POUŽITÍ KONEČNÝCH SMĚSÍ PRO MODELOVÁNÍ PŘÍJMOVÝCH ROZDĚLENÍ

Ivana Malá*

Zkoumání a modelování rozdělení mezd a příjmů z různých úhlů pohledu a různými metodami je dlouhodobě aktuálním tématem a problémem. Velký zájem dokumentuje také bohatá literatura týkající se nejrozumnějších přístupů (obsáhlý přehled Kleiber, Kotz, 2003, nebo McDonald, 1984). Na základě pravděpodobnostních modelů lze činit závěry nejen o číselných charakteristikách rozdělení (jako jsou charakteristiky polohy, variability nebo šikmosti), ale odpovídat na podrobnější otázky týkající se rizika chudoby či aspektů kvality života založených na příjmech. Nejčastěji je třeba sledovat vývoj rozdělení příjmů (různých charakteristik) nebo srovnávat rozdělení příjmů jednak v různých částech populace, definovaných kritérii jako vzdělání, zaměstnání, odvětví či region bydliště, jednak příjmy porovnávat na mezinárodní úrovni.

Jako model pro příjmy (nebo mzdy) jsou v literatuře používána různá pravděpodobnostní rozdělení. Kromě logaritmicko-normálního rozdělení se dvěma či třemi parametry (například Bílková, 2008) mezi příjmová rozdělení patří také Dagumovo rozdělení, zobecněné lambda, zobecněné beta, Pareto rozdělení a další (Kleiber, Kotz, 2003, McDonald, 1984, Flachaire, Nunez, 2007, Milanovic, 2002). Vhodné modely pro příjmy v České republice byly hledány například v Bartošová, Bína, 2009, Bartošová, 2006, Pavelka, 2009, Pacáková, Sipková, 2007.

V tomto textu je rozdělení čistých ročních příjmů na jednu osobu českých domácností popsáno směsí logaritmicko-normálních rozdělení. Domácnosti jsou děleny podle vzdělání osoby v čele domácnosti (faktor o 5 hodnotách – základní, střední, úplné střední, bakalářské a magisterské vzdělání) a dále podle toho, zda v domácnosti jsou děti (dvě hodnoty ano/ne), a podrobněji podle toho, kolik dětí v domácnosti je (faktor o 5 hodnotách 0, 1, 2, 3, více než 3 děti). Logaritmicko-normální rozdělení je použito pro rozdělení ve skupině a rozdělení příjmů pro všechny domácnosti je konstruováno jako směs zmíněných rozdělení v podskupinách s vahami rovnými procentnímu zastoupení podskupin ve výběru. Tento model poskytuje informaci pro každou podskupinu zvlášť i pro celou populaci.

Dále je použit model bez informace o příslušnosti ke složce pro dvě až pět umělých složek. V tomto případě jsou složky konstruovány tak, aby výsledný model co nejlépe vystihoval empirické hodnoty. V takovém případě je třeba hledat interpretaci jednotlivých složek (podskupin) a je třeba si uvědomit, že postup nedává (jednoznačně)

* Vysoká škola ekonomická v Praze, Fakulta informatiky a statistiky (malai@vse.cz).

Článek byl podpořen grantovým projektem IGS 24/2010 Vysoké školy ekonomické v Praze.

příslušnost jednotlivých domácností ke složkám směsi, je možné takové pravděpodobnosti pro každou domácnost odhadnout. V tomto textu nebyly použity žádné vysvětlující proměnné pro pravděpodobnosti příslušnosti domácností do složek.

1. Metodologie

1.1 Konečné směsi pravděpodobnostních rozdělání

Předpokládejme, že X je náhodná veličina se spojitým pravděpodobnostním rozdělením a toto rozdělení lze popsat jako směs K pravděpodobnostních rozdělání s hustotami $f_j(x; \theta_j)$, $j = 1, \dots, K$. V tomto textu budeme předpokládat, že tyto hustoty pocházejí ze stejného rozdělení, liší se pouze v hodnotách p neznámých parametrů, které obsahuje (p -rozměrný) vektor parametrů, který pro j -tou složku označíme θ_j . Zvolme dále váhy π_j , $j = 1, \dots, K$ splňující podmínky

$$\sum_{j=1}^K \pi_j = 1, \quad 0 \leq \pi_j \leq 1, \quad j = 1, \dots, K. \quad (1)$$

Konečnou směs rozdělání pak definujeme jako vážený průměr hustot f_j s vahami π_j , budeme tedy předpokládat, že náhodná veličina X má hustotu

$$f(x; \psi) = \sum_{j=1}^K \pi_j f_j(x; \theta_j). \quad (2)$$

Hustota definovaná pomocí závisí na vektoru parametrů ψ

$$\psi = (\pi_1, \dots, \pi_{K-1}, \theta_1, \dots, \theta_K), \quad j = 1, \dots, K, \quad (3)$$

který obsahuje $(K-1)$ parametrů π_j a Kp parametrů obsažených ve vektorech θ_j . Z definice ihned plyne, že distribuční funkce F náhodné veličiny X je směsí distribučních funkcí jednotlivých složek, a totéž platí i o střední hodnotě. Tyto jednoduché vztahy ovšem neplatí pro rozptyl (a vyšší momenty) rozdělení směsi, stejně jako pro kvantily. 100P% kvantil x_p náhodné veličiny X je třeba počítat z definice kvantilu jako řešení rovnice

$$F(x_p; \psi) = \sum_{j=1}^K \pi_j F_j(x_p; \theta_j) = P, \quad 0 < P < 1. \quad (4)$$

Tuto rovnici je obvykle třeba řešit numericky, přičemž vážený průměr odpovídajících skupinových kvantilů x_{jp} , $j = 1, \dots, K$, jednotlivých složek směsi $\sum_{j=1}^K \pi_j x_{jp}$ může sloužit jako velmi dobrá počáteční hodnota pro aproximační proces. Rozptyl směsi lze opět určit ze středních hodnot, rozptylů a pravděpodobností jednotlivých složek, nikoli však jako směs rozptylů, ale z definice rozptylu.

