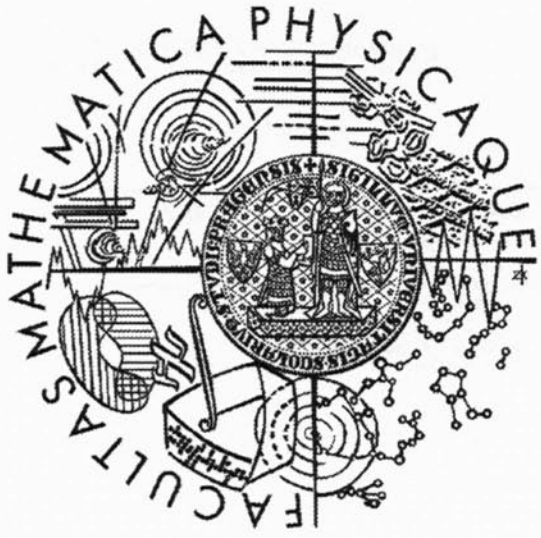
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE



Jaroslav Ševčík

SMĚSOVÉ MODELY PRAVDĚPODOBNOSTNÍCH DISTRIBUCÍ

Katedra pravděpodobnosti a matematické statistiky
Vedoucí diplomové práce: Doc. Petr Volf, CSc.
Studijní program: Matematika, Matematická statistika

Rád by som týmto poďakoval vedúcemu mojej diplomovej práce docentovi Petrovi Volfovi, ktorý mi pomohol pri spracovaní tejto neľahkej témy. Vďaka mu patrí za cenné rady, návrhy a pripomienky, ktoré mi veľmi pomohli, a tiež za dôveru a trpezlivosť, ktorú so mnou mal pri dokončovaní práce.

Z celého srdca by som chcel poďakovať svojím rodičom za ich podporu, ktorou ma sprevádzali počas celého štúdia.

Osobitnú vďaku by som chcel vysloviť svojej sestre za pomoc s jazykovou korektúrou a za jej povzbudenia pri dokončovaní práce.

Prehlasujem, že som svoju diplomovú prácu napísal samostatne a výhradne s použitím citovaných prameňov. Súhlasím so zapožičiavaním práce.

V Prahe dňa 11. augusta 2006

Jaroslav Ševčík

Obsah

Abstrakt

Názov práce: Směsové modely pravděpodobnostních distribucí
Autor: Jaroslav Ševčík
Katedra: Katedra pravděpodobnosti a matematické statistiky
Vedúci diplomovej práce: Doc. Petr Volf, CSc.
E-mail vedúceho: volf@utia.cas.cz
Abstrakt: Táto práca sa zaoberá zmesovými distribučnými modelmi, pričom ťažisko práce je zamerané na odhad počtu komponentov tohoto modelu. Prvá kapitola sa zaoberá modelom konečnej zmesi, základnými otázkami s ním spojenými a niektorými jeho aplikáciami. Druhá kapitola je venovaná odhadu zložitosti. Podrobne je rozobrané testovanie hypotézy o počte komponentov metódou pomeru vierohodností, metóda maximálnej penalizovanej vierohodnosti, metóda minimálnej penalizovanej vzdialenosti a pozornosť je venovaná aj Bayesovkým metodám. V závere kapitoly je predstavená penalizácia vierohodnosti, ktorá vedie ku konzistentnému odhadu počtu komponentov zmesi. V poslednej kapitole sú prezentované implementácie niektorých metód na odhad počtu komponentov v prostredí R, pričom pozornosť je venovaná informačným kritériám a metódam založeným na minimálnej vzdialenosti. Na rozsiahlej simulácii sú prezentované dobré vlastnosti modifikovaného BIC kritéria v prípade výberov z normálnych zmesí s rozsahom maximálne 1000. Na záver je za účelom klasifikácie zhlukov aplikovaný model zmesi na reálne dvojrozmerné dáta.
Kľúčové slová: zmesové distribučné modely, počet komponentov, konzistencia

