SHLUKOVÁNÍ PROSTŘEDNICTVÍM KONEČNÝCH SMĚSÍ

ROBUST 2014

Jetřichovice 19.–24. ledna

Jitka Bartošová

katedra exaktních metod

Vysoká škola ekonomická v Praze
Fakulta managementu

Jindřichův Hradec

Abstrakt

- Snahy o modelování velkých náhodných výběrů končí často zjištěním, že není vhodné žádné rozdělení, které by bylo známé a analyticky schůdné.
- Proto se k modelování empirického rozdělení četností používají **konečné směsi hustot,** které lze vcelku jednoduše odhadnou pomocí EM-algoritmu.
- **Jednotlivé komponenty konečných směsí**, získané například prostřednictvím procedury *mclust* (kombinaci algoritmů EM / BIC) v prostředí R, **mohou být chápány jako latentní shluky.**
- Vyvstává však otázka, **nakolik mohou být tyto komponenty skutečně považovány za shluky a do jaké míry je přiřazení objektů do shluků „objektivní“?**
- To souvisí s otázkou: **„Do jaké míry jsou tyto shluky separované?“**
- Budeme tedy hledat **pravidlo, podle kterého budeme schopni rozhodnout, zda určitá podmnožina (subpopulace), popsaná rozdělením četností, je shlukem.**

Cíle

- **prezentace jednoho z nástrojů shlukování**,
- **posouzení míry separace**,
- **aplikace** na datové soubory nesoucí informaci o finančním potenciálu domácností.

Úvod

Tento příspěvek vychází z potřeb **kvantitativního hodnocení finančního potenciálu obyvatelstva**, tj. z potřeb aplikačních, a nemá žádné teoretické ambice.

Znalost finančního potenciálu je jedním z výchozích bodů pro

- **posuzování životní úrovně, úrovně sociálního zabezpečení** a **sociální spravedlivosti**,
- **rozhodování v oblasti hospodářské politiky,**
- **plánování** v oblasti **státního rozpočtu**, především při t**vorbě daňové**, **sociální a důchodové politiky**,
- **plánování regionálního rozvoje**,
- **rozhodování o alokaci dotací** v rámci ČR i v rámci EU,
- **tvorbu legislativních úprav v oblasti sociálního, zdravotnického, důchodového a daňového systému**
- **sledování úrovně** a **struktury výdajů, nákupních úmyslů, schopnosti splácet úvěry a hypotéky**,
- **rozhodování podniků o alokaci výrobních prostředků**, **vstupu na trh** apod.

Datová základna

EU SILC (*European Union – Statistics on Income and Living Conditions*) – (reprezentativní) vzorek dat

- Rozsáhlé výběrové šetření, které je povinné pro všechny členské státy Evropské unie.
- Jednotná metodika ve všech zemích Unie zaručuje vzájemnou srovnatelnost získaných výsledků napříč EU.

Soubory obsahují:

- **průřezová data**,
- **longitudinální data**.

Pracovní definice finančního potenciálu

Aktuální (nominální, resp. reálná) hodnota celkových disponibilních příjmů (v paritě kupní síly), která zahrnuje:

- **příjmy z práce,**
- **soukromé příjmy z investic a převodů vlastnictví mezi domácnostmi,**
- **sociální transfery přijaté v hotovosti, včetně starobních důchodů.**

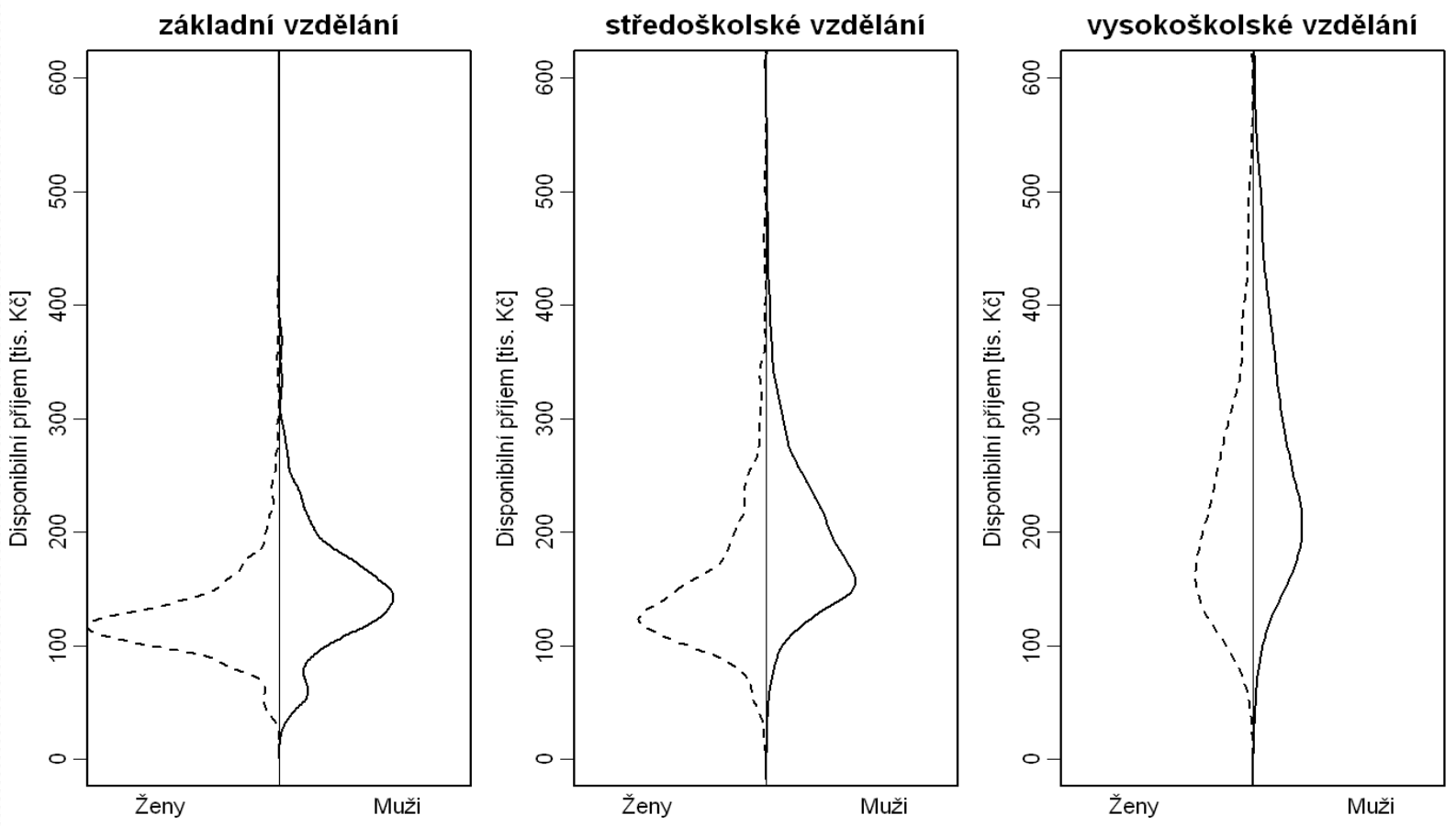
V datových souborech jsou k dispozici pouze informace týkající se:

- současného (resp. minulého) příjmu,
- struktury příjmů.

