POUŽITÍ KONEČNÝCH SMĚSÍ LOGARITMICKO-NORMÁLNÍCH ROZDĚLENÍ PRO MODELOVÁNÍ PŘÍJMŮ ČESKÝCH DOMÁCNOSTÍ

Ivana Malá, Vysoká škola ekonomická v Praze

Úvod

Zkoumání a modelování pravděpodobnostního rozdělení mezd a příjmů z různých úhlů pohledu a různými ekonometrickými nebo statistickými metodami je dlouhodobě aktuálním tématem a problémem, kterému je věnována velká pozornost. Informace o příjmech osob a domácností jsou nedílnou součástí analýz týkajících nejširšího spektra ekonomických, demografických i dalších problémů, neboť velikost, skladba, vývoj příjmů a také očekávání jejich vývoje silně ovlivňují chování domácností. Znalost rozdělení dává o příjmech hlubší znalost, než jsou jen běžně používané charakteristiky polohy, variability, šikmosti nebo špičatosti. Tyto charakteristiky lze samozřejmě z modelového rozdělení snadno určit, je možné ale porovnávat rozdělení pro různé skupiny obyvatel definované například bydlištěm, pohlavím, vzděláním či zaměstnáním, sledovat vývoj a změny v čase, srovnávat regiony nebo státy. Příjmy, jejich velikost a veličiny od nich odvozené (jako například míra chudoby) jsou také považovány za jeden z určujících faktorů ovlivňujících kvalitu života obyvatel. Proto jsou součástí různých indexů snažících se kvalitu života v daném regionu či státu popsat. Výsledky takových zkoumání a analýz jsou pak využívány nejen odborníky a analytiky, ale příjmy a hlavně jejich výše jsou objektem zájmu nejširší veřejnosti.

Velký zájem o příjmy a jejich rozdělení dokumentuje také bohatá literatura týkající se nejrůznějších přístupů (obsáhlý přehled McDonald, 1984 nebo Kleiber, Kotz, 2003). Jako pravděpodobnostní model pro příjmy (nebo mzdy) jsou v obsáhlé literatuře používána různá pravděpodobnostní rozdělení, která se také někdy nazývají příjmová. Logaritmicko-normálního rozdělení, použité také v tomto textu, bylo pro příjmy a platy v České republice použito například v Bartošová (2009), Bartošová, Bína (2009) nebo Bílková (2012). Mezi často a s úspěchem používaná příjmová rozdělení patří také Dagumovo rozdělení (Kleiber, 2007; Dagum, 2008; pro příjmy v České republice Malá, 2011), zobecněné lambda rozdělení (Pacáková, Sipková, 2007 pro Slovensko), zobecněné beta (McDonald, Xu, 1995), Paretovo rozdělení pro velké příjmy a další pravděpodobnostní rozdělení (Kleiber, Kotz, 2003; McDonald, 1984; Milanovic, 2002). V práci Pittau, Zelli (2006) je místo parametrického přístupu k problému hledání vhodného pravděpodobnostního rozdělení (pomocí odhadu parametrů ve zvoleném modelu) použit neparametrický jádrový odhad hustoty pravděpodobnosti. V obsáhlé literatuře se setkáváme s podrobnými a obsáhlými analýzami velikosti

a variability příjmů v různých státech konstruovanými na základě různých zdrojů dat. V práci Prieto-Alaiz, Victoria-Feser (1996) se analýza týká Španělska, článek autorů Wu a Perloff (2005) podrobně zkoumá vývoj příjmů v Číně. Flachaire a Nunez (2007) použili data o příjmech domácností ve Velké Británii ke konstrukci modelů používajících směs pravděpodobnostních rozdělení. Práce Pittau, Zelli (2006) nebo Milanovic (2002) srovnávají příjmy mezi státy či regiony. Morley (1981) použil data týkající se vývoje příjmů v Brazílii ke zkoumání vlivu změny struktury (a velikosti) populace na změnu příjmů a také odlišnou dynamiku vývoje pro různé skupiny obyvatel.

Cílem tohoto textu je provést analýzu ročních příjmů domácností v České republice v letech 2005–2010 pomocí modelů směsi logaritmicko-normálních rozdělení se dvěma až čtyřmi složkami. Použití směsi rozdělení umožňuje hledat v příjmově velmi nehomogenní populaci českých domácností podmnožiny s podobnými příjmy. Vzhledem k tomu, že jsou použita data o příjmech za šest let, je v článku sledován také vývoj velikosti homogenních skupin domácností podle příjmů. Všechny výpočty byly provedeny pro příjem na jednu osobu a pro příjem na spotřební jednotku určený podle metodiky Evropské unie (EU). Je proto možné porovnat rozdíly vývoje podle toho, jaký příjem je sledován. S různými výpočty průměrného příjmu domácnosti také souvisí také zkoumání vývoje velikosti českých domácností posuzované počtem členů či počtem spotřebních jednotek. Giniho koeficient je použit pro posouzení příjmové nerovnosti v populaci domácností, stejně jako v jejích částech.

V dalším textu je použit pro modelování rozdělení příjmů v České republice model směsi dvou až čtyř složek s logaritmicko-normálním rozdělením (bez informace o příslušnosti domácností ke složce). V takovém případě jsou složky konstruovány tak, aby výsledný model co nejlépe vystihoval empirické hodnoty. Není zřejmá interpretace složek a také tento postup nedává jednoznačně příslušnost jednotlivých domácností ke složkám směsi, je ovšem možné takové pravděpodobnosti pro každou domácnost odhadnout. V tomto textu se budeme zabývat pouze odhady parametrů a charakteristik rozdělení, nikoliv odhady příslušnosti domácností ke složkám. Jde tedy o úlohu nalezení co nejvýstižnějších složek a odhad jejich parametrů, nikoliv o shlukovou nebo diskriminační analýzu třídící domácnosti do homogenních skupin podle velikosti příjmu (Hebák a kol., 2007). Nebyly také použity žádné vysvětlující proměnné pro pravděpodobnosti příslušnosti domácností do složek. Pracujeme pouze s čistými ročními příjmy na jednoho člena domácnosti a na jednu spotřební jednotku.