Abstract

Title: Mixture models
Author: Jaroslav Ševčík
Department: Department of Probability and Mathematical Statistics
Supervisor: Doc. Petr Volf, CSc.
Supervisor's e-mail address: volf@utia.cas.cz
Abstract: The thesis deals with mixture models, especially problem of estimation of the number of components in mixture is discussed. First chapter concerns on finite mixture models. Basic questions coupled with this model are examined and some applications are given. Second chapter is focused on estimation of the mixture complexity. Detailed description of testing the number of components using the likelihood ratio test statistic, maximum penalized likelihood estimation and penalized minimum-distance estimation are given. Bayesian approach is also included. An appropriate penalization for maximum penalized likelihood estimator which leads to consistent estimation is demonstrated at the end of the chapter. The last chapter presents some scripts written in R computing and statistic environment as a part of the thesis and implements many of the methods mentioned in the second chapter, especially information criterions and minimum-distance methods. Good results of the modified BIC criterion in the case of normal mixtures samples of maximum lengths 1000 was achieved in simulation study. At the very end of the thesis real example of 2-dim. data is solved using mixture model.
Keywords: mixture models, number of components, consistency

Nie sú to vrcholky hôr čo zdolávame,
sme to my sami.

- SIR EDMUND HILLARY -

Úvod

Pozícia štatistika pri analýze dát, odhadovaní, predikovaní a získavaní informácií z komplexných systémov je dnes veľmi silná vďaka existencii širokého spektra pravdepodobnostných rozdelení a vďaka dostupnosti rozličných, čoraz silnejších výpočtových metód. Fascinujúcu ilustráciu tohoto aspektu predstavujú zmesové modely pravdepodobnostných distribúcií.

História zmesových distribučných modelov sa začala písať už pred viac ako sto rokmi, no do popredia širšieho záujmu sa dostala až koncom osemdesiatych rokov minulého storočia, čo súviselo predovšetkým s príchodom počítačov a expanziou výpočtovej sily, ktorá je nutnou súčasťou aplikácie týchto modelov. S veľkým entuziazmom začali byť zmesové distribučné modely používané v praxi. Dnes je ich pole pôsobnosti skutočne široké, modely zmesí sú úspešne aplikované v biológii, medicíne, fyzike, informatike, ekonómii aj marketingu.

Model so zmesou hustôt zahŕňa modely s konečným aj nekonečným počtom komponentov a umožňuje, aby komponenty pochádzali z rovnakej alebo rôznych rodín pravdepodobnostných rozdelení. Najčastejšie sa však v praxi uplatňujú konečné zmesi pravdepodobnostných distribúcií rovnakého typu, a práve tieto tvoria ťažisko mojej práce. Krása a flexibilita konečných zmesí spočíva v tom, že vrámci parametrickej rodiny poskytujú veľmi dobrú aproximáciu neparametrických modelov používaných na popis dát. V tejto súvislosti potom hovoríme o semiparametrickom prístupe modelovania komplikovaných distribučných tvarov. Najväčší potenciál modelu zmesi však spočíva v tom, že popisuje heterogénnu štruktúru dát v kontexte zhlukovej analýzy. Model zmesi ako metódu zhlukovej analýzy mnohí štatistici považujú za jediný správny štatistický prístup k hľadaniu a klasifikácii zhlukov v dátach.

Medzi základné otázky, s ktorými sa v súvislosti s modelom zmesi stretávame, patrí odhad spojitých parametrov modelu, odhad diskrétneho parametra zložitosti modelu (počet jeho komponentov) a otázka konzistencie a robustnosti týchto odhadov. Neexistujú ale žiadne explicitné formule na nájdenie týchto odhadov, preto sa musíme vždy uspokojiť s nejakým iteratívnym riešením. Spravidla teda na nájdenie odhadu používame nejaký iteratívny algoritmus, pri ktorom nás potom zaujímajú jeho vlastnosti - kvalita výsledného riešenia a časová náročnosť.

Prvá kapitola tejto práce je venovaná zmesovému distribučnému modelu. Sú v nej zavedené hlavné definície, popísané hlavné termíny a problémy, ktoré s modelom zmesi súvisia. Predstavím a popíšem základné metódy odhadu parametrov pre model s pevným počtom komponentov. Pokúsim sa tiež prezentovať potenciál zmesi pri modelovaní dát

a v závere kapitoly ukážem konkrétne prípady, v ktorých našiel model zmesi uplatnenie. Zároveň je táto kapitola prechodovým mostíkom do ďalších častí, pre ktoré sú uvedené poznatky základným východiskom.