Chybí informace ohledně:

- očekávaného budoucího příjmu,
- vyčíslení majetku,
- charakteru příjmu apod.

Ukázka jádrových odhadů empirické hustoty příjmů domácností



Rozložení ekvivalizovaných hodnot ročních **příjmů** českých domácností v závislosti na

- **pohlaví**,
- **dosaženém vzdělání**.

plná čára – domácnosti v čele s mužem, *přerušovaná čára* – domácnosti v čele se ženou.

Shluková analýza

- Cílem *shlukové analýzy* je roztřídit (*klasifikovat*) objekty do navzájem disjunktních podmnožin – *shluků* (*clusters*).

Přístupy:

- **vzdálenostní / podobnostní (*distance-based clustering*) – *pevné shlukování*** – přiřazuje ke shlukům jednoznačně:
  - metoda *k-means*,
  - metody *hierarchického shlukování* (metody *nejbližších / nejvzdálenějších sousedů*)
- **modelový (*model-based clustering*) – *pravděpodobnostní* (*bayesovské*) *shlukování*** – přiřazuje pouze pravděpodobnosti příslušnosti k jednotlivým (latentním) shlukům

Modelový (bayesovský) přístup

Do této kategorie se řadí **směsové modely *(mixture models)***

- **konečné směsi hustot (*finite mixtures*)**, založené na aposteriorních pravděpodobnostech,
- **směsi regresních modelů (*mixtures of regression models*))**, založené na podmíněných aposteriorních pravděpodobnostech.

Bayesovské shlukování patří mezi *nehierarchické postupy*, které vyžadují , aby **počet shluků** (komponent směsi) byl **předem určený**. K tomu slouží:

- **expertní odhady**
- **penalizovaná informační kritéria**

Konečné směsi hustot

- Hustota vícerozměrné náhodné veličiny X je **konečnou směsí K hustot**, pokud

$$f^{(K)}\left(X;\boldsymbol{\Psi}^{(K)}\right)=\sum_{k=1}^{K}p_k^{(K)}f_k^{(K)}\left(\mathbf{x};\boldsymbol{\theta}_k^{(K)}\right)$$

kde $p_k^{(K)}>0$ pro $k=1,\ldots,K$ a $\sum_{k=1}^{K}p_k^{(K)}=1$ jsou *váhy* (*marginální pravděpodobnosti*) komponent,

$\boldsymbol{\Psi}^{(K)}=(p_1^{(K)},\ldots,p_{K-1}^{(K)},\boldsymbol{\theta}_1^{(K)},\ldots,\boldsymbol{\theta}_K^{(K)})$ je *vektor* neznámých *parametrů*.

(viz např. McLachlan a Peel, 2000)

Směsi lineárních modelů (s pevnými a náhodnými efekty)

- Konečné směsi hustot lze jednoduše rozšířit na **konečné směsi lineárních modelů (*finite mixtures of linear models – FMLM*)** nahrazením nepodmíněného rozdělení podmíněným rozdělením odezvy.
- Hustota podmíněného rozdělení odezvy *Y* je dána vztahem

$$Y|\mathbf{x} \sim f_Y\left(y|\mathbf{x}, \mathbf{z}, \boldsymbol{\Psi}^{(K)}\right) = \sum_{k=1}^{K} \pi_k^{(K)} f_k^{(K)}\left(y|\mathbf{x}, \mathbf{z}, \boldsymbol{\beta}_k, \sigma_k, \mathbf{D}_k\right)$$

kde **x** a **z** jsou *vektory prediktorů*, které se vztahují k *fixním* a *náhodným efektům*

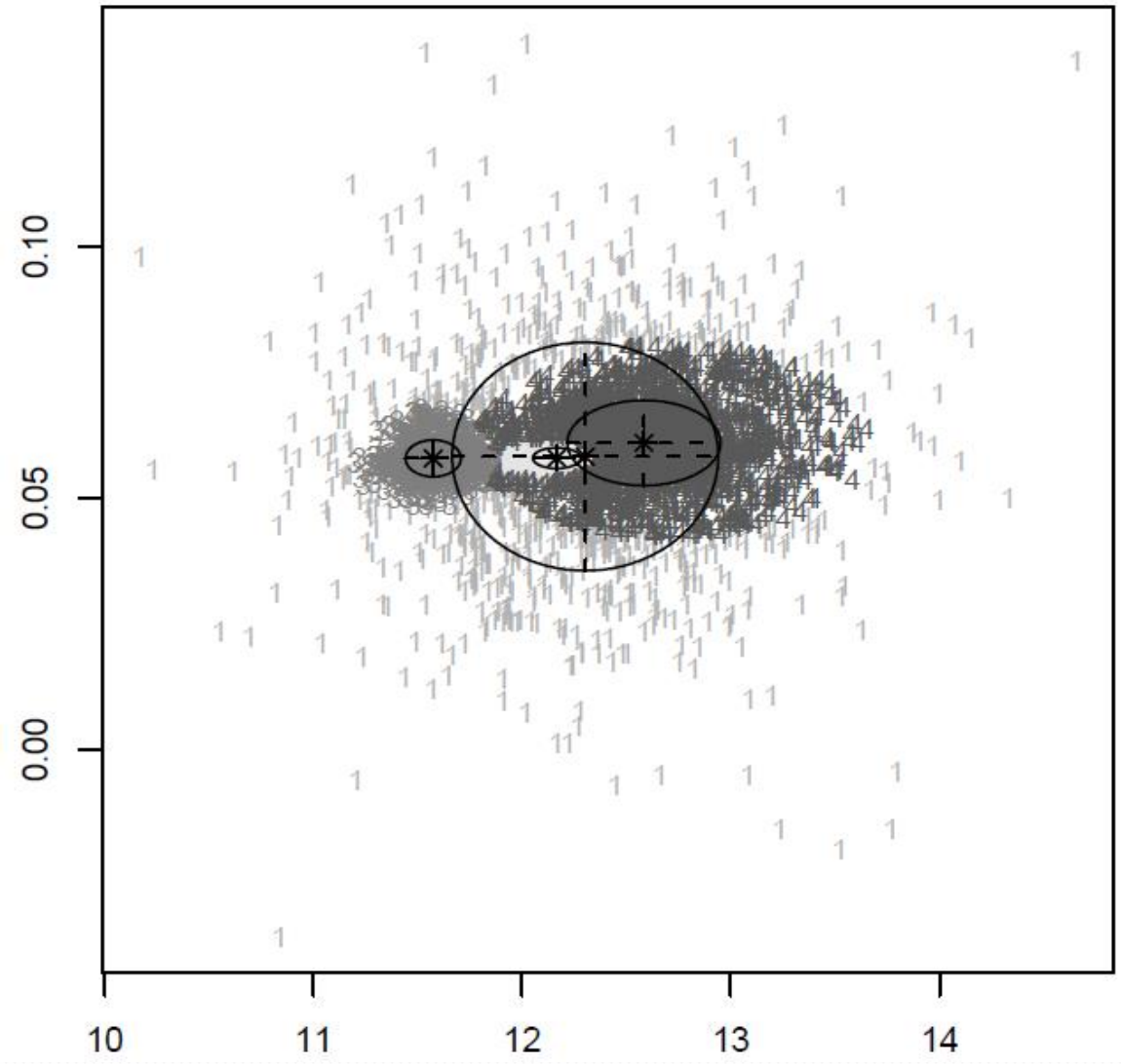
$$\boldsymbol{\Psi}^{(K)} = (\pi_1, \ldots, \pi_{K-1}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K, v_1, \ldots, v_K, \sigma_1, \ldots, \sigma_K, \mathbf{D}_1, \ldots, \mathbf{D}_K)$$