Předpokládejme, že máme k dispozici náhodný výběr x_i , $i = 1, \dots, n$ z rozdělení s hustotou (2). Odhad neznámého vektoru parametrů ψ metodou maximální věrohodnosti znamená najít bod $\hat{\psi}$, ve kterém nabývá maxima věrohodnostní funkce

$$L(\boldsymbol{\psi}) = \prod_{i=1}^n f(x_i; \boldsymbol{\psi}) = \prod_{i=1}^n \left(\sum_{j=1}^K \pi_j f_j(x_i; \boldsymbol{\theta}_j) \right). \quad (5)$$

1.2 Známa příslušnost pozorování ke složce ve směsi

Předpokládejme nyní, že u každého pozorování ve výběru známe složku směsi, ze které pozorování pochází. Data tohoto typu se vyskytují tehdy, skládá-li se populace, ze které vybíráme, z K podskupin. Rozdělení náhodné veličiny X v těchto podskupinách tvoří složky směsi. Pokud jsme schopni určit ke každému pozorování podskupinu, do které pozorování patří, budeme hovořit o modelu s úplnými daty. V takovém případě pozorování pocházející z j -té složky přispívá do celkové věrohodnostní funkce pouze hodnotou $\pi_j f(x_i; \boldsymbol{\theta}_j)$. Definujme nyní vektory $\mathbf{z}_i = (z_{i1}, z_{i2}, \dots, z_{iK})$, $i = 1, \dots, n$ předpisem

$$\begin{aligned} z_{ij} &= 1 \quad \text{pozorování } x_i \text{ pochází z } j\text{-té složky hustoty směsi} \\ &= 0 \quad \text{jinak.} \end{aligned} \quad (6)$$

Věrohodnostní funkci (5) pak lze přepsat ve tvaru (Titterton, 1985)

$$L(\boldsymbol{\psi}) = \prod_{i=1}^n \left(\prod_{j=1}^K \pi_j^{z_{ij}} f_j(x_i; \boldsymbol{\theta}_j)^{z_{ij}} \right), \quad (7)$$

kde z_{ij} jsou složky vektorů $\mathbf{z}_i$ definovaných v (6).

Logaritmováním věrohodnostní funkce (7) dostáváme logaritmickou věrohodnostní funkci l ve tvaru

$$l(\boldsymbol{\psi}) = \ln L(\boldsymbol{\psi}) = \sum_{i=1}^n \sum_{j=1}^K z_{ij} \ln \pi_j + \sum_{i=1}^n \sum_{j=1}^K z_{ij} \ln f_j(x_i; \boldsymbol{\theta}_j). \quad (8)$$

Logaritmická věrohodnostní funkce se tedy skládá ze dvou částí, z nichž první závisí pouze na neznámých vahách v definici směsi a druhá pouze na hustotách jednotlivých složek. Vektory $\mathbf{z}$ jsou podle předpokladu známy. Obě části pak lze maximalizovat samostatně. Připomeňme, že z definice vektorů $\mathbf{z}$ plyne, že pro pevné j je součet $\sum_{i=1}^n z_{ij}$

roven absolutní četnosti pozorování ve výběru pocházejících z j -té složky. Maxima první části v (8) je tedy dosaženo pro

$$\hat{\pi}_j = \frac{\sum_{i=1}^n z_{ij}}{n}, \quad 1 \leq j \leq K. \quad (9)$$

Část druhou lze rozdělit na jednotlivé složky a nalézt maximálně věrohodné odhady parametrů $\boldsymbol{\theta}$ v každé složce zvlášť maximalizací logaritmické věrohodnostní funkce

$$\sum_{i: z_{ij}=1} \ln(f_j(x_i; \boldsymbol{\theta}_j)), \quad j = 1, \dots, K. \quad (10)$$

1.3 Neznámá příslušnost pozorování ke složce ve směsi

Předpokládejme nyní, že příslušnost pozorování k některé složce ve směsi není známa. V takovém případě hovoříme o modelu s neúplnými daty, logaritmickou věrohodnostní funkci nelze zapsat ve tvaru (8) a při odhadu parametrů směsi rozdělení je třeba vycházet ze vzorce (5).

Hodnoty vektorů definovaných v (6) nejsou známy, jedná se o náhodné vektory s multinomickým rozdělením s parametry K a $(\pi_1, \pi_2, \dots, \pi_K)$. V tomto textu nebudeme uvažovat závislost pravděpodobnostního rozdělení na indexu i .

Při logaritmování výrazu (5) dostáváme logaritmickou věrohodnostní funkci ve tvaru

$$l(\boldsymbol{\psi}) = \ln \left(\prod_{i=1}^n \left(\sum_{j=1}^K \pi_j f_j(x_i; \boldsymbol{\theta}_j) \right) \right) = \sum_{i=1}^n \ln \left(\sum_{j=1}^K \pi_j f_j(x_i; \boldsymbol{\theta}_j) \right), \quad (11)$$

a výraz nelze upravit tak, aby nebylo nutné logaritmovat součet. Vhodným nástrojem pro nalezení odhadů směsi je EM-algoritmus, numerický iterační postup, ve kterém se každá iterace skládá ze dvou kroků. Je třeba vyjít z počátečních přiblížení neznámého vektoru parametrů $\boldsymbol{\psi}$, dále se opravují zvlášť odhady pravděpodobností a parametry rozdělení. První krok (**E-krok** podle expectation) hledá novou aproximaci vektoru pravděpodobností $\boldsymbol{\pi}$, druhý krok (**M-krok** podle maximization) hledá maximálně věrohodné odhady parametrů $\boldsymbol{\theta}$ rozdělení složek pro pevné hodnoty pravděpodobností složek získané v kroku prvním. Tyto dva kroky jsou opakovány tak dlouho, až již nedochází ke změně hodnot parametrů, a lze tedy nalezené parametry považovat řešení optimalizační úlohy (Titterton, 1985, Pavelka, 2009). EM algoritmus je běžně implementován ve statistických programech, obvykle však obsahuje pouze směsi normálních rozdělení. K nalezení odhadů v tomto textu byl použit balíček *flexmix* v programu R 2.13.1.