1. Konečné směsi pravděpodobnostních rozdělení

Podrobné informace o českých domácnostech poskytuje šetření Životní podmínky (jedná se národní modul šetření s názvem European Union – Statistics on Income and Living Conditions, EU-SILC). Data z tohoto šetření provedeného v letech 2005 až 2011 Českým statistickým úřadem obsahují podrobné údaje o příjmech českých domácností v letech 2004 až 2010, které jsou použity v tomto textu. Kromě mnoha znaků domácností (CZSO) obsahují soubory dat celkový roční čistý příjem domácností a informaci

o počtu členů a jejich složení. Pro porovnání domácností lze použít tento celkový příjem anebo příjem na jednoho člena domácnosti, jinak též příjem na osobu nebo per capita. Další možností je sledovat příjmy na spotřební neboli ekvivalentní jednotku, tento postup zohledňuje nejen velikost, ale také demografické složení domácnosti. Výpočet těchto jednotek je konstruován tak, aby odrážel úspory plynoucí ze sdílení domácnosti více osobami, tedy úspory na nákladech na předměty a služby, které slouží většímu počtu členů domácnosti (domácí spotřebiče, elektřina apod.). Standardně se používají dvě stupnice spotřebních jednotek. Stupnice Organizace pro hospodářskou spolupráci a rozvoj (dále OECD) přiřazuje první dospělé osobě v domácnosti váhu 1,0, dalším osobám starším 13 let váhu 0,7 a dětem do 13 let včetně váhu 0,5. Ve stupnici EU (dále *ej*) jsou více zohledněny úspory ze sdílení výdajů, váhy jsou definovány jako 1,0 pro první dospělou osobu, 0,5 pro další osoby starší 13 let věku a váhu 0,3 pro všechny děti mladší 13 let věku (CZSO). Celkový příjem domácnosti vztažený na spotřební jednotku podle metodiky Evropské Unie budeme nazývat ekvivalizovaným příjmem. Podle předchozího je zřejmé, že největší hodnotu má počet osob, pak počet spotřebních jednotek podle OECD a nejmenší je hodnota počtu spotřebních jednotek podle metodiky EU. Pro přepočítané příjmy je nerovnost obrácená, nejvyšší je příjem přepočtený na spotřební jednotku podle EU, menší je pro příjem na jednu spotřební jednotku podle metodiky OECD a nejmenší je příjem na jednu osobu. Rovnosti je dosaženo pro jednočlenné domácnosti.

Vzhledem k tomu, že domácnosti v České republice netvoří homogenní soubor, použijeme v tomto textu pro rozdělení příjmů na osobu a ekvivalizovaného příjmu rozdělení směsi pravděpodobnostních rozdělení se dvěma, třemi a čtyřmi složkami, které budou mít logaritmicko-normální rozdělení se dvěma parametry (McLachlan, Peel, 2000). Pro všechny složky tedy předpokládáme stejné pravděpodobnostní rozdělení a jednotlivé složky se liší pouze parametry. Pro popis příjmového rozdělení se obyčejně pouze dvouparametrické logaritmicko-normální rozdělení nepoužívá, v případě více složek dochází ovšem ke zřetelnému zlepšení již při několika málo složkách a rozdělení lze brát pouze dvouparametrické. Budeme tedy předpokládat, že hustoty pravděpodobnosti ekvivalizovaného příjmu (budeme používat značení s indexy *ej*) $f_{ej}(x;\boldsymbol{\psi}_{ej})$ lze zapsat jako

$$f_{ej}(x;\boldsymbol{\psi}_{ej}) = \sum_{j=1}^{K} \pi_j^{ej} f(x;\mu_j^{ej},\sigma_j^{ej2}) = \sum_{j=1}^{K} \pi_j^{ej} \frac{1}{x} \varphi\left(\frac{\ln(x) - \mu_j^{ej}}{\sigma_j^{ej}} \right), x \in R, \tag{1}$$

kde $f(x;\mu_j^{ej},\sigma_j^{ej2})$ je hustota logaritmicko-normálního rozdělení s parametry ($j = 1, ..., K$), μ_j^{ej} a σ_j^{ej2}, φ je hustota normovaného normálního rozdělení a $\pi_j^{ej}, j = 1, ..., K$ jsou váhy složek směsi splňující podmínky

$$\sum_{j=1}^{K} \pi_j^{ej} = 1,\ 0 \le \pi_j^{ej} \le 1, j = 1, ..., K.$$

Vektor $\boldsymbol{\psi}_{ej}$ obsahuje neznámé parametry, tedy K parametrů μ_j^{ej}, K parametrů σ_j^{ej2}a $(K - 1)$ volných parametrů π_j^{ej}. Budeme uvažovat počet složek K rovný 1 (logarit-

micko-normální rozdělení), 2, 3 a 4. Budeme tedy hustotu zkoumaného příjmu hledat jako vážený průměr hustot jednotlivých složek. Obdobně označíme hustotu příjmů přepočítaných na jednu osobu (per capita) $F_{pc}(x;\boldsymbol{\psi}_{pc})$ a všechny další funkce a parametry obdobně indexem *pc*.

V dalších úvahách budou využity známé vztahy pro logaritmicko-normální rozdělení a jeho vztah k rozdělení normálnímu, neboť náhodná veličina má logaritmicko-normální rozdělení právě když její logaritmus má normální rozdělení a navíc parametry logaritmicko-normálního rozdělení jsou střední hodnota a rozptyl normálního rozdělení logaritmů. Z definice hustoty (1) ihned plyne, že distribuční funkce $F_{ej}(x;\boldsymbol{\psi}_{ej})$ náhodné veličiny s rozdělením směsi je váženým průměrem distribučních funkcí $F(x;\mu_j^{ej},\sigma_j^{ej2})$ složek s vahami π_j^{ej}. Totéž platí i o střední hodnotě, neboť střední hodnota směsi $\mathrm{E}(X^{ej})$ je váženým průměrem středních hodnot složek $E(X_j^{ej})$ s vahami π_j^{ej}. Dostáváme tedy

$$F_{ej}(x;\boldsymbol{\psi}_{ej}) = \sum_{j=1}^{K} \pi_j^{ej} F(x;\mu_j^{ej},\sigma_j^{ej2}) = \sum_{j=1}^{K} \pi_j^{ej} \Phi\left(\frac{\ln(x)-\mu_j^{ej}}{\sigma_j^{ej}}\right), x \in \mathrm{R}, \tag{2}$$

$$E(X^{ej}) = \sum_{j=1}^{K} \pi_j^{ej} E\left(X_j^{ej}\right) = \sum_{j=1}^{K} \pi_j^{ej} e^{\mu_j^{ej}+\frac{1}{2}\sigma_j^{ej2}}, \tag{3}$$

kde F je distribuční funkce logaritmicko-normálního rozdělení a Φ je distribuční funkce normovaného normálního rozdělení. Rozptyl směsi byl určen podle vztahu

$$D(X^{ej}) = \sum_{j=1}^{K} \pi_j^{ej} E\left((X_j^{ej})^2\right) - \left(E(X^{ej})\right)^2 = \sum_{j=1}^{K} \pi_j^{ej} e^{2\mu_j^{ej}+2\sigma_j^{ej2}} - \left(\sum_{j=1}^{K} \pi_j^{ej} e^{\mu_j^{ej}+\frac{1}{2}\sigma_j^{ej2}}\right)^2. \tag{4}$$

$100P\%$ kvantil x_P^{ej} $(0 < \mathrm{p} < 1)$ rozdělení ekvivalizovaného příjmu popsaného směsí je třeba počítat z definice kvantilu jako řešení rovnice (použijeme (1), (2))

$$F_{ej}(x_P^{ej};\boldsymbol{\psi}_{ej}) = \sum_{j=1}^{K} \pi_j F(x_P^{ej};\mu_j^{ej},\sigma_j^{ej2}) = \sum_{j=1}^{K} \pi_j^{ej} \Phi\left(\frac{\ln(x_P^{ej})-\mu_j^{ej}}{\sigma_j^{ej}}\right) = P. \tag{5}$$

Tato rovnice nemá explicitní řešení, vážený průměr odpovídajících skupinových kvantilů $x_{j,P}^{ej}$, $j = 1, ..., K$ jednotlivých složek směsi může sloužit jako vhodná počáteční hodnota pro aproximační numerický proces.