je *vektor* odhadovaných *parametrů* modelu.
(viz např. Celeux a Laverone , 2005)

EM-algoritmus

- Na principu **bayesovského shlukování** pracuje známý **EM-algoritmus** (***expectation-maximization***), který je založen na předpokladu směsového charakteru datových souborů.
- Shlukování se provádí na základě odhadu parametrů jednotlivých komponent směsi prostřednictvím iterační procedury, která nalezne **lokální maximum** příslušné **věrohodnostní funkce**. (viz Dempster, Laird a Rubin, 1977)
- Při shlukování prostřednictvím směsových modelů **nedocílíme dokonalé separace** objektů či proměnných – vzniklé shluky se částečně nebo zcela překrývají

Separace shluků náhodných parametrů v regresním modelu (odhad EM-algoritmem)

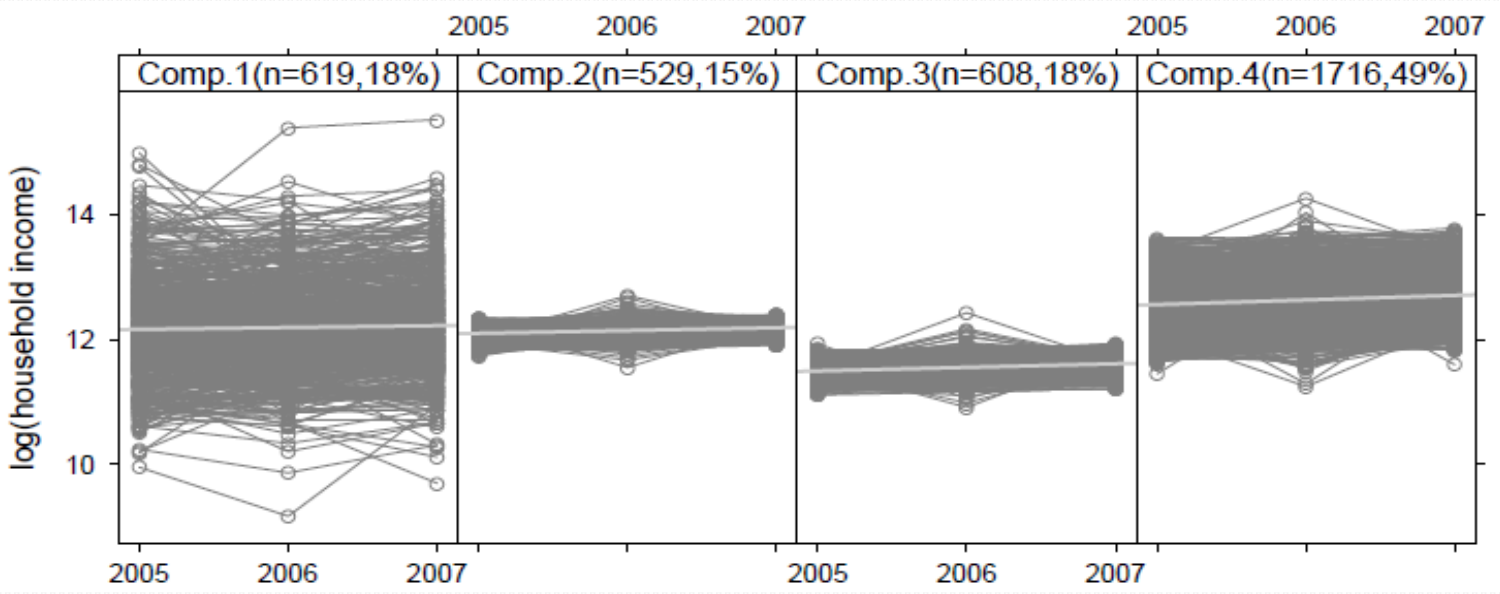


Odhad **lineárního smíšeného modelu s autoregresní náhodnou složkou (*LMM-AR)***:

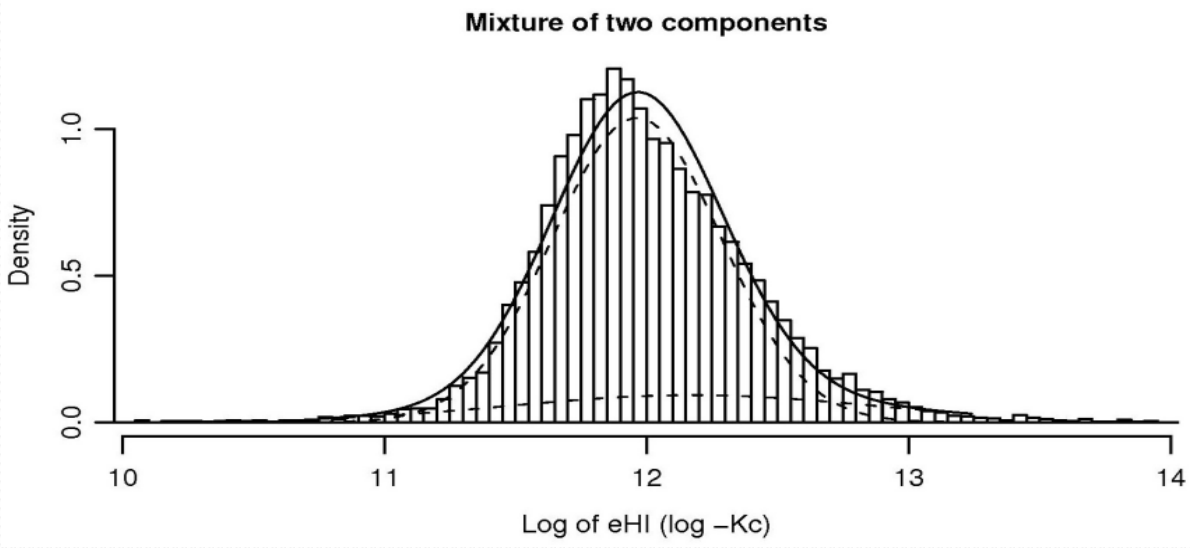
- **Překrývání shluků** odpovídá kovariancím komponent.
- **Body** v grafu odpovídají odhadům náhodných parametrů modelu