1.4 Logaritmicko-normální rozdělení

V tomto textu budeme pro popis rozdělení jednotlivých složek používat logaritmicko-normální rozdělení, které se běžně používá pro modelování rozdělení příjmů. Ukážeme, že i dvouparametrické rozdělení lze velmi dobře použít, pokud se místo jedné hustoty použije směs takových hustot.

Předpokládejme, že rozdělení náhodné veličiny X lze popsat směsí K dvouparametrických logaritmicko-normálních rozdělení. Hustota směsi je popsána (po dosazení do (2)) výrazem

$$f(x; \boldsymbol{\psi}) = \sum_{j=1}^K \pi_j f_j(x; \mu_j, \sigma_j^2) = \sum_{j=1}^K \frac{\pi_j}{\sqrt{2\pi}\sigma_j} \exp \left(-\frac{(\ln x - \mu_j)^2}{2\sigma_j^2} \right). \quad (12)$$

Vektor neznámých parametrů $\boldsymbol{\psi}$ má $(K-1)+2K$ složek a platí

$$\boldsymbol{\psi} = (\pi_j, j=1, \dots, K-1, \mu_j, \sigma_j^2, j=1, \dots, K). \quad (13)$$

Veličina $Z = \ln X$ má rozdělení popsané směsí normálních rozdělení ve tvaru

$$f(z; \Psi) = \sum_{j=1}^K \frac{\pi_j}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{(z - \mu_j)^2}{2\sigma_j^2}\right), \quad (14)$$

tedy směsí normálních rozdělení $N(\mu_j; \sigma_j^2)$ s pravděpodobnostmi složek π_j . Tento vztah využijeme pro odhad parametrů rozdělení v případě neúplných dat.

V případě známých příslušností do skupin (část 1.2) byly odhady neznámých parametrů $\hat{\pi}_j$, $\hat{\mu}_j$, $\hat{\sigma}_j^2$ v (13) nalezeny jako ($j = 1, \dots, K$) (Bílková, 2009)

$$\hat{\pi}_j = \frac{n_j}{n}, \quad \hat{\mu}_j = \frac{\sum_{i=1}^n \ln x_i}{n}, \quad \hat{\sigma}_j^2 = \frac{\sum_{i=1}^n (\ln x_i - \hat{\mu}_j)^2}{n}, \quad (15)$$

kde n_j je absolutní četnost j -té složky ve výběru, n rozsah výběru. Obdobně je možné v citované práci nalézt potřebné vztahy pro charakteristiky logaritmicko-normálního rozdělení.

Pro neznámou příslušnost ke složce byly parametry (pro 1–5 složek) odhadnuty v souboru hodnot $\ln x$. Pro hledání parametrů normálních rozdělení pro daný počet složek ($K = 1-5$) byl použit EM algoritmus obsažený v balíčku *flexmix* (Grün, Leisch, 2008). Aproximační procedura byla ve všech případech spuštěna z různých počátečních podmínek vzhledem k tomu, že EM algoritmus může konvergovat k lokálnímu extrému, a nikoliv ke globálnímu maximu. Počet počátečních podmínek byl volen tak, aby se zvětšoval s počtem složek směsi. Byl využit jednak náhodný generátor příslušnosti (obsažený ve *flexmix*), který každému pozorování přiřadí náhodně příslušnost ke složce a dále příslušnost odhadnutá pomocí uspořádaného výběru hodnot $\ln x$ pro různé volby pravděpodobností jednotlivých složek. Zvláště v případě většího počtu složek (4 a 5) opravdu algoritmus našel různá řešení z různých počátečních podmínek, případně konvergence zcela selhala i po velkém počtu iterací. V případě existence různých řešení bylo vybráno řešení s maximální dosaženou hodnotou věrohodnostní funkce a tato volba ještě ověřena spuštěním EM algoritmu z počátečního řešení daného třemi dosaženými nejlepšími aproximacemi. Kvalita modelů se stejným počtem složek byla srovnávána hodnotou logaritmické věrohodnostní funkce v řešení, modely s různým počtem složek (a tedy i parametrů) pak pomocí hodnoty Akaikova informačního kritéria AIC definovaného vztahem

$$AIC = -2\ln(L(\Psi)) + 2\text{počet parametrů modelu}. \quad (16)$$

V případě směsi normálních rozdělení jsou počty parametrů modelu uvedeny v tabulce 1. Z tabulky je patrný rychlý nárůst počtu parametrů s každou složkou zahrnutou do modelu.

Tabulka 1**Počet parametrů pro modely se složkami 1–5**

K parametry	1		2		3		4		5	
	π_j	μ_j, σ_j^2	π_j	μ_j, σ_j^2	π_j	μ_j, σ_j^2	π_j	μ_j, σ_j^2	π_j	μ_j, σ_j^2
Počet parametrů	0	2	1	4	2	6	3	8	4	10

Pramen: Vlastní výpočty.