Pro výpočet Giniho koeficientu pro jednotlivé složky byl využit vztah pro Giniho koeficient logaritmicko-normálního rozdělení s parametry μ (na tomto parametru koeficient nezávisí) a σ^2, $2\Phi(\sigma/\sqrt{2}) - 1$ podle Kleiber, Kotz (2003). V případě směsí byl Giniho koeficient určen jako (Young, 2011)

$$G = \sum_{i=1}^{K}\sum_{j=1}^{K} \frac{\pi_i^{ej}\pi_j^{ej}E(X_i^{ej})}{E(X^{ej})}\left(2\Phi\left(\frac{\left(\ln E(X_i^{ej}) - \ln E(X_j^{ej}) + 0{,}5\sigma_i^{ej2} + 0{,}5\sigma_j^{ej2}\right)}{\sqrt{\sigma_i^{ej2}+\sigma_j^{ej2}}}\right) - 1\right). \tag{6}$$

Vzorec obsahuje střední hodnoty jednotlivých složek směsi (tyto hodnoty podle (3) závisí na všech parametrech) a Giniho koeficient směsi tedy nelze určit pouze pomocí směrodatných odchylek jednotlivých složek.

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

$$L(\boldsymbol{\psi}^{ej}) = \prod_{i=1}^{n} f_{ej}(x_i; \boldsymbol{\psi}^{ej}) = \prod_{i=1}^{n} \left(\sum_{j=1}^{K} \pi_j^{ej} f(x_i; \mu_j^{ej}, \sigma_j^{ej\,2}) \right). \tag{7}$$

Logaritmická věrohodnostní funkce (logaritmus (7)) má tvar

$$l(\boldsymbol{\psi}^{ej}) = \ln\left(\prod_{i=1}^{n} \left(\sum_{j=1}^{K} \pi_j^{ej} f(x_i; \mu_j^{ej}, \sigma_j^{ej\,2}) \right) \right) = \sum_{i=1}^{n} \ln\left(\sum_{j=1}^{K} \pi_j^{ej} f(x_i; \mu_j^{ej}, \sigma_j^{ej\,2}) \right), \tag{8}$$

a výraz nelze upravit tak, aby nebylo nutné logaritmovat součet. Vhodným nástrojem pro nalezení odhadů neznámých parametrů směsi je EM-algoritmus (McLachlan, Peel, 2000), 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 (vah) 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 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 $\boldsymbol{\psi}$ a lze tedy nalezené parametry považovat za řešení optimalizační úlohy.

Pro odhad parametrů modelů směsí byly použity balíčky *mixtool*s (RMIXTOOLS) a *flexmix* (RFLEXMIX) v programu *R* 2.13.1 (RPROGRAM). Tyto programy hledají odhady parametrů ve směsích normálních rozdělení, proto byly použity na hodnoty logaritmů analyzovaných příjmů. Odhady parametrů pro počet složek do 3 byly dosaženy oběma programy bez numerických problémů, čtyři složky již byly numerickým problémem a procesy i po velkém počtu iterací nekonvergovaly, nebo konvergovaly k řešením s nulovým (nulovými) rozptyly složek.

Kvalitu modelů se stejným počtem složek lze porovnat srovnáním hodnoty logaritmické věrohodnostní funkce v dosaženém maximu, modely s různým počet složek (a tedy i parametrů) pak pomocí hodnoty Akaikova informačního kritéria AIC definovaného vztahem

$$AIC = -2\ln(L(\boldsymbol{\psi})) + 2 \text{ počet parametrů modelu.} \tag{9}$$

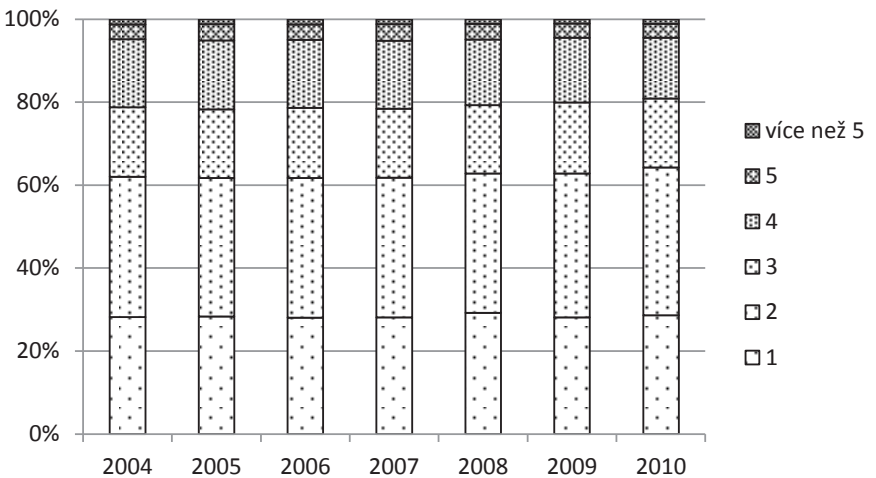
V případě jednoho dvouparametrického rozdělení je počet neznámých parametrů 2, pro dvě složky 5, tři složky 8 a čtyři složky 11 parametrů. Připomeňme, že pomocí tohoto kritéria lze porovnávat kvalitu modelů vždy v rámci jednoho roku, v našem případě 2004–2010.

2. Data a výsledky

Modely popsané v první části nyní použijeme pro popis rozdělení čistého ročního příjmu (v Kč) českých domácností v letech 2004 až 2010. Již bylo zmíněno, že 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 2011, zahrnují však příjmy z let 2004–2010. Z rozsáhlého šetření příjmů byly použity pouze informace o celkovém ročním čistém příjmu domácnosti (v Kč), počtu členů domácnosti, počtu spotřebních jednotek podle metodiky Evropské unie (*ej*) a OECD (*sj*). Příjem na jednu osobu byl určen pro každou domácnost jako podíl celkového ročního čistého příjmu domácnosti a počtu členů domácnosti. Ekvivalizované příjmy byly určeny jako podíl celkových příjmů a počtu spotřebních jednotek podle obou metodik. Dále byly využity koeficienty (váhy) poskytované pro vybrané domácnosti a umožňující odstranit vliv způsobu výběru (nejedná se o prostý náhodný výběr) a přepočítat hodnoty na celou populaci českých domácností. Rozsahy výběrů šetření SILC-EU byly postupně 4 351, 7 483, 9 675, 11 294, 9 911, 9 098 a 8 066 domácností. Odhady tedy byly pořízeny na základě velkých výběrů a studie zahrnuje sedm let.

Na obrázku 1 je znázorněn vývoj rozložení počtu členů českých domácností. Ve všech sledovaných letech byla modální hodnotou domácnost se dvěma členy, se zastoupením mezi 33 a 35 procenty domácností.