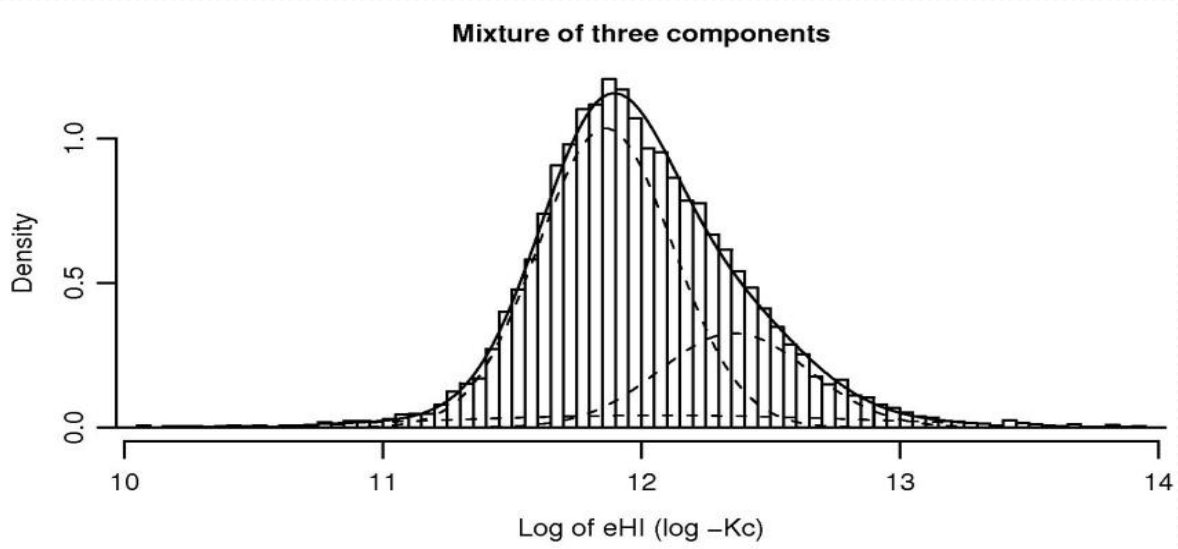
(**označení** symboly **1 – 4**, odpovídá příslušnosti k jednotlivým shlukům – komponentám modelu)

LMM-AR MODEL VÝVOJE LOGARITMŮ PŘÍJMŮ V JEDNOTLIVÝCH KOMPONENTÁCH

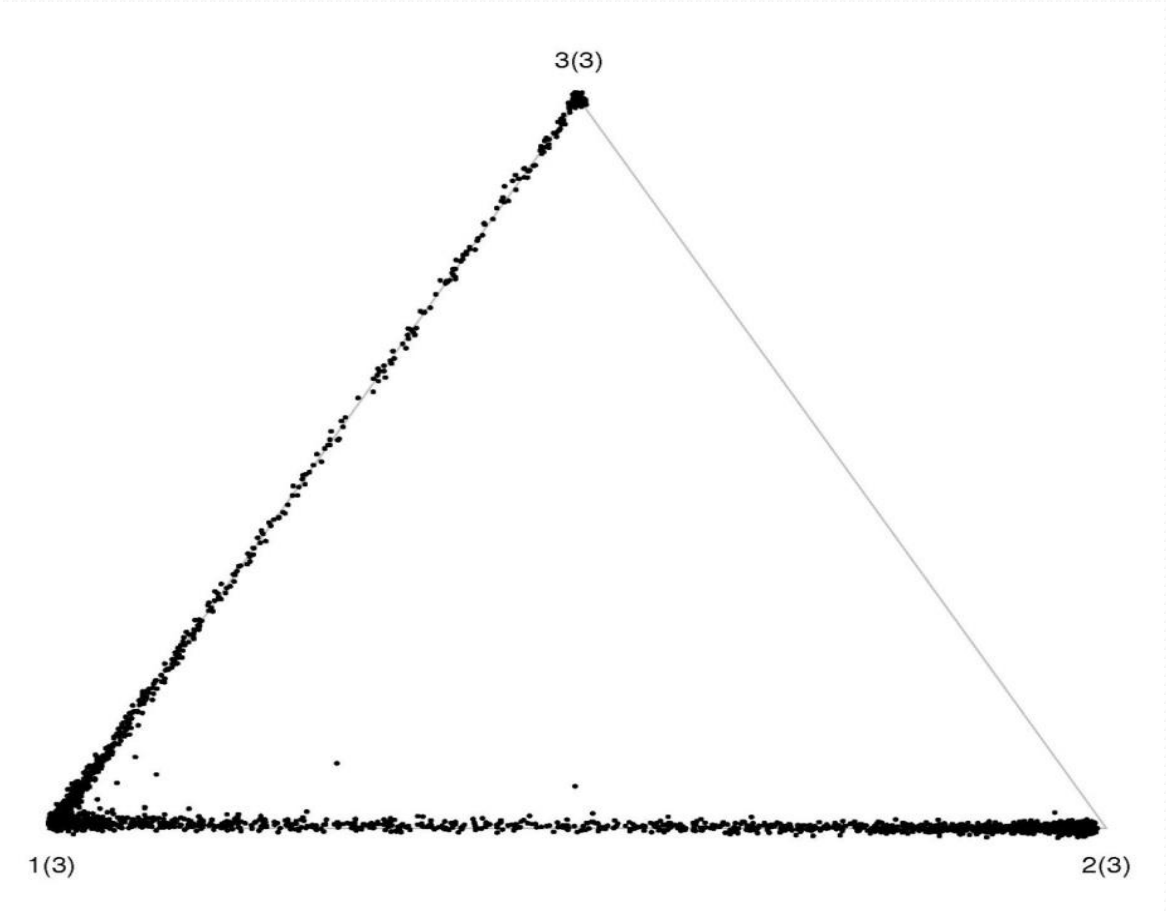


Aplikace směsí hustot





Zobrazení pravděpodobnostního přiřazení objektů do komponent



Pravděpodobnosti přiřazení domácností do jednotlivých komponent v *třísložkové směsi hustot*

Specifikace počtu shluků

Penalizovaná informační kritéria

Penalizace vnáší do modelu „pokutu“ za zvyšování počtu parametrů v modelu – růst jeho složitosti.

