacetiletým výročím založení České statistické společnosti v tomto roce 2010 je vhodné si připomenout i činnost Československé statistické společnosti /ČSS/ v předválečném období, v období první Československé republiky v letech 1918–1938.

Při popisu činnosti ČSS není bez zajímavosti uvést, jak došlo k jejímu založení a co bylo na programu její první schůze /prvním valném shromáždění/. Je nutno dodat, že založení ČSS spadá do období, kdy k zakládání statistických společností dochází v řadě dalších evropských zemí.

Z podnětu presidenta Státního úřadu statistického prof. Dr. F. Weyra /1879–1951/ byl ustaven přípravný výbor, který 28. listopadu 1928 projednal návrh stanov statistické společnosti, které byly vzaty na vědomí Zemským úřadem v Praze.

Dne 30. ledna 1929 se uskutečnilo první ustavující valné shromáždění zakládajícího výboru, které podle stanov zvolilo předsednictvo společnosti,

Doručeno redakci: 6. 7. 2010, imprimatur: 7. 12. 2010.
MSC2010: 00A99, DOI: 10.5300/IB/2011-1/37

30 řádných a 22 dopisujících členů. Předsedou byl zvolen prof. Dr. V. Mildschuh, profesor Karlovy univerzity a uznávaný statistický odborník. Na schůzi předsednictva dne 22. března 1929 byl projednán program společnosti pro počáteční období.

První valné shromáždění ČSS se sešlo 26. 4. 1929. Na schůzi byl zdůrazněn význam založení statistické společnosti i její předpokládaný přínos. Založení společnosti, bylo zdůrazněno, je akt důvěry a vychází z předpokladu, že je dostatek povolaných pracovníků, kteří by mohli sloužit její veliké myšlence /idei/ – tj. úsilí o vysokou vědeckou úroveň statistického bádání. Při hodnocení stále většího významu statistiky pro hospodářskou praxi bylo poukázáno, že „její služba platí snaze poznat co nejdůkladněji vše, co nás obklopuje, poznat vše konkrétně, pravdivě a vědecky bezpečně“. Zvláštní důraz byl také položen na otázky metodické „ježto /podle jednatele a místopředsedy dr. B. Živanského/ se v tomto směru vyskytuje u nás dosud mnoho poklesů při užívání statistických výsledků nezřídka i ve vládních návrzích zákonů“.

Po zprávě jednatele byl schválen navržený jednací řád a zvoleni noví tři členové společnosti. Po jejich zvolení se uskutečnila přednáška místopředsedy společnosti dr. Boháče „Náš populační program a statistika“, na kterou navázala obsáhlá diskuse na dalších schůzkách společnosti. Programem pozdějších schůzek společnosti byly další přednášky, a to dr. G. Reifa „O metodě a programu statistiky mzdové u nás“ a dr. B. Živanského „Pokud naše úřední statistika může přispět ke zjištění příčin dnešní zemědělské tísně“.

Předsednictvo společnosti se v období do druhého valného shromáždění sešlo čtyřikrát. Na svých schůzích se zabývalo činností ČSS a programem přednáškové činnosti v budoucnu. Jednatel společnosti spolu s dr. R. Kollarem byli pověřeni přípravou cyklu přednášek na téma: „Význam statistiky pro řešení národohospodářských otázek přítomnosti“. Vedle přednášek v uvedeném cyklu se předpokládala přednáška o problémech sčítání lidu, československé úmrtnostní tabulce a o matematické statistice.

ČSS byl po jejím ustavení navázán také styk se zahraničními statistickými společnostmi. Výročních schůzí některých zahr. statistických společností se zúčastňovali také členové ČSS a podávali zprávy o průběhu jejich jednání.

Druhé valné shromáždění ČSS se konalo 27. května 1930. Vedle zprávy předsednictva o činnosti /viz výše/ se uskutečnila volba nového předsednictva společnosti.

Předsedou byl zvolen dr. V. Mildschuh, místopředsedy dr. B. Živanský a dr. A. Boháč, jednatelem dr. J. Janko, dále byli zvoleni revizoři účtů, pokladník a noví členové společnosti. Po provedených volbách se konala přednáška dr. V. Verunáče „Zásady vědecké organizace práce v praxi a statistika“.

Třetí valné shromáždění ČSS se uskutečnilo 19. června 1931 za předsednictví prof. Dr. V. Mildschuha s následujícím programem: 1. Zápis o jednání II. valného shromáždění, 2. Zpráva předsednictva o činnosti Společnosti v uplynulém roce, 3. Zpráva účetní a pokladní, 4. Volné návrhy, 5. Přednáška doc. Dr. J. Janka „Statistika a matematika".

V rámci zprávy o činnosti byli účastníci informování i o 2 mimořádných schůzích Společnosti, na nichž byly předneseny přednášky a organizovány příslušné diskuse. Přednášky jsou zveřejňovány ve Statistickém obzoru. Diskuse k příslušným bodům programu byla zaměřena na způsob zveřejňování přednášek a diskusí, placení členských příspěvků, honorářů za přednášky a zavedení systému přednáškových cyklů.

Čtvrté valné shromáždění ČSS se uskutečnilo dne 27. května 1932 se shodnou strukturou programu jako u třetího shromáždění. Přednášku „K nové organizaci studia konjunktury" přednesl dr. P. Smutný. V rámci zprávy předsednictva bylo poukázáno na 4 mimořádné schůze, na nichž byly předneseny přednášky k aktuálním statistickým problémům, které také vyšly tiskem.