Obrázek 1
Rozdělení počtu členů domácnosti v letech 2004–1010

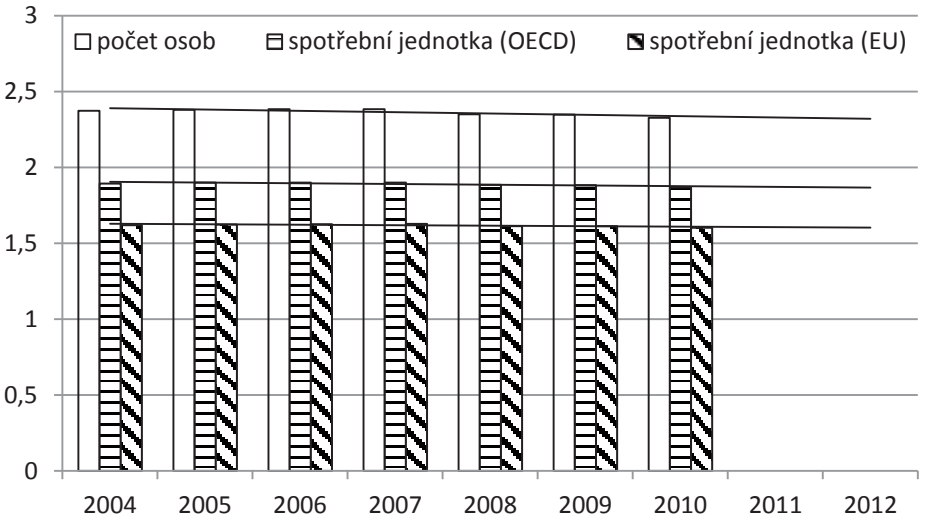


Zdroj: vlastní výpočty

Na dalším obrázku (obrázek 2) je znázorněn vývoj průměrného počtu členů domácnosti a průměrného počtu spotřebních jednotek podle metodik EU a OECD.

Z grafu je patrné již zmíněné seřazení všech tří zkoumaných charakteristik velikosti domácnosti. Ve všech třech charakteristikách dochází k pomalému poklesu velikosti domácnosti (kvantifikované počtem členů či počtem spotřebních jednotek), regresní přímky jsou znázorněny jen jako orientační, byly sestrojeny z určených průměrů, nikoliv z panelových dat výběrů a mohou znázorňovat (a porovnat) rychlost zmenšování domácností. Za sledované období došlo k poklesu průměrného počtu osob v jedné domácnosti ze 2,37 na 2,26, počet spotřebních jednotek na jednu domácnost podle metodiky EU poklesl z 1,62 na 1,60 a počet spotřebních jednotek podle metodiky OECD z 1,90 na 1,87.

Obrázek 2
Průměrné počty členů domácnosti, počtu spotřebních jednotek podle metodik EU a OECD v letech 2004–2010



Zdroj: vlastní výpočty

V dalším textu budeme předpokládat, že příjmy mají přibližně logaritmicko-normální rozdělení. Pro posouzení lineární závislosti mezi příjmy přepočítanými na osobu a na spotřební jednotku podle EU a OECD proto použijeme korelační koeficient na logaritmy příjmů. Výsledky jsou ve sledovaných letech přibližně stálé, korelace mezi příjmem na jednoho člena a příjmem na spotřební jednotku je rovna přibližně 0,86, mezi příjmem na jednoho člena a na jednotku OECD 0,95 a nejsilněji jsou vázány ekvivalizované příjmy na jednotku podle metodik EU a OECD s korelačním koeficientem přibližně 0,97.

Ke konstrukci složek nebudeme používat známé vysvětlující proměnné (jako například v Malá (2012) vzdělání osoby v čele domácnosti nebo počet dětí v domácnosti), ale budeme je tvořit tak, aby výsledný model co nejlépe vystihoval pozorovaná data. V prvním případě je možné problém rozložit na odhad v jednotlivých (známých) složkách a výsledné rozložení je pak směsí odhadnutých rozdělení s vahami, jejichž

maximálně věrohodným odhadem je relativní četnost prvků složek ve výběru. Dostáváme tedy také informaci o jednotlivých složkách a výsledky lze interpretovat se znalostí vzniku složek.

V případě neznámých příslušností ke složkám je odhad i interpretace výsledků složitější, dostáváme ale lepší odhad hledaného rozdělení. Pro předkládanou analýzu nebyly využity pro třídění do složek žádné další vysvětlující proměnné a cílem bylo pouze odhadnout rozdělení směsi, nikoliv třídit jednotlivé domácnosti do složek a odhadovat pravděpodobnosti příslušnosti jednotlivých domácností do složek. V následujících tabulkách jsou uvedeny jednotlivé složky v pořadí podle odpovídající střední hodnoty rozdělení. Vzhledem k tomu, že střední hodnota logaritmicko-normálního rozdělení závisí na obou parametrech, je toto pořadí jiné, než jaké by bylo dosaženo porovnáním středních hodnot logaritmů příjmů (odhadnutých parametrů μ^{ej} a μ^{pc}) Všechny údaje jsou v nominálních cenách. V tabulkách je proto v posledním řádku uveden index změny za sledované období. Inflace v České republice ve sledovaných letech byla celkem 17,0 procenta (CZSO), meziroční pak postupně 1,9, 2,5, 2,8, 6,3, 1,0 a 1,5 procenta.

Nejprve se budeme zabývat pouze celkovým rozdělením sledovaných příjmů. V tabulce 1 jsou uvedeny výběrové hodnoty charakterizující polohu (aritmetický průměr a medián) a variabilitu (výběrová směrodatná odchylka a Giniho koeficient) pro příjmy na jednoho člena a v tabulce 2 hodnoty získané pro ekvivalizovaný příjem podle metodiky EU.

Tabulka 1
Porovnání výběrových charakteristik polohy a variability a odhadů na základě logaritmicko-normálního rozdělení, příjem na jednoho člena domácnosti (hodnoty kromě Giniho koeficientu v Kč)

rok	Výběrové hodnoty				Maximálně věrohodný odhad			
	průměr	medián	směrodatná odchylka	Gini	střední hodnota	medián	Gini	směrodatná odchylka
2004	111 023	97 050	77 676	0,255	109 779	99 042	0,252	52 483
2005	114 945	100 640	74 502	0,233	113 728	102 971	0,247	53 327
2006	123 806	146 548	74 578	0,250	122 716	111 614	0,242	56 080
2007	132 877	117 497	73 982	0,243	132 015	120 789	0,234	58 225
2008	145 277	126 595	93 397	0,237	143 638	131 479	0,234	63 187
2009	150 853	132 794	88 171	0,240	143 638	136 244	0,241	68 233
2010	154 159	134 815	89 651	0,245	153 023	138 899	0,244	70 742
Index	1,389	1,389	1,154		1,394	1,402		1,348

Zdroj: vlastní výpočty

Již bylo uvedeno, že úroveň ekvivalizovaných příjmů je vyšší než úroveň příjmů na jednoho člena domácnosti a toto platí také pro směrodatnou odchylku, která kvantifikuje absolutní variabilitu. Variabilitu příjmů přepočítaných podle obou přístupů lze porovnat pomocí relativní variability (určené jako podíl směrodatné odchylky

a průměru). Variační koeficienty (neuvedené v tabulkách) ukazují větší relativní variabilitu pro příjmy na jednu osobu o 1–6 procentních bodů. Pokud budeme uvažovat Giniho koeficient jako míru nerovnosti, pohybují se hodnoty od 0,230 do 0,255 a není patrný jednoznačný vývoj. Odhady hodnot těchto charakteristik (určené z jednoho logaritmicko-normálního rozdělení) ukazují jasně slabiny použití logaitmicko-normálního rozdělení pro popis rozdělení příjmů. Rozdělení dobře vystihuje charakteristiky polohy a Giniho koeficient, v případě směrodatné odchylky však dochází ke značnému podhodnocení. Tento problém je odstraněn použitím více logaritmicko-normálních hustot ve směsi.