2. Použitá data a výsledky

Modely popsané v první části nyní použijeme pro popis rozdělení čistého ročního příjmu (v Kč) na jednu osobu českých domácností v letech 2004 až 2008. Použitá data pocházejí z šetření Životní podmínky, které provádí od roku 2005 každoročně Český statistický úřad. Jedná se o národní modul šetření s názvem European Union – Statistics on Income and Living Conditions (zkratka EU-SILC), šetření byla provedena v letech 2005 až 2009, zahrnují však příjmy z let 2004–2009. Z rozsáhlého šetření příjmů byly použity pouze informace o celkovém čistém příjmu domácnosti a o počtu členů domácnosti. Z těchto hodnot byl vydělením určen čistý příjem na jednu osobu. Dále byly využity koeficienty (váhy) poskytované pro vybrané domácnosti. Rozsahy výběrů byly postupně 4 351, 7 483, 9 675, 11 294 a 9 911 domácností.

Pro třídění do skupin se známými příslušnostmi ke složkám (úplná data podle 1.2) byly použity údaje o vzdělání osoby v čele domácnosti (základní – Z, střední – S, úplné střední s maturitou – ÚS, bakalářské – BS a magisterské a doktorské – MS) a byl použit model směsi s 5 složkami logaritmicko-normálního rozdělení představujícími rozdělení příjmu domácností s osobami v čele s daným dosaženým vzděláním. K posouzení vlivu přítomnosti či nepřítomnosti vyživovaných dětí v domácnosti (případně jejich počtu) na příjem domácnosti byly domácnosti tříděny na bezdětné a s dětmi (dvě složky) a podle počtu dětí na žádné, jedno, dvě, tři a čtyři a více dětí (tedy pět složek). I po spojení má skupina 4 a více dětí v domácnosti i ve velkých výběrech šetření, které byly k dispozici, velmi malou četnost výskytu (jak je vidět v tabulce 3 na relativních četnostech $\hat{\pi}$). Další skupinou s relativně malými četnostmi je skupina s bakalářským vzděláním vzhledem k tomu, že titul bakalář se neuděluje dlouho a navíc málo studujících tímto titulem končí a nepokračuje na magisterském studiu.

Tabulka 2**Odhadnuté parametry složek směsi, složky definované vzděláním osoby v čele domácnosti**

2004	Z	S	ÚS	BS	MS
$\hat{\mu}$	11,342	11,422	11,558	11,652	11,844
$\hat{\sigma}$	0,345	0,405	0,460	0,518	0,516
$\hat{\pi}$	0,127	0,451	0,293	0,015	0,114
2005	Z	S	ÚS	BS	MS
$\hat{\mu}$	11,377	11,459	11,61	11,825	11,862
$\hat{\sigma}$	0,353	0,394	0,444	0,522	0,522
$\hat{\pi}$	0,126	0,460	0,290	0,018	0,107
2006	Z	S	ÚS	BS	MS
$\hat{\mu}$	11,46	11,546	11,683	11,851	11,933
$\hat{\sigma}$	0,365	0,388	0,436	0,517	0,488
$\hat{\pi}$	0,122	0,460	0,292	0,017	0,109
2007	Z	S	ÚS	BS	MS
$\hat{\mu}$	11,539	11,636	11,758	11,891	11,987
$\hat{\sigma}$	0,378	0,372	0,421	0,456	0,472
$\hat{\pi}$	0,017	0,108	0,459	0,293	0,123
2008	Z	S	ÚS	BS	MS
$\hat{\mu}$	11,629	11,711	11,849	11,981	12,076
$\hat{\sigma}$	0,36	0,363	0,42	0,527	0,507
$\hat{\pi}$	0,125	0,452	0,297	0,017	0,109

Pramen: Vlastní výpočty.

Kromě využití informace o příslušnosti ke složce směsi byly dále domácnosti tříděny do složek jen na základě hodnoty sledovaného příjmu (model s neúplnými daty 1.3). Pro předkládanou analýzu nebyly využity pro třídění do složek žádné další vysvětlující proměnné.

V tabulkách 2 až 4 jsou uvedeny odhady parametrů určené podle (15) (tabulky 2 a 3) a pomocí EM algoritmu v tabulce 4. Odhady parametru μ udávají odhad střední hodnoty logaritmu příjmu v dané podskupině, odhad parametru σ směrodatnou odchylku tohoto logaritmu příjmu.

Tabulka 3**Odhadnuté parametry složek směsi, složky definované přítomností (či počtem dětí) v domácnosti**

			Počet dětí			
2004	Děti ne	Děti ano	1	2	3	4 a více
$\hat{\mu}$	11,624	11,258	11,381	11,21	10,89	10,886
$\hat{\sigma}$	0,391	0,473	0,472	0,435	0,431	0,342
$\hat{\pi}$	0,674	0,326	0,148	0,148	0,025	0,005
2005	Děti ne	Děti ano	1	2	3	4 a více
$\hat{\mu}$	11,662	11,298	11,444	11,246	10,934	10,755
$\hat{\sigma}$	0,385	0,461	0,449	0,415	0,413	0,463
$\hat{\pi}$	0,671	0,329	0,144	0,152	0,026	0,006
2006	Děti ne	Děti ano	1	2	3	4 a více
$\hat{\mu}$	11,734	11,392	11,529	11,325	11,092	10,891
$\hat{\sigma}$	0,381	0,452	0,428	0,427	0,437	0,376
$\hat{\pi}$	0,675	0,325	0,147	0,147	0,025	0,006
2007	Děti ne	Děti ano	1	2	3	4 a více
$\hat{\mu}$	11,802	11,49	11,625	11,426	11,182	10,992
$\hat{\sigma}$	0,376	0,435	0,419	0,391	0,460	0,382
$\hat{\pi}$	0,682	0,318	0,142	0,1452	0,026	0,005
2008	Děti ne	Děti ano	1	2	3	4 a více
$\hat{\mu}$	11,877	11,586	11,728	11,508	11,312	11,026
$\hat{\sigma}$	0,377	0,442	0,432	0,396	0,431	0,338
$\hat{\pi}$	0,689	0,311	0,143	0,137	0,026	0,004

Pramen: Vlastní výpočty.