- ***Akaikovo* (*Akaike information criterion* –** *AIC*, viz Akaike, 1974), které vychází z konceptu *informační entropie:*

$$AIC = 2m - 2\ln(maximized\ L)$$

  kde *m* je počet neznámých parametrů a *maximized L* je maximum věrohodnostní funkce pro daný mode

- ***Bayesovské* (*Bayesian information criterion* –** *BIC*, viz Schwarz, 1978):

$$BIC = m\ln(n) - 2\ln(maximized\ L)$$

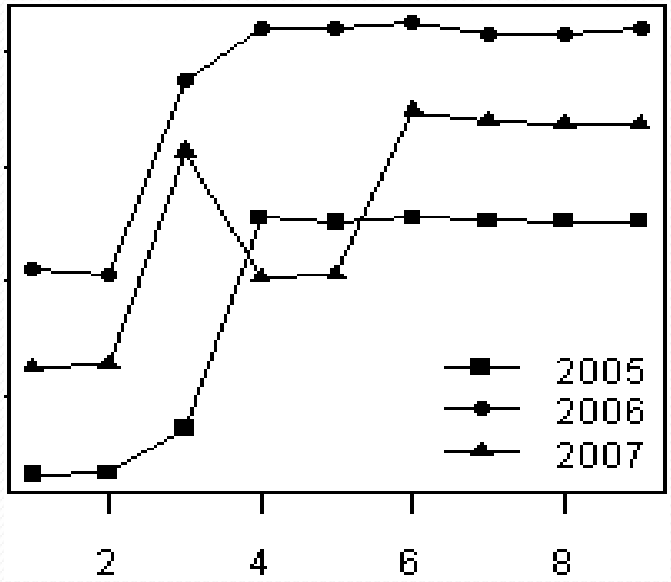
  kde *m* je počet neznámých parametrů a *n* je počet všech pozorování.

Kritéria pouze **porovnávají dvojice odhadnutých modelů**, z nichž jeden má vždy o složku méně, než druhý.

(viz Fraley a Raftery, 2003)

Odhady počtu komponent

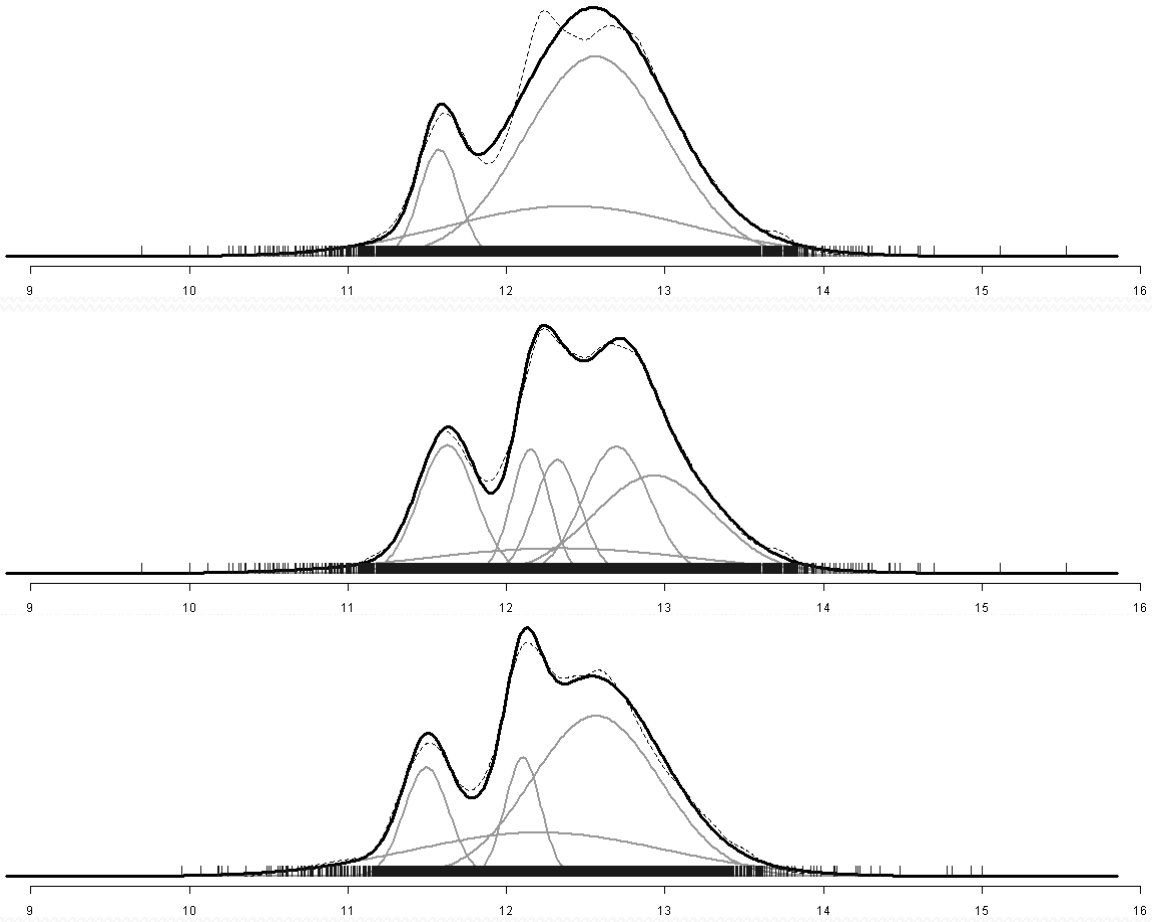
- K odhadům se často využívá **procedura *mclust*** (viz Fraley a Raftery, 2006), který je **součástí knihovny programu R**.
- Procedura ***mclust*** je **kombinací algoritmů *EM* / *BIC***. Zahrnuje *EM*-algoritmus pro vícerozměrné normální směsi hustot a *BIC* odhad počtu komponent.



***BIC* jako funkce počtu složek:**
Hodnoty kritéria *BIC* pro logaritmy čistých ročních příjmů českých domácností. (EU-SILC 2005 – 2007)

Modely s různým počtem komponent

- Tří složková směs
- Šesti složková směs
- Čtyř složková směs



Aplikace konečných směsí hustot na **modelování rozložení logaritmů** čistých měsíčních **příjmů** českých domácností

Problematika odhadu počtu komponent

Odhad často převyšuje skutečný počet shluků – potřeba **slučování.**

Existuje celá řada **slučovacích konceptů** (pro normální směsi). Přehled (včetně vlastních návrhů, odlišných od návrhů využívajících zjišťování unimodality) uvádí např. Hennig (2010).