V druhej kapitole sa budem venovať modelu zmesi s neznámym počtom komponentov a predovšetkým odhadu zložitosti takéhoto modelu, ako niekedy nazývame počet komponentov zmesi. Toto je hlavná časť, na ktorú som sa v práci zameral. Ide o náročnú problematiku, ktorá stále nebola uspokojivo vyriešená. Existuje však veľké množstvo prístupov, v ktorých som sa pokúsil zorientovať. Možno ich rozdeliť na dva prúdy. V prípade prvého má odhad zložitosti modelu dobré teoretické vlastnosti, ale nemáme efektívny algoritmus na jeho nájdenie. V druhom prípade máme zase efektívny algoritmus, ale absentujú potrebné teoretické vlastnosti. Pokúsim sa ponúknuť prehľad tejto problematiky a v závere kapitoly sa budem venovať moderným výsledkom z tejto oblasti.

V tretej kapitole sa zamerám na praktické použitie zmesí. Budem sa zaoberať možnosťou aplikácie modelov zmesí v štatistickom prostredí R. Pokúsim sa implementovať niektoré metódy pre odhad modelu zmesi a simuláciou porovnám niektoré kritériá pre výber počtu komponentov modelu. Na záver aplikujem model zmesi za účelom odhalenia a klasifikácie zhlukov na reálnom príklade viacrozmerných dát.

Kapitola 1

Stručný úvod do zmesí

Na úvod tejto práce by som chcel čitateľa zoznámiť s problematikou zmesových distribučných modelov, previesť ho akýmsi poznávacím zájazdom po svete zmesí, aby mohol potom istejšie listovať v nasledujúcich kapitolách.

1.1 Ako to všetko začalo?

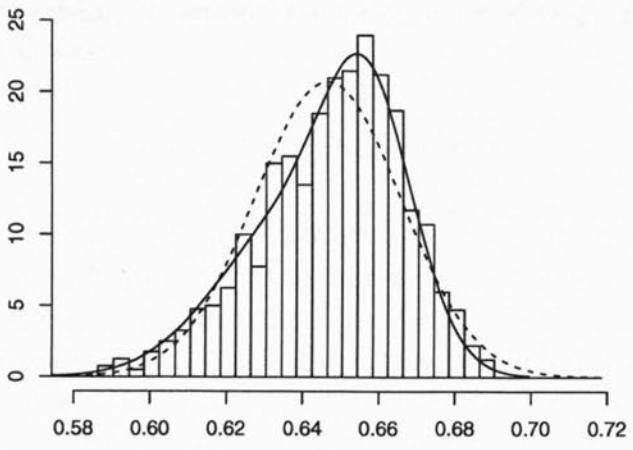
Prvé významné použitie zmesových modelov sa objavilo už pred viac ako sto rokmi, a to v práci slávneho štatistika a biometrika *Karla Pearsona*. Tento sa v roku 1894 zaoberal úlohou o kraboch, ktorú mu predložil jeho kolega *Weldon*. Weldon skúmal pomer dĺžky prednej časti hlavy k dĺžke tela krabov zo zátoky Naples na Floride. Merania boli uskutočnené na výbere o rozsahu 1000 krabov a zaznamenané do 29 intervalov. Výsledky sú znázornené na obrázku 1.1 spolu s hustotou jednoduchého normálneho rozdelenia odhadnutou z týchto dát. Weldon (1983) sa nazdával, že asymetria v histograme dát by mohla byť signálom akejsi evolúcie v populácii krabov. Domnieval sa, že v tejto populácii krabov sa vyvíjajú dva nové poddruhy[1]. Pearson (1984) použil na popísanie dát model so zmesou dvoch normálnych heteroskedastických rozdelení a na základe tohoto napokon prehlásil, že v populácii krabov sú prítomné dve subpopulácie. Na obrázku 1.1 je znázornená hustota odhadnutej zmesi[2]. Môžeme vidieť, že Pearsonovi sa týmto modelom podarilo dobre zachytiť zošikmenie dát.

1.2 Základná definícia

Ako býva už zvykom, každá matematická práca začína definovaním určitých pojmov, zavedením určitých štruktúr, objektov, poprípade zavedením pevnej notácie následne používanej v celej práci. Formálny matematický jazyk je pre prácu matematika nevyhnutný a nebude tomu inak ani v tomto prípade.