Tabulka 2
Porovnání výběrových charakteristik polohy a variability a odhadů na základě logaritmicko-normálního rozdělení, příjem na spotřební jednotku podle metodiky EU (hodnoty kromě Giniho koeficientu v Kč)

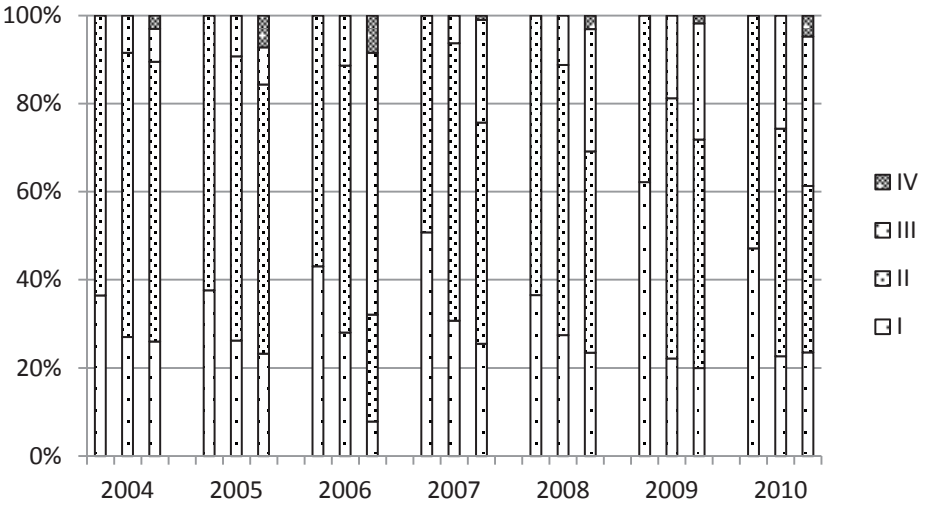
rok	Výběrové hodnoty				Maximálně věrohodný odhad			
	průměr	medián	směrodatná odchylka	Gini	střední hodnota	medián	Gini	směrodatná odchylka
2004	148 261	127 500	94 052	0,251	146 611	133 306	0,242	67 118
2005	153 412	132 613	92 829	0,250	151 721	138 649	0,236	67 419
2006	165 498	143 548	93 689	0,245	163 991	150 197	0,233	71 879
2007	178 097	156 267	96 166	0,240	176 882	162 348	0,230	76 502
2008	170 510	147 658	104 914	0,234	168 513	155 344	0,225	70 840
2009	201 454	176 273	116 977	0,230	168 513	182 536	0,235	88 387
2010	204 607	178 969	112 484	0,240	203 015	185 350	0,237	90 723
Index	1,842	1,844	1,448		1,849	1,871		1,728

Zdroj: vlastní výpočty

V dalším textu budeme uvažovat pro všechny roky a oba přepočítané příjmy modely se dvěma, třemi a čtyřmi složkami. Hodnoty Akaikova kritéria určeného podle (8) nejsou uvedeny v textu, ale výsledek dává pro všechny roky a oba příjmy stále zmenšování až do modelu se čtyřmi složkami. Pokles je značný mezi hodnotou pro jednu a dvě složky, neboť již použití dvou složek výrazně vylepšuje kvalitu modelu. Velikost poklesu hodnoty Akaikova kritéria se zmenšuje s počtem použitých složek tak, jak zlepšení modelu je stále více vyváženo zvětšením počtu odhadovaných parametrů (vždy 3 parametry na jednu přidanou složku). Vzhledem k velkému počtu pozorování ve výběrech, na jejichž základě byly parametry modelů odhadovány, všechny testy dobré shody modelů byly statisticky významné (P<0,05).

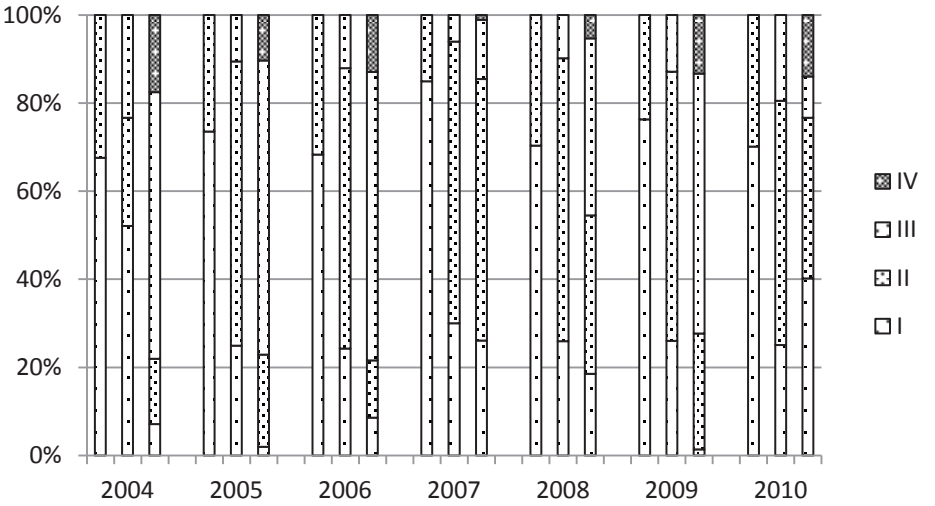
Na obrázcích 3 a 4 jsou znázorněny odhadnuté váhy (pravděpodobnosti) odhadnutých složek, od složky s nejmenší střední hodnotou (vždy I. složka) až do složky s největší odhadnutou střední hodnotou (složky II. až IV. podle počtu složek). Obrázek 3 obsahuje výsledky pro příjmy na jednu osobu a obrázek 4 na spotřební jednotku.