Jedná se např. o:

- **grafické diagnostiky unimodality** a využití **hierarchického slučování (*hierarchical methods*)** založeného na statistice ***dip*** (viz Hartigan a Hartigan, 1985), tj. na hodnotě *maxima diference* mezi empirickou distribuční funkcí a *jednovrcholovou* (*unimodal*) teoretickou distribuční funkcí, která toto maximum minimalizuje (viz Tantrum, Murua a Stuetzle, 2003)
- **hierarchické slučování (*non-hierarchical methods*)** – např.
  - tvorba směsí normálních směsí,
- metodu **k-průměrů (*k-means*)** apod.

Problematika odhadu počtu komponent

- Existuje rovněž **explicitní výraz** pro **míru vzdálenosti dvou vícerozměrných normálních rozdělení** (viz Fukunaga, 1990), který zahrnuje determinant a inverzní matici ke kovarianční matici komponent. Jeho hodnotu však nelze přesně odhadnout a získané výsledky mohou být (v případě téměř singulárních matic) vychýlené.
- Velké množství literatury se zabývá také metodami **určení počtu shluků**.

Index záměny

Položme si otázku:

- Můžeme komponenty konečných směsí hustot příjmů, získané např. prostřednictvím procedury *mclust* (kombinaci algoritmů EM/BIC) v prostředí R považovat za shluky?
- Naším cílem je najít **pravidlo pro zjišťování, zda určitá podmnožina dat či subpopulace** (popsaná rozdělením četností) **je, či není shlukem – index záměny pro měření separace a shlukování**.

Index záměny

- Mějme pro každou komponentu určenou **třídu** (spojitých) **rozdělení** a soubor těchto tříd nazvěme ***bází směsi***. (V běžném směsovém modelu nechť je báze tvořena konečným počtem normálních složek.)
- Jednotlivé **báze** pak **generují třídy rozdělení prostřednictvím *smíšení***, přičemž ***marginální pravděpodobnosti*** **jsou** zde **v roli souřadnic**.
- Využijme úvahu, že **shluk by měl být zřetelně separován od jiného shluku** a v případě, že tomu tak je, pak **pravděpodobnost přiřazení** určitého pozorování ke svému „správnému“ **shluku je velká**, byť s určitostí nevíme, do kterého shluku patří (shluk je latentní).

  (Více viz **Hennigův koncept slučování založený na přímém odhadu pravděpodobnosti nesprávné klasifikace** (Hennig, 2010).)

Index záměny

- Pro naše účely **budeme považovat za shluk komponentu směsi** a budeme ho charakterizovat dvojicí (p,D), kde ***p* je marginální pravděpodobnost** (váha shluku) a ***D* je rozdělení četností** (dané hustotou).
- Pro definici postavení dvojic shluků je zásadní **koncept záměny**.

  Pro komponenty $A = (p_A, D_A)$ a $B = (p_B, D_B)$ bude tvořit **index záměny (*confusion index*)** $r_{A/B}$ pravděpodobnost, že tyto dvě komponenty jsou nesprávně zvolené a dojde k záměně.

  - Proces přiřazování je dán **Bayesovou větou**, kde p_A a p_B jsou pravděpodobnosti jevů A, resp. B.

    Dvě komponenty budeme považovat za shluk, budou-li jejich indexy záměny, $r_{A/B}$ a $r_{B/A}$, menší než určitá předem zvolená úroveň (práh – *threshold T*).

Index záměny

- **Index záměny** komponenty *A* za *B* reprezentuje pravděpodobnost, že náhodně vybrané pozorování *x* z *B* je klasifikováno do *A*.
- **Klasifikační pravidlo**, tj. podmíněná pravděpodobnost, že libovolné pozorování *x* z definičního oboru hustot f_A (f_B) pochází z komponenty *A*, resp. *B* je definováno vztahem

$$r_{x,A|AB} = \frac{p_A f_A(x)}{p_A f_A(x) + p_B f_B(x)}$$

(současně $r_{x,B|AB} = 1 - r_{x,A|AB}$)

Index záměny

- Jednotlivá *přiřazení* množiny objektů (domácností či jedinců) jsou *navzájem nezávislá* a **index záměny $r_{A/B}$ je** tedy **dán očekávanou pravděpodobností, že komponentě *A* bude přidělen náhodný výběr z rozdělení D_B**, takže platí

$$r_{A|B} = \int \frac{p_A f_A(x) f_B(x)}{p_A f_A(x) + p_B f_B(x)} dx \, .$$

- **Analytické vyjádření neexistuje** pro žádný netriviální pár rozdělení (s výjimkou dvou rovnoměrných rozdělení).
- **Lze** však **aproximovat s libovolnou přesností jeho empirickou verzi**, pokud použijeme velký náhodný výběr z *B* a zprůměrujeme pravděpodobnosti ve vztahu pro $r_{x,A/B}$.
- **Za specifických předpokladů lze index záměny aproximovat** matematicky **prostřednictvím Taylorova rozvoje**.

Vlastnosti indexu

- pro libovolný pár p_A a p_B **závisí** *index záměny* pouze **na podílu** p_B / p_A
- *index záměny* **není symetrický**, **splňuje** však **identitu**