V tabulce 4 jsou uvedeny odhady parametrů pro neúplná data pro modely s počtem složek jedna až čtyři. Složky v tabulce jsou uspořádány vzestupně podle hodnoty odhadu parametru μ , tedy podle odhadnuté střední hodnoty logaritmu příjmu na jednu osobu v dané složce. I když střední hodnota logaritmicko-normálního rozdělení závisí nejen na hodnotě μ , ale také na rozptylu σ^2 , pořadí odhadnutých středních hodnot složek původních (nelogaritmovaných příjmů) v tabulce 8 je stejné, neboť podle tabulky 4 vyšší hodnotu parametru μ doprovází větší hodnota parametru σ .

Tabulka 4

Odhady parametrů logaritmicko-normálního rozdělení jednotlivých složek pro $K = 2-4$

2004	$\hat{\pi}_1$	$\hat{\mu}_1$	$\hat{\sigma}_1$	$\hat{\pi}_2$	$\hat{\mu}_2$	$\hat{\sigma}_2$	$\hat{\pi}_3$	$\hat{\mu}_3$	$\hat{\sigma}_3$	$\hat{\pi}_4$	$\hat{\mu}_4$	$\hat{\sigma}_4$
$k = 2$	0,370	11,469	0,162	0,630	11,524	0,557						
$k = 3$	0,259	11,460	0,127	0,616	11,508	0,445	0,125	11,571	0,793			
$k = 4$	0,245	11,467	0,585	0,372	11,467	0,585	0,371	11,542	0,379	0,012	12,447	0,891
2005	$\hat{\pi}_1$	$\hat{\mu}_1$	$\hat{\sigma}_1$	$\hat{\pi}_2$	$\hat{\mu}_2$	$\hat{\sigma}_2$	$\hat{\pi}_3$	$\hat{\mu}_3$	$\hat{\sigma}_3$	$\hat{\pi}_4$	$\hat{\mu}_4$	$\hat{\sigma}_4$
$k = 2$	0,382	11,506	0,160	0,618	11,564	0,522						
$k = 3$	0,257	11,497	0,121	0,631	11,539	0,435	0,112	11,661	0,808			
$k = 4$	0,322	11,396	0,452	0,248	11,49	0,118	0,308	11,663	0,358	0,122	11,729	0,776
2006	$\hat{\pi}_1$	$\hat{\mu}_1$	$\hat{\sigma}_1$	$\hat{\pi}_2$	$\hat{\mu}_2$	$\hat{\sigma}_2$	$\hat{\pi}_3$	$\hat{\mu}_3$	$\hat{\sigma}_3$	$\hat{\pi}_4$	$\hat{\mu}_4$	$\hat{\sigma}_4$
$k = 2$	0,438	11,580	0,176	0,562	11,657	0,557						
$k = 3$	0,270	11,565	0,131	0,586	11,637	0,411	0,144	11,672	0,766			
$k = 4$	0,254	11,561	0,127	0,308	11,625	0,584	0,416	11,649	0,361	0,022	11,951	1,021
2007	$\hat{\pi}_1$	$\hat{\mu}_1$	$\hat{\sigma}_1$	$\hat{\pi}_2$	$\hat{\mu}_2$	$\hat{\sigma}_2$	$\hat{\pi}_3$	$\hat{\mu}_3$	$\hat{\sigma}_3$	$\hat{\pi}_4$	$\hat{\mu}_4$	$\hat{\sigma}_4$
$k = 2$	0,518	11,648	0,205	0,482	11,759	0,563						
$k = 3$	0,307	11,628	0,146	0,63	11,732	0,427	0,064	11,758	0,919			
$k = 4$	0,01	11,522	1,334	0,254	11,624	0,133	0,494	11,709	0,360	0,242	11,777	0,602
2008	$\hat{\pi}_1$	$\hat{\mu}_1$	$\hat{\sigma}_1$	$\hat{\pi}_2$	$\hat{\mu}_2$	$\hat{\sigma}_2$	$\hat{\pi}_3$	$\hat{\mu}_3$	$\hat{\sigma}_3$	$\hat{\pi}_4$	$\hat{\mu}_4$	$\hat{\sigma}_4$
$k = 2$	0,624	11,736	0,237	0,376	11,871	0,605						
$k = 3$	0,275	11,681	0,144	0,613	11,822	0,382	0,112	11,85	0,830			
$k = 4$	0,231	11,670	0,134	0,439	11,813	0,315	0,297	11,82	0,545	0,033	11,957	1,044

Pramen: Vlastní výpočty.

V tabulce 5 jsou všechny modely srovnány podle hodnoty Akaikova kritéria ((16) a tabulka 1). Připomeňme, že vzhledem k různému počtu pozorování v jednotlivých letech nelze přímo srovnávat kvalitu modelů mezi jednotlivými roky. Ve všech sledovaných letech hodnota kritéria klesá pro počty složek jedna až čtyři a pro pět složek začíná růst. V případě pěti složek penalizace za počet parametrů v (16) převáží růst hodnoty věrohodnostní funkce. Proto v tabulce 4 nejsou uvedeny parametry modelu s pěti složkami. Popisovat rozdělení příjmů jedním logaritmicko normálním rozdělením také není vhodné (tabulka 5) a proto hodnoty odhadnutých parametrů tohoto modelu nejsou uvedeny v tabulce 4. Hodnota Akaikova kritéria pro tento model je však zajímavá pro srovnání s hodnotami modelů směsí konstruovaných na základě vzdělání a počtu dětí. Je vidět, že použití těchto směsí nevede k vylepšení modelu celkového rozdělení příjmů. Dostáváme tedy zajímavou informaci o podskupinách, smísení těchto rozdělení do rozdělení celkového ale nepřináší výrazné zlepšení modelu. Na druhé straně konstrukce umělých podskupin (jak plyne také z jejich konstrukce) vede k vylepšení modelu. Již dvě složky ve směsi znamenají výrazné vylepšení aproximace empirického rozdělení modelem, výsledky modelů se třemi a čtyřmi složkami jsou již velmi podobné

a velmi dobře vystihují empirické rozdělení pozorování příjmů. Vzhledem k velikosti výběrů však testy dobré shody vycházejí ve všech případech významně.