Obrázek 3
Odhadnuté váhy pro příjmy na jednu osobu, modely se 2, 3 a 4 složkami



Zdroj: vlastní výpočty

Obrázek 4
Odhadnuté váhy pro příjmy na spotřební jednotku, modely se 2, 3 a 4 složkami



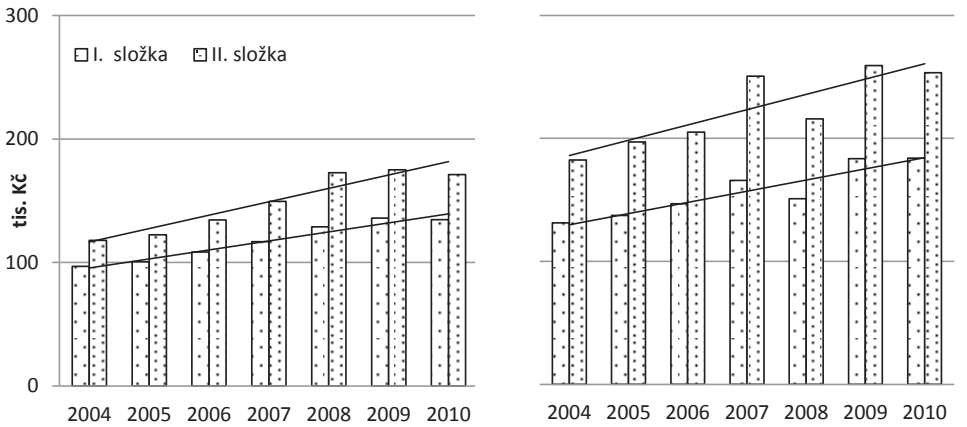
Zdroj: vlastní výpočty

Jak již bylo zmíněno, modely se třemi a méně složkami vykazují v čase poměrně stabilitu, modely se čtyřmi složkami již takové nejsou. Tyto vlastnosti jsou patrné ve všech dále prezentovaných výsledcích.

V případě dvou složek se soubor dělí na dvě části, pro příjem na jednoho člena je menší složka domácností s nižšími příjmy (30 % až 50 %, s výjimkou 62 % v roce 2009), pro ekvivalizovaný průměr je naopak menší složka s vyššími příjmy (23 % až 32 % s výjimkou pouze 15 % v roce 2007). Pro model se čtyřmi složkami pro příjmy na jednoho člena má čtvrtá složka nejvyšších příjmů zastoupení 3 % až 8 %. První složka nízkopříjmových domácností obsahuje kolem čtvrtiny domácností (s výjimkou 8 procent v roce 2006). Pro ekvivalizované příjmy je rozložení složek v jednotlivých letech (obrázek 4) velmi proměnlivé. Odhadnuté pravděpodobnosti pro třísložkové modely jsou uvedeny v tabulce 3.

Odhadnuté střední hodnoty složek pro modely (a jejich vývoj v čase) jsou ukázány na obrázcích 5–7 pro 2, 3 a 4 složky a příjem na jednu osobu (levý obrázek) a na spotřební jednotku (pravý obrázek). Sloupcové grafy jsou doplněny trendovou přímkou proloženou odhadnutými hodnotami. Tato přímka naznačuje předpověď vývoje na další období, pokud by přímka byla protažena.

Obrázek 5
Odhadnuté střední hodnoty příjmů pro modely se dvěma složkami, příjem na jednu osobu (vlevo) a na spotřební jednotku (vpravo).



Zdroj: vlastní výpočty

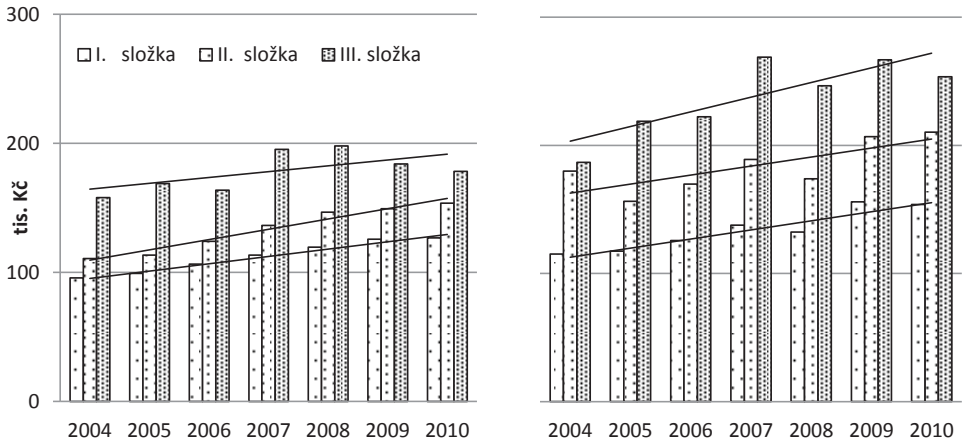
Všimněme si, že pro model s dvěma složkami dochází v letech 2004–2009 k růstu střední hodnoty v obou složkách, v posledním sledovaném roce pak je vidět pokles středních hodnot (pro příjem na jednu osobu v první složce ze 135 956 Kč na 134 550 Kč, ve druhé ze 175 106 Kč na 171 238 Kč), pro příjem na spotřební jednotku pouze ve druhé složce z 259 100 Kč na 253 350 Kč. Došlo však k růstu celkové střední hodnoty příjmu na jednoho člena ze 150 729 Kč na 153 919 Kč, neboť se změnila struktura složek (obrázky 3 a 4) a odhadnuté váhy z 0,62 a 0,38 na 0,53 a 0,47. Obdobně pro příjem na spotřební jednotku vzrostl z 201 348 Kč na 204 613 Kč změnou vah 0,77 a 0,23 na 0,7 a 0,3. Znamená to, že ve skupině domácností s vyššími příjmy bylo odhadnuto o 7 procent více domácností a tedy do skupiny domácností s menšími příjmy o 7 procent méně.

V případě tří složek (obrázek 6) nastává pokles středních hodnot složek již od roku 2008, je však opět vyvážen změnou rozložení pravděpodobností a zvýšením procenta domácností ve složkách s vyššími příjmy. Růst popsaný přímkou vykazuje o trochu rychlejší růst prostřední složky, pomalejší byl růst příjmů složky nízkopříjmových domácností a nejpomaleji rostly střední hodnoty ve skupině vysokých příjmů.

V případě ekvivalizovaných příjmů (obrázek vpravo) dochází ke kolísání středních hodnot. Růst středních hodnot dvou složek s nižšími příjmy je v tomto případě lineární s téměř stejnými směrnicemi.

Obrázek 6

Odhadnuté střední hodnoty příjmů pro modely se třemi složkami, příjem na jednu osobu (vlevo) a na spotřební jednotku (vpravo).