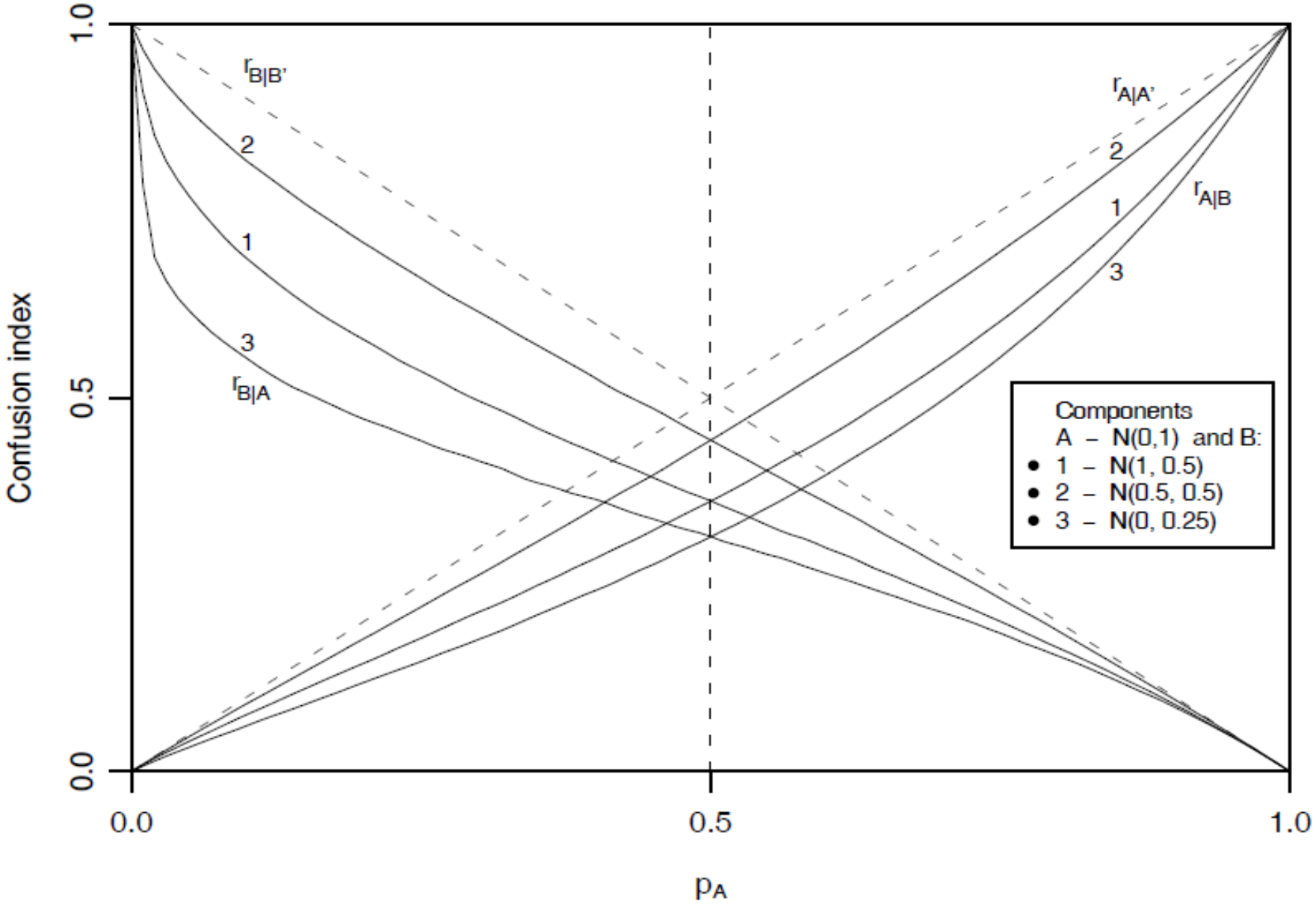
$$p_B r_{A|B} = p_A r_{B|A}$$

- pro $p_A = p_B$ je $r_{A/B} = r_{B/A}$

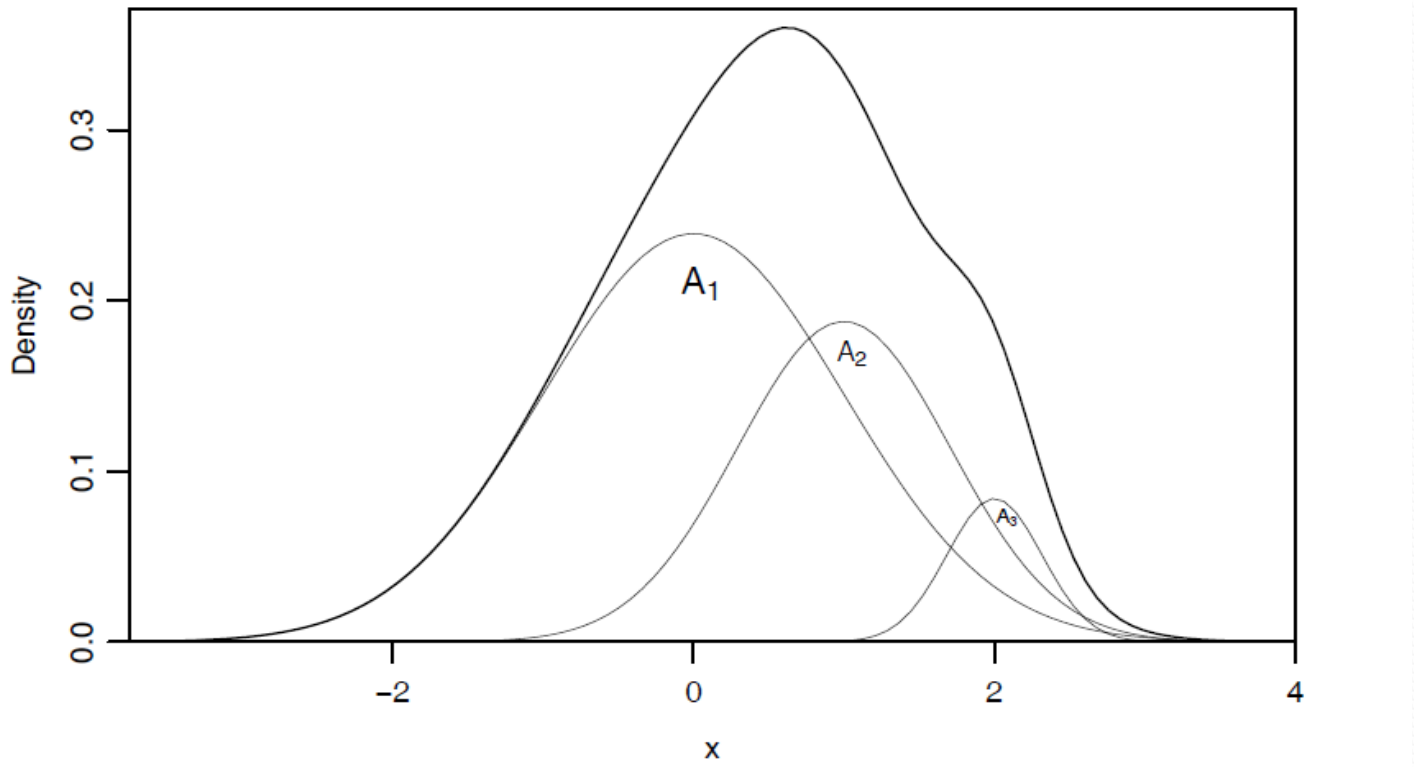
Následující graf zobrazuje **indexy záměny (*confusion indexes*)** $r_{A/B}$ a $r_{B/A}$ jako funkce pravděpodobnosti p_A

- A' (B') označuje komponentu 1 - p_A , D_A (1 - p_B , D_B)

Vlastnosti indexu



Směs tří normálních komponent



$A1$ ($p1 = 0{,}6$; $D1$: $N(0; 1)$) $A2$ ($p2 = 0{,}333$; $D2$: $N(1; 0{,}5)$)

$A3$ ($p3 = 0{,}067$; $D3$: $N(2; 0{,}1)$)

Vzhledem k tomu, že jednotlivé marginální hustoty se výrazně překrývají, nemůžeme ani jednu komponentu považovat za shluk, a to i přesto, že $A1$ a $A3$ jsou od sebe poměrně daleko.