Tabulka 5

Srovnání kvality uvažovaných modelů, hodnoty Akaikova kritéria (16)

Model	2004	2005	2006	2007	2008
$K = 1$	105 575	181 891	236 278	276 868	244 598
$K = 2$	105 078	181 000	235 073	275 458	243 669
$K = 3$	105 031	180 862	234 945	275 264	243 067
$K = 4$	105 026	180 860	234 940	275 223	243 040
$K = 5$	105 032	181 234	234 937	285 382	243 159
Děti (ano/ne)	110 349	190 074	246 951	289 447	259 606
Děti (0, 1, 2, 3, >3)	113 033	194 660	251 790	296 107	260 956
Vzdělání	99 481	172 001	223 170	279 644	230 186

Pramen: Vlastní výpočty.

V tabulkách 6 až 8 jsou uvedeny maximálně věrohodné odhady základních charakteristik polohy (střední hodnota) a variability (směrodatná odchylka) pro jednotlivé složky na základě odhadnutých parametrů (tabulky 2 až 4). Při interpretaci charakteristik jednotlivých složek by bylo vhodné vzít v úvahu inflaci v České republice ve sledovaných letech. Inflace byla celkem 14,1 procenta (podle CZSO), meziroční pak postupně 1,9, 2,5, 2,8 a 6,3 procenta. V posledním řádku tabulek je uveden koeficient růstu charakteristiky za sledované období.

Střední hodnoty pro dělení podle počtu dětí (tabulka 6) vzrostly o 15 % (domácnosti se 4 a více dětmi) až 52 % (domácnosti se 3 dětmi), zastoupení obou skupin je však velmi malé. Ostatní hodnoty se pohybovaly v mnohem užším pásmu mezi 28 a 32 procenty. Pro směrodatné odchylky došlo k nárůstu (vynecháme-li dvě výše zmíněné skupiny) o 19 až 23 procent. Pro dělení podle vzdělání osoby v čele domácnosti (tabulka 7) střední hodnoty složek vzrostly o 31–40 procent, směrodatné odchylky o 17–42 procent hodnoty roky 2004.

V tabulce 6 je patrný vliv počtu dětí v domácnosti na sledovaný příjem na jednu osobu. Podle očekávání střední hodnota příjmu klesá s počtem dětí v domácnosti. Pro obě charakteristiky je rozdíl mezi hodnotou pro bezdětné domácnosti a domácnosti se třemi dětmi přibližně 60 tisíc korun. Rozdíl by byl menší, pokud by byl analyzován příjem na ekonomickou jednotku, který dětem dává váhu 0,3 místo 1, jako ve výpočtu na jednu osobu a druhé dospělé osobě v domácnosti váhu 0,5. Proto u úplné rodiny se dvěma dětmi se celkový příjem dělí jen počtem ekonomických jednotek 2,1 místo počtem členů 4. U domácnosti se dvěma dospělými a čtyřmi dětmi je to dokonce 2,7 místo 6, a tedy výsledný přepočítaný příjem je více než dvakrát větší.

V tabulce 7 je vidět očekávaný pozitivní vliv vzdělání na příjem, i když v tomto textu jde o vliv vzdělání osoby v čele domácnosti. Ve sledovaných letech byla odhadnutá

střední hodnota příjmu pro magisterské vzdělání o 62–77 procent větší než pro domácnosti s osobou v čele se základním vzděláním (nebo bez vzdělání), zatímco směrodatné odchylky byly 2,5–3krát větší.

Tabulka 6

Odhady středních hodnot a směrodatných odchylek složek ve směsi logaritmicko-normálních rozdělení, dělení podle počtu dětí (index charakterizuje změnu 2004–2008)

Rok	Děti ne		Děti ano		Počet dětí 1		Počet dětí 2		Počet dětí 3		Počet dětí	
	$E(X)$	$\sqrt{D(X)}$	$E(X)$	$\sqrt{D(X)}$	$E(X)$	$\sqrt{D(X)}$	$E(X)$	$\sqrt{D(X)}$	$E(X)$	$\sqrt{D(X)}$	$E(X)$	$\sqrt{D(X)}$
2004	120625	49026	86670	49026	97968	49026	81195	37059	58858	26593	56641	26593
2005	125005	49967	89703	49967	103239	49967	83459	36182	61040	26324	52166	26324
2006	134132	53016	98140	53016	111368	53016	90777	40598	72221	33129	57624	33129
2007	143299	55842	107432	55842	122123	55842	98957	40219	79842	38758	63893	38758
2008	154518	60386	118620	60386	136123	60386	107625	44346	89759	40554	65064	40554
index	1,28	1,23	1,37	1,23	1,39	1,23	1,33	1,20	1,53	1,52	1,15	1,52

Pramen: Vlastní výpočty.