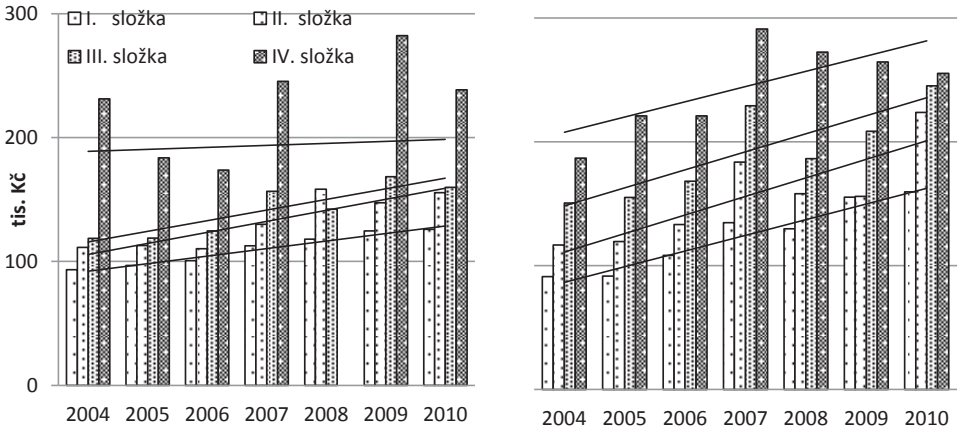


Zdroj: vlastní výpočty

Na obrázku 7 jsou znázorněny odhadnuté střední hodnoty v modelu se čtyřmi složkami. Tyto modely již nevykazují jasný vývoj. Proto, i když (jak bylo zmíněno) dochází pro čtyři složky ještě k poklesu hodnoty Akaikova kritéria, je patrně lepší omezit se na model se třemi složkami definujícími složky domácností s příjmy nízkými, středními a vysokými. Obrázky 3 a 4 doplníme tabulkou pravděpodobností pro tyto modely (tabulka 3).

Obrázek 7

Odhadnuté střední hodnoty příjmů pro modely se čtyřmi složkami, příjem na jednu osobu (vlevo) a na spotřební jednotku (vpravo).



Zdroj: vlastní výpočty

Tabulka 3

Odhadnutá procentní zastoupení složek (%) pro modely se třemi složkami.

rok	2004	2005	2006	2007	2008	2009	2010
složka	příjmy na osobu						
I.	27,0	26,23	28,1	30,7	27,5	22,1	22,7
II.	64,5	64,48	60,6	63,0	61,3	59,1	51,6
III.	8,5	9,27	11,4	6,3	11,2	18,8	25,7
	příjmy na spotřební jednotku						
I.	52,1	24,91	24,3	30,0	26,0	26,0	25,1
II.	24,5	64,47	63,6	63,9	64,2	61,1	55,5
III.	23,3	10,60	12,1	6,1	9,8	12,9	19,5

Pro příjmy na jednu osobu kolísá procento domácností ve složce s nízkými příjmy v intervalu 22,1–30,7 procent, ve složce středních příjmů v intervalu 51,6–64,5 procent a procento domácností s vyššími příjmy je od 6,3 procenta v roce 2007 do 25,7 procent v roce 2010. Pro příjmy na spotřební jednotku se vymyká rok 2005, dále pak jsou rozložení velmi podobná a znamenají kolem 25 procent ve složce nižších příjmů, kolem 60 procent ve složce středních příjmů a zbylých přibližně 10 procent ve složce vyšších příjmů. Rok 2010 se ukazuje být opět rozdílný a teprve další data týkající se let 2011 a 2012 ukáží, jak bude vývoj pokračovat.

Na závěr popíšeme rozdílnost příjmů v jednotlivých složkách (variabilitu nebo nerovnost příjmů). V tabulkách 4 a 5 je patrný rozdíl v modelech se dvěma složkami mezi Giniho indexem v jednotlivých složkách. Index je malý v první a velký ve druhé

složce, toto platí také pro směrodatné odchylky (v textu neuvedeno). Složky domácností s nízkými příjmy jsou tedy daleko homogennější než složky domácností s příjmy vyššími. Obdobně je to pro tři složky, neplatí to však vždycky, výjimkou je pro oba typy příjmů rok 2004.

Tabulka 4
Odhad Giniho koeficientu pro složky nalezené pro příjem na jednu osobu

	K=2			*K*=3				*K*=4				
	I.	II.	celkem	I.	II.	III.	celkem	I.	II.	III.	IV	celkem
2004	0,090	0,305	0,251	0,453	0,073	0,257	0,253	0,090	0,068	0,281	0,504	0,255
2005	0,090	0,305	0,246	0,073	0,257	0,453	0,250	0,068	0,281	0,090	0,504	0,250
2006	0,089	0,303	0,242	0,069	0,246	0,447	0,242	0,063	0,259	0,094	0,459	0,243
2007	0,099	0,305	0,236	0,075	0,236	0,430	0,235	0,044	0,091	0,249	0,451	0,234
2008	0,113	0,308	0,237	0,082	0,238	0,486	0,236	0,075	0,203	0,332	0,657	0,238
2009	0,135	0,334	0,242	0,081	0,213	0,443	0,240	0,076	0,179	0,307	0,546	0,241
2010	0,138	0,344	0,246	0,076	0,200	0,406	0,243	0,073	0,183	0,345	0,565	0,245

Zdroj: vlastní výpočty

Tabulka 5
Odhad Giniho koeficientu pro složky nalezené pro příjem na spotřební jednotku

	K=2			*K*=3				*K*=4				
	I.	II.	celkem	I.	II.	III.	celkem	I.	II.	III.	IV	celkem
2004	0,169	0,334	0,249	0,127	0,365	0,140	0,248	0,066	0,069	0,202	0,389	0,248
2005	0,173	0,338	0,244	0,108	0,214	0,417	0,244	0,045	0,096	0,214	0,418	0,244
2006	0,163	0,321	0,240	0,098	0,208	0,404	0,240	0,072	0,067	0,204	0,398	0,240
2007	0,182	0,385	0,236	0,114	0,212	0,484	0,236	0,109	0,196	0,344	0,630	0,235
2008	0,142	0,331	0,231	0,076	0,197	0,443	0,230	0,065	0,148	0,242	0,501	0,230
2009	0,170	0,359	0,241	0,106	0,203	0,418	0,240	0,020	0,114	0,203	0,415	0,241
2010	0,167	0,336	0,243	0,108	0,191	0,376	0,242	0,118	0,232	0,100	0,397	0,242

Zdroj: vlastní výpočty

Již bylo zmíněno, že model se čtyřmi složkami je složitější a také interpretace výsledků není jasná. Nepodařilo se nalézt v datových souborech složky snadno interpretovatelné pomocí výše příjmů. Byly nalezeny například složky se skoro stejnými hodnotami parametru μ a velice rozdílnými parametry σ, algoritmus měl snahu zmenšit počet složek nebo nebylo dosaženo řešení ani při „rozumných“ počátečních podmínkách (například dosažené aproximaci při jiném pokusu) a velkém množství (řádově tisících) iterací. Časová náročnost těchto výpočtů nebyla zanedbatelná, jednalo se přibližně o minuty až desítky minut.

Dále je vidět, že Giniho koeficienty určené ze směsí v tabulkách 4 a 5 (podle Young, 2011) jsou srovnatelné s ostatními modely i s výběrovými hodnotami (tabulka 1).

Závěr

V textu bylo ukázáno, že model konečné směsi pravděpodobnostních rozdělení je užitečný a dobře aplikovatelný při modelování nehomogenních rozdělení příjmů. Hustotu složitého rozdělení pravděpodobnosti, 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 různá pravděpodobnostní rozdělení náhodné veličiny v jednotlivých složkách. Tato rozdělení jsou pak popsána hustotami, z kterých je 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 anebo nelze pozorovat příslušnost k některé ze složek. Nejčastěji je pro všechny složky směsi použito, tak jako v tomto textu, stejné rozdělení (a složky se liší jen volbou parametrů).