Hodnocení pomocí indexu záměny

Pro daný *práh* (*threshold T*) lze komponenty *A* a *B* považovat za:

- **dostatečně separované**, pokud $r_{A/B}$ i $r_{B/A} < T$
- **zaměnitelné**, pokud $r_{A/B}$ i $r_{B/A} > T$
- komponenta *A* bude **satelitem** *B*, pokud $r_{A/B} < T$ a $r_{B/A} > T$

  (nutná podmínka k tomu, aby komponenta *A* mohla být satelitem *B*, je platnost nerovnosti $p_A < p_B$)

Z analýzy provedené pro spojitý interval $0 < T < 1$ vyplývá, že pro:

- $T < 0{,}042$ nebo $T = 0{,}042$ jsou **všechny** komponenty **zaměnitelné**
- $0{,}042 < T < 0{,}106$ je $A3$ **satelitem** $A1$ a ostatní páry jsou **zaměnitelné**
- $0{,}106 < T < 0{,}272$ je $A3$ **satelitem** $A1$ a $A2$, ostatní jsou **zaměnitelné**

Hodnocení pomocí indexu záměny

Úplný seznam všech teoretických možností v případě libovolných komponent *A*, *B*, *C*:

- *párová separace*
- *párová záměna*
- *A* a *B* jsou *separované* a *C*
  - je *zaměnitelná* s oběma
  - je *zaměnitelná* s jednou a separovaná od druhé
  - je *zaměnitelná* s jednou a je satelitem druhé
  - je *satelitem* oběma
  - je *satelitem* jedné a je *separovaná* od druhé
- *A* je *satelitem B* a
  - *B* je *satelitem C* a *A* je *satelitem C*
  - *B* je *satelitem C* a *A* je *zaměnitelné* s *C*
  - *C* je *zaměnitelné* s oběma

Agregace komponent a index záměny

- **Dostatečné separace množiny komponent lze mnohdy docílit agregací (sloučením) některých komponent**.
- **Matici záměny (*confusion matrix*) sloučených komponent** však **nelze získat z původní matice záměny.**

  Nová (alternativní) definice shluku po agregaci vyžaduje, aby byl shluk (*cluster*) dostatečně separován od sjednocení všech ostatních komponent.

  (Definice je podobná párové separaci, není však identická.)

Aplikace indexu záměny

Pro aplikaci byly použity *longitudinální datové soubory* z šetření **EU-SILC dvanácti vybraných zemí Evropy**:

Rakousko, Belgie, Estonsko, Španělsko, Finsko, Francie, Island, Itálie, Lucembursko, Norsko, Portugalsko, Švédsko

- **Matice záměny (*confusion matrix*)** pro čtyřkomponentní směsi (hodnoty **1000** *r*)

Rakousko

$$\begin{pmatrix} 1000 & 124 & 6 & 0 \\ 310 & 1000 & 47 & 3 \\ 117 & 388 & 1000 & 17 \\ 3 & 51 & 43 & 1000 \end{pmatrix}$$

Belgie

$$\begin{pmatrix} 1000 & 153 & 3 & 0 \\ 204 & 1000 & 32 & 0 \\ 37 & 288 & 1000 & 4 \\ 0 & 3 & 17 & 1000 \end{pmatrix}$$

Estonsko

$$\begin{pmatrix} 1000 & 162 & 8 & 0 \\ 39 & 1000 & 68 & 0 \\ 7 & 256 & 1000 & 3 \\ 0 & 1 & 28 & 1000 \end{pmatrix}$$

Španělsko

$$\begin{pmatrix} 1000 & 104 & 84 & 0 \\ 323 & 1000 & 308 & 0 \\ 192 & 228 & 1000 & 1 \\ 0 & 1 & 7 & 1000 \end{pmatrix}$$

Finsko

$$\begin{pmatrix} 1000 & 314 & 63 & 2 \\ 82 & 1000 & 194 & 6 \\ 14 & 167 & 1000 & 50 \\ 1 & 20 & 177 & 1000 \end{pmatrix}$$

Francie

$$\begin{pmatrix} 1000 & 225 & 18 & 0 \\ 108 & 1000 & 118 & 0 \\ 16 & 227 & 1000 & 1 \\ 0 & 3 & 37 & 1000 \end{pmatrix}$$

Aplikace indexu záměny

- **Matice záměny** (***confusion matrix***) pro čtyřkomponentní směsi (hodnoty **1000** *r*)

Island

$$\begin{pmatrix} 1000 & 163 & 4 & 0 \\ 41 & 1000 & 129 & 0 \\ 100 & 218 & 1000 & 1 \\ 120 & 48 & 10 & 1000 \end{pmatrix}$$

Itálie

$$\begin{pmatrix} 1000 & 66 & 4 & 0 \\ 17 & 1000 & 68 & 0 \\ 1 & 130 & 1000 & 1 \\ 0 & 0 & 10 & 1000 \end{pmatrix}$$

Lucembursko

$$\begin{pmatrix} 1000 & 208 & 9 & 0 \\ 269 & 1000 & 51 & 0 \\ 63 & 292 & 1000 & 0 \\ 0 & 1 & 8 & 1000 \end{pmatrix}$$

Norsko

$$\begin{pmatrix} 1000 & 134 & 3 & 0 \\ 109 & 1000 & 41 & 0 \\ 12 & 159 & 1000 & 2 \\ 0 & 1 & 22 & 1000 \end{pmatrix}$$

Portugalsko

$$\begin{pmatrix} 1000 & 148 & 9 & 34 \\ 193 & 1000 & 65 & 5 \\ 18 & 116 & 1000 & 0 \\ 216 & 24 & 1 & 1000 \end{pmatrix}$$

Švédsko

$$\begin{pmatrix} 1000 & 47 & 13 & 0 \\ 310 & 1000 & 89 & 7 \\ 16 & 227 & 1000 & 5 \\ 40 & 163 & 69 & 1000 \end{pmatrix}$$

Výsledky aplikace

Strukturu shluků byla zkoumána na **čtyřsložkových modelech logaritmů ekvivalizovaných příjmů domácností**.

Ve většině zemí **byla identifikována**

- **velká (hlavní) komponenta**, kterou lze charakterizovat jako skupinu **domácností s vysoce stabilním příjmem** (malým rozptylem a vysokou korelací),
- **menší, ale stále značně početná komponenta** zahrnující domácnosti **s poněkud méně stabilním příjmem**,
- **jedna** či **dvě komponenty s vysoce nestabilním příjmem** (velkým rozptylem a malou korelací).

  Tyto dvě komponenty zřejmě obsahují především chyby a nesrovnalosti v měření, protože nelze očekávat, že hodnoty ekvivalizovaných příjmů domácností se budou měnit z roku na rok tak výrazně a často, jak je tomu v těchto komponentách.

  (Pro další podrobnosti viz Bartošová, 20013)

Děkuji za pozornost