Tabulka 7

Odhady středních hodnot a směrodatných odchylek složek ve směsi logaritmicko-normálních rozdělení, dělení podle vzdělání (index charakterizuje změnu 2004–2008)

Rok	Základní		Střední		Úplné střední		Bakalářské		Magisterské	
	$E(X)$	$\sqrt{D(X)}$	$E(X)$	$\sqrt{D(X)}$	$E(X)$	$\sqrt{D(X)}$	$E(X)$	$\sqrt{D(X)}$	$E(X)$	$\sqrt{D(X)}$
2004	89457	31804	99113	41844	116285	56450	131421	72909	159075	87 862
2005	92902	33843	102398	41962	121609	56768	156567	87625	162469	90 928
2006	101378	38270	111443	44920	130359	59648	160275	88720	171454	88907
2007	110243	43206	121427	47454	139612	61480	161938	77854	179583	89711
2008	119826	44574	130207	48866	152848	67134	183481	103813	199691	108111
index	1,34	1,40	1,31	1,17	1,31	1,19	1,40	1,42	1,26	1,23

Pramen: Vlastní výpočty.

V tabulce 8 jsou uvedeny charakteristiky pro umělé složky směsi pro model se dvěma a třemi komponentami. Dvě složky by bylo možné interpretovat tak, že domácnosti jsou rozděleny do dvou skupin, na skupinu s nižšími a na skupinu s vyššími příjmy (a také větší variabilitou popsanou například variačním koeficientem). Odhadnuté hodnoty vah (tabulka 4) pak přibližují podíl jednotlivých složek mezi domácnostmi. Je patrné, že zastoupení první složky s nižšími příjmy roste s časem. V případě tří složek se zdá, že se z domácností s vyššími příjmy vydělí přibližně 12 procent do skupiny s vysokými příjmy a zbytek utvoří středně příjmové pásmo. Při této interpretaci má smysl uvažovat také změnu v letech 2004–2008, která je indexem popsána v posledním

řádku tabulky. Odhadnuté střední hodnoty rostou stejně jako směrodatné odchylky. V případě dvou složek rostou variační koeficienty první složky od 16 % v roce 2004 do 24 % v roce 2008, druhá složka od 60 procent do 66 procent. Pro model se třemi složkami podobně roste variační koeficient od složky charakterizující příjmy domácností s nízkými příjmy (růst od 12 do 14 procent v roce 2008), přes složku charakterizující středně příjmové domácnosti (kolísání v intervalu 40–45 procent) až po složku popisující vysoké příjmy s kolísáním variačního koeficientu kolem 100 procent.

Tabulka 8

Odhady středních hodnot a směrodatných odchylek složek ve směsi logaritmicko-normálních rozdělení, neznámá příslušnost ke složce (index charakterizuje změnu 2004–2008)

Rok	K = 2				K = 3					
	První složka		Druhá složka		První složka		Druhá složka		Třetí složka	
	$E(X)$	$\sqrt{D(X)}$	$E(X)$	$\sqrt{D(X)}$	$E(X)$	$\sqrt{D(X)}$	$E(X)$	$\sqrt{D(X)}$	$E(X)$	$\sqrt{D(X)}$
2004	96967	15812	118081	71218	95613	12192	109866	51413	145136	135797
2005	100589	16198	120601	67496	99143	12040	112827	51496	160722	154246
2006	108607	19264	134878	81349	106253	13979	123187	52845	157216	140455
2007	116893	24217	149864	91523	113398	16645	136375	60990	194908	224521
2008	128551	30900	171787	114208	119535	17303	146527	58079	197689	196849
index	1,33	1,95	1,45	1,60	1,25	1,42	1,33	1,13	1,36	1,45

Pramen: vlastní výpočty.

Závěr

Model konečné směsi rozdělení je užitečný při modelování pravděpodobnostního rozdělení náhodných veličin. Hustotu složitého rozdělení často s více vrcholy popisuje pomocí váženého průměru (obecně velkého počtu) hustot jednoduchých, vhodných rozdělení se známými vlastnostmi. Pokud sledovaná veličina je definována na základním souboru, který se skládá z disjunktních podmnožin, lze si představit různé pravděpodobnostní rozdělení náhodné veličiny v jednotlivých složkách. Tato rozdělení jsou pak popsána hustotami, z kterých je pak pomocí vah odrážejících zastoupení dané podmnožiny v celé sledované populaci konstruováno rozdělení náhodné veličiny. Problém odhadu parametrů pak závisí na tom, zda pro konkrétní pozorování lze, nebo nelze pozorovat příslušnost k některé ze složek. Nejčastěji je pro všechny složky směsi použito stejné rozdělení (a složky se liší jen volbou parametrů), tento postup byl použit i v tomto textu.

Další možností použití směsi rozdělení je popsat rozdělení směsí různých rozdělení tak, aby bylo možné použít například jiná rozdělení pro střední část hodnot náhodné veličiny a jiné (například Paretovo) pro hodnoty malé či velké.

Při modelování příjmů na osobu dochází k výraznému vylepšení (samostatně nepřilíhš vhodného) logaritmicko-normálního rozdělení již při použití dvou složek,

přičemž již 3 složky poskytly poměrně velmi dobrou aproximaci pro příjmy v letech 2004–2008.

Pokud byl použit model se známými příslušnostmi ke složkám, byla získána užitečná informace o jednotlivých podskupinách domácností odrážející vliv vzdělání či počtu dětí, přístup však nevedl k dobrému modelu příjmu mezi všemi domácnostmi (nedošlo k výraznému zlepšení Akaikeho kritéria proti modelu s jednou složkou, odhadnutá hustota velmi špatně vystihuje empirické rozložení pravděpodobností /graf není obsažen v textu/).

Oproti tomu hledání umělých komponent tak, aby výsledný model co nejlépe vystihoval data, vedlo k velmi dobrému vystižení rozdělení pozorovaných hodnot. V tomto případě ale docházelo (zvláště pro počty skupin 4 a 5) k numerickým problémům s odhadem a k problémům s identifikací parametrů. Obecně při použití konečných směsí s mnoha složkami dochází k velmi obtížné identifikaci parametrů (složek).