Aby byl prohlouben styk se zahraničními statistickými společnostmi, provedlo valné shromáždění ČSS podle návrhu předsednictva volbu čestných členů zahraničních. Soubor těchto členů obsahuje řadu vynikajících mezinárodně uznávaných statistických odborníků, jako např. prof. I. Fishera, prof. G. U. Yuleho, prof. Zahna, dr. A. Julina, dr. H. W. Methorsta a dalších.

Na uvedeném shromáždění bylo také schváleno, aby noví členové /řádní i mimořádní/ byli voleni zpravidla po absolvování přednášky na mimořádné schůzi společnosti.

Páté řádné valné shromáždění ČSS se uskutečnilo 16. června 1933 za předsednictví prof. Dr. V. Mildschuha. Struktura programu byla obdobná jako u dřívějších zasedání.

Při jednání společnosti bylo vzpomenuto úmrtí členů v ČSR i čestného člena Lucien/a/ March/a/, budovatele a organizátora francouzské oficiální statistiky a aktivního člena Mezinárodního statistického institutu.

Po schválení zprávy předsednictva o činnosti, zprávy účetní a pokladní byly vykonány volby na další tříleté období. Hlavní funkcionáři zůstávají stejní – prof. Dr. V. Mildschuh, předseda, místopředsedové: dr. B. Živanský, doc. Dr. A. Boháč.

Následující valná shromáždění ČSS se konala v příslušných letech se stejnou strukturou jejich programů, nejsou proto jednotlivě popisována. Vedle přednášek konaných na valných shromážděních, byly pořádány přednášky statistických odborníků i na mimořádných schůzích společnosti.

V uvedeném období se konala dne 6. června 1934 zvláštní smuteční schůze statistické společnosti, která byla věnována památce zesnulého řádného člena

ČSS a vícepresidenta Státního úřadu statistického doc. Dr. J. Mráze. Smuteční shromáždění se konalo za účasti členů rodiny zesnulého, zástupců vysokých škol i dalších institucí. S projevy vystoupil president SÚS dr. Auerhan a řada dalších vedoucích pracovníků statistického úřadu.

Poslední předválečné shromáždění Československé statistické společnosti se uskutečnilo v roce 1938. Toto desáté valné shromáždění se konalo za předsednictví prof. V. Mildschuha dne 24. června 1938. Na shromáždění byl schválen zápis o jednání devátého valného shromáždění a zpráva předsednictva o činnosti společnosti v uplynulém roce.

Podle zprávy předsednictva měla Československá statistická společnost ke konci uvedeného období 38 stálých členů, 24 členů dopisujících a 13 členů čestných. Program jednání zahrnoval také uctění památky čestného člena prof. Fr. Žižka, profesora statistiky na univerzitě ve Frankfurtu nad Mohanem, který zemřel 20. května 1938. Prof. Žižek byl považován za jednoho z nejlepších německých znalců statistické metodologie.

Shromáždění bylo také informováno o činnosti předsednictva, které na svých dvou schůzích projednávalo podrobně přednáškový program a další administrativní otázky, a o mimořádných schůzích s příslušnými přednáškami. Přednášky byly zveřejněny ve Statistickém obzoru. Podle návrhu předsednictva se uskutečnila volba dvou nových členů společnosti. Po schválení pokladní zprávy následovala zpráva profesora ČVUT dr. J. Janka: „Rozbor některých nových dat naší populační statistiky“ s následnou diskusí.

Stručný popis činnosti Československé statistické společnosti v období před 2. světovou válkou může nepochybně vést k určitým zamyšlení nad další činnosti naší České statistické společnosti, jejímu zaměření, jejímu vztahu k odborně blízkým institucím, zejména Českého statistického úřadu, obsahu, zaměření a organizaci přednášek členů, zveřejňování informací o činnosti České statistické společnosti a činnostech jiných národních statistických společností, zapojování společnosti do statistiky významných akcí ke zvýšení jejího přínosu ve statistických zkoumáních, při využívání statistických výsledků a jejich interpretace v důležitém veřejném zájmu.

Určité rezervy lze spatřovat i v získávání nových členů tak, aby všechny obory statistiky na tomto úseku byly přiměřeně zastoupeny a odráželo se to i ve vyvážené činnosti statistické společnosti ve vztahu ke struktuře činnosti Mezinárodního statistického institutu.

Zvýšení úsilí na tomto úseku má své opodstatnění i v porovnání s předválečnou situací, kdy statistická společnost byla silně „statovsky“ zaměřena a volila za své členy špičkové statistické odborníky státní statistické služby, univerzitních pracovišť a vedoucí statistické pracovníky jiných institucí.

Contents / Obsah

Informační Bulletin České statistické společnosti vychází čtyřikrát do roka v českém vydání. Příležitostně i mimořádné české a anglické číslo.

Časopis je zařazen do seznamu Rady pro výzkum, vývoj a inovace, více viz server `http://www.vyzkum.cz/`

The Bulletin of the Czech Statistical Society is published quarterly. Most of the contributions are published in Czech and Slovak languages.

ISSN 1210–8022, DOI: 10.5300/IB

DOI je přiřazováno ve spolupráci s Čs. sdružením uživatelů TEXu.
Toto číslo bylo vytištěno s laskavou podporou Českého statistického úřadu.