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 nepříliš vhodného) logaritmicko-normálního rozdělení již při použití dvou složek, přičemž již tři složky poskytly poměrně velmi dobrou aproximaci pro příjmy v letech 2004-2008.

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 čtyři složky anebo v textu neuvedených pěti složek) k numerickým problémům s odhadem a k problémům s identifikací parametrů. Obecně při použití konečných směsí s mnoha složkami dochází k obtížné identifikaci parametrů (složek) dokonce i ve velkých výběrech.

Stejně jako u jiných problémů parametrických modelů je důležitá volba vhodného pravděpodobnostního rozdělení. Je to základní problém zvláště tehdy, pokud používáme předem definované dělení do složek anebo model s malým počtem složek s neznámou příslušností. Pokud bychom nebyli omezeni počtem složek, bylo by možné rozdělení libovolně přesně aproximovat například jen směsí normálních rozdělení.

Literatura:

BARTOŠOVÁ, J. 2009. Analysis and Modelling of Financial Power of Czech Households. Bratislava 03. 02. 2009 – 06. 02. 2009. In *8th International Conference APLIMAT 2009*. Bratislava: Slovak University of Technology, pp. 717–722.

BARTOŠOVÁ, J.; BÍNA, V. 2009. Modelling of Income Distribution of Czech Households in Years 1996 – 2005. *Acta Oeconomica Pragensia*. Vol. 17, No. 4, pp. 3−18.

BÍLKOVÁ, D. 2012. Recent Development of the Wage and Income Distribution in the Czech Republic. *Prague Economic Papers*, Vol. 21, No. 2, pp. 233–250.

DAGUM, C. A 2008. New Model of Personal Income Distribution: Specification and Estimation, In *Modeling Income Distributions and Lorenz Curves*, Economic Studies in Equality, Social Exclusion and Well-Being, Vol. 5, pp. 3–25.

FLACHAIRE E.; NUNEZ O. 2007. Estimation of the Income Distribution and Detection of Subpopulations: an Explanatory Model. *Computational Statistics & Data* Analysis. 2007. Vol. 51, No.7, pp. 3368–3380.

HEBÁK, P.; HUSTOPECKÝ, J.; PECÁKOVÁ, I.; PLAŠIL, M.; PRŮŠA, P.; VLACH P.; SVOBODOVÁ, A. A. 2007. *Vícerozměrné statistické metody [3].* 2. vyd. Praha: Informatorium, 2007.

KLEIBER, C.; KOTZ, S. 2003. *Statistical Size Distributions in Economics and Actuarial Sciences.* New York: Wiley-Interscience, 2003.

KLEIBER, C. A. 2007. Guide to the Dagum Distributions. Working paper 23/07, Wirtschaftswissenechaftliches Zentrum (WWZ) der Univeritat Basel, Dostupné na http://wwz.unibas.ch/uploads/tx_x4epublication/23_07.pdf.

MALÁ, I. 2011. Distribution of Incomes Per Capita of the Czech Households from 2005 to 2008. Bratislava 01.02.2011 – 04.02.2011. In *Aplimat 2011 [CD-ROM].* Bratislava: Slovak University of Technology, pp. 1583–1588.

MALÁ, I. 2012. Použití konečných směsí pro modelování příjmových rozdělení. *Acta Oeconomica Pragensia*, Vol. 20, No. 4, pp. 26–39.

MCDONALD, J. B. 1984. Some Generalized Functions for the Size Distributions of Income. *Econometrica,* Vol. 52, pp. 647–663.

MCDONALD, J. B.; XU, Y. J. 1995. A Generalization of the Beta Distribution with Applications. *Journal of Econometrics*, Vol. 66, No. 6, pp. 133–152.

MCLACHLAN, G. J.; PEEL, D. 2000. *Finite Mixture Models.* New York: Wiley series in Probability and Mathematical Statistics: Applied Probability and Statistics Section.

MILANOVIC, B. 2002. True World Income Distribution, 1988 and 1993: First Calculation Based on Household Surveys Alone. *The Economic Journal*, 2002. Vol. 112, pp. 51–92.

MORLEY, S. A. 1981. The Effect of Changes in the Population on Several Measures of Income Distribution. *The American Economic Review*, Vol. 71, No. 3, pp. 285–294.

PACÁKOVÁ, V.; SIPKOVÁ, L. 2007. Generalized Lambda Distributions of Households Incomes. *E + M Ekonomie a Management*, Vol.10, No.1, pp. 98–107.

PITTAU, M. G.; ZELLI, R. 2006. Empirical Evidence of Income Dynamics across EU Regions. *Journal of Applied Econometrics*, Vol. 21, No. 5, pp. 605–628.

PRIETO-ALAIZ, M.; VICTORIA-FESER, M. P. 1996. Modelling Income Distribution in Spain: A Robust Parametric Approach. STICERD - Distributional Analysis Research Programme Papers 20, Suntory and Toyota International Centres for Economics and Related Disciplines, LSE.

YOUNG, Y. 2011. The Gini Coefficient for a Mixture of Ln-Normal Populations, Working Paper LSE, London. Dostupné na http://personal.lse.ac.uk/YoungA/MixtureGini.pdf.

WU, X.; PERLOFF, J. M. 2005. China's Income Distribution, 1985-2001. *The Review of Economics and Statistics*, Vol. 87, No. 4, pp. 763–775.

Internetové zdroje:

CZSO. Czech Statistical Office. Dostupné na www.czso.cz.

RPROGRAM. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2012. Dostupné na http://www.R-project.org/.

MIXTOOLS. Tatiana Benaglia, Didier Chauveau, David R. Hunter, Derek Young. mixtools: An R Package for Analyzing Finite Mixture Models. Journal of Statistical Software, 32(6), 1-29. 2009. Dostupné na http://www.jstatsoft.org/v32/i06/.

THE USE OF FINITE MIXTURES OF LOGNORMAL DISTRIBUTION FOR THE MODELLING OF HOUSEHOLD INCOME DISTRIBUTIONS IN THE CZECH REPUBLIC

Ivana Malá, University of Economics, Prague, W. Churchill Sq. 4, CZ – 130 67 Prague 3
(malai@vse.cz)

Abstract
In the text finite mixtures of lognormal distributions are used for the modelling of net annual income per capita and equivalized income of the Czech households (in CZK) in 2004-2010. The development of distribution of number of members of households is analysed and the characteristics of standardized units according to EU and OECD methodologies are given. Data from the survey EU-SILC organized by the Czech Statistical Office from 2005-2011 (dealing with incomes from 2004-2010) are used for the analysis. Models (with incomplete data) with two to four artificial components are used in order to fit the distribution of incomes; the development of their characteristics is shown. All estimates in the text are maximum likelihood estimates, EM algorithm in the program R is used for the optimalization. Models are compared with the use of Akaike criterion.

Keywords
finite mixture of distributions, income distribution, income inequality, Gini coefficient, EM algorithm

JEL Classification
C13, C51, O15