Stejně jako u jiných problémů parametrických modelů je důležitá volba rozdělení. Je to základní problém zvláště tehdy, pokud používáme předem definované dělení do složek, nebo pro model s neznámými složkami malý počet složek. Pokud bychom nebyli omezeni počtem složek, bylo by možné rozdělení aproximovat směsí například jen normálních rozdělení.

Literatura

- BARTOŠOVÁ, J. Logarithmic-Normal Model of Income Distribution in the Czech Republic. *Austrian Journal of Statistics*. 2006, vol. 35, s. 215–222.
- BARTOŠOVÁ J.; BÍNA V. Modelling of Income Distribution of Czech Households in Years 1996 – 2005. *Acta Oeconomica Pragensia*. 2009, vol. 17, no. 4, s. 3–18.
- BÍLKOVÁ, D. Application of Lognormal Curves in Modeling of Wage Distributions. *Journal of Applied Mathematics*. 2008, vol. 1, no. 2, s. 341–352.
- CZSO. Český statistický úřad. www.czso.cz.
- FLACHAIRE E.; NUNEZ O. Estimation of the Income Distribution and Detection of Subpopulations: an Explanatory Model. *Computational Statistics & Data Analysis*. 2007, vol. 51, no. 7, s. 3368–3380.
- GRÜN, B.; LEISCH, F. Flexmix version 2: Finite mixtures with concomitant variables and varying and constant parameters. *Journal of Statistical Software*. 2008, vol. 28, no. 4, s. 1–35.
- KLEIBER, C.; KOTZ, S. *Statistical Size Distributions in Economics and Actuarial Sciences*. New York : Wiley-Interscience, 2003.
- MCDONALD, J. B. Some Generalized Functions for the Size Distribution of Income. *Econometrica*. 1984, vol. 52, no. 3, s. 647–665.
- MILANOVIC, B. True World Income Distribution, 1988 and 1993: First Calculation Based on Household Surveys Alone. *The Economic Journal*. 2002, vol. 112, s. 51–92.
- PACÁKOVÁ, V.; SIPKOVÁ, L. Generalized Lambda Distributions of Household's Incomes. *E + M Ekonomie a Management*. 2007, roč. 10, č. 1, s. 98–107.
- PAVELKA, R. Application of density mixture in the probability model construction of wage distributions, In Proceedings: Applications of Mathematics and Statistics in Economy, AMSE 2009, Uherské Hradiště, 2009, s. 341–350.
- TITTERINGTON, D. M.; SMITH, A. F.; MAKOV, U. E. *Statistical analysis of finite mixture distributions*. Wiley, 1985. ISBN 9780471907633.

THE USE OF FINITE MIXTURES OF LOGNORMAL DISTRIBUTION FOR THE MODELLING OF INCOME DISTRIBUTIONS

Abstract: In the text finite mixtures of lognormal distributions are used for the modelling of net annual per capita income of the Czech households in Czech crowns in 2004–2008. All the households are divided into subgroups with observed group membership according to the attained education of the head of a household (a factor with five levels: basic, secondary, complete secondary, bachelor, magister education), the existence of children in the household (a factor with two levels: children yes or no) and the number of children (a factor with five levels: 0–3, more than 3 children). Then, models with incomplete data (unobserved component membership) were used for 2–5 components. All the estimates in the text are maximum likelihood estimates; explicit formulas exist for the complete data models; an EM algorithm for normal mixtures incorporated in the *flexmix* package in R was used for the incomplete data models. The models are compared with the use of the Akaike criterion. Maximum likelihood estimates of mixing proportions, expected values and standard deviations of the components are given and their development in the five-year period analysed is discussed. The program R was used for all the computations.

Keywords: finite mixture of distributions, income distribution, lognormal distribution

JEL Classification: C13, C51

NĚMEC, J. a kol. *Komparace politických systémů: Politické systémy Latinské Ameriky*. 1. vyd. Praha : VŠE, Nakladatelství Oeconomica, 2010. 279 s. ISBN 978-80-245-1740-7.

Publikace z řady komparací politických systémů se věnuje systémům vybraných zemí Latinské Ameriky. Vedle úvodní kapitoly, která vytyčuje základní témata dosavadního studia a politologického výzkumu Latinské Ameriky a vymezuje některé obecné pojmy a koncepty, jsou v knize obsaženy případové studie Argentiny, Bolívie, Brazílie, Chile, Kolumbie, Mexika a Venezuely. Výběr zemí není vyčerpávající, nicméně zvolené případy reprezentují určité fenomény, jejichž rozbor je klíčový pro pochopení politického dění v celém regionu. Jedním z centrálních témat publikace je otázka, zda se politické systémy těchto zemí ve svých rysech přibližují, či zda se s postupem času od sebe vzdalují. Tedy zda je možné hovořit o latinskoamerické identitě i z pohledu fungování politických systémů.

STROSSA, P. *Počítačové zpracování přirozeného jazyka*. 1. vyd. Praha : VŠE, Nakladatelství Oeconomica, 2011. 316 s. ISBN 978-80-245-1777-3.

Jednotlivé kapitoly této knihy se věnují hlavním aplikačním oblastem počítačového zpracování přirozeného jazyka (zejména textů) v přirozené posloupnosti: automatickému indexování textů, automatizaci tvorby tezaurů, automatizaci referování (tvorby abstraktů), strojovému překladu, automatickému získávání znalostí z textů, automatizované korektuře textů a komunikaci mezi člověkem a počítačem v přirozeném jazyce. Kapitoly mají různý rozsah, což odpovídá pokročilosti vědeckého poznání a technických nástrojů v jednotlivých oblastech. Čtyři přílohy detailněji ilustrují vybrané problémy a nástroje.