[1] Tento článok je výnimočný v tom, že ako prvý používa štatistickú analýzu ako primárnu metódu pre štúdium biologických javov.

[2] Odhad som získal metódou maximálnej vierohodnosti. Pearson ale svojho času použil metódu momentov, ktorá si vyžadovala riešenie polynomickej rovnice deviateho stupňa.



Obrázok 1.1: Graf pomeru dĺžky prednej časti hlavy k dĺžke tela pre výber 1000 krabov a hustota zmesového modelu normálnych rozdelení s jednou (prerušovaná čiara) a dvoma komponentami (plná čiara)

Nech $\mathbf{X}_1, \ldots, \mathbf{X}_n$ je náhodný výber o rozsahu n, kde $\mathbf{X}_j$ je p-rozmerný náhodný vektor s hustotou $f(\mathbf{x}_j)$ na $\mathbb{R}^p$ vzhľadom k nejakej σ-konečnej miere μ. Nech $\boldsymbol{X} = (\mathbf{X}_1^T, \ldots, \mathbf{X}_n^T)^T$ označuje celý náhodný výber a malými písmenami označme jednotlivé realizácie, teda $\boldsymbol{x} = (\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T)^T$ nech značí realizáciu celého náhodného výberu, kde $\mathbf{x}_j$ je pozorovaná hodnota náhodého vektora $\mathbf{X}_j$. Hoci hovoríme o hustote náhodného vektora $\mathbf{X}_j$, nemusí tento striktne reprezentovať len spojitý náhodný vektor. Použitie čítacej miery nám totiž umožňuje hovoriť o hustote aj v prípade diskrétneho náhodného vektora. Predpokladajme, že hustotu $f(\mathbf{x}_j)$ náhodného vektora $\mathbf{X}_j$ môžeme zapísať v tvare

$$f(\mathbf{x}_j) = \sum_{i=1}^{g} \pi_i f_i(\mathbf{x}_j), \tag{1.1}$$

kde $f_i(\mathbf{x}_j)$ sú hustoty a π_i sú nezáporné čísla, ktorých súčet je jedna, teda

$$0 \leq \pi_i \leq 1 \ (i = 1, \ldots, g) \ \text{ a } \ \sum_{i=1}^{g} \pi_i = 1. \tag{1.2}$$

Čísla $\pi_1, \ldots, \pi_g$ sa nazývajú *proporcie* alebo *váhy* zmesi. Keďže funkcie $f_1(\mathbf{x}_j), \ldots, f_g(\mathbf{x}_j)$ sú hustoty, je zrejmé, že (1.1) definuje hustotu. Funkcie $f_i(\mathbf{x}_j)$, $i = 1, \ldots, g$ nazývame *hustoty komponentov* zmesi. Hustotu (1.1) nazývame aj hustota *konečnej zmesi pravdepodobnostných distribúcií* s g komponentami. V doterajšej formulácii je počet komponentov zmesi g považovaný za fixný, resp. daný. Často sa však stretávame práve s prípadmi, keď je hodnota počtu komponentov neznáma a musí byť odhadnutá z dát spolu s proporciami zmesi a parametrami hustôt prislúchajúcimi jednotlivým komponentom v zmesi. Niekedy sa stretávame s modelmi, pri ktorých je dokonca povolené,

aby počet komponentov g rástol s rozsahom výberu[3].

Hustota (1.1) je len špeciálnym príkladom hustoty *všeobecnej zmesi pravdepodobnostných distribúcií*, ktorá má tvar

$$p_F(\mathbf{x}) = \int_\Lambda f(\mathbf{x};\ \lambda) dF(\lambda), \tag{1.3}$$

kde Λ je merateľný priestor a F je ľubovoľná pravdepodobnostná miera na tomto priestore, ktorú nazývame *distribučná funkcia zmesi*, alebo *zmesová distribúcia*[4]. Iným špeciálnym príkladom všeobecnej zmesi distribúcií sú zmesi, ktorých zmesová distribúcia F patrí do nejakej parametrickej rodiny. Typickým príkladom takejto zmesi je gamma zmes Poissonových rozdelení, ktorú predstavím v nasledujúcej kapitole.

Zavedenie všeobecnej zmesi nás v súvislosti s konečnou zmesou prirodzene privádza aj k otázke *nekonečnej zmesi*, teda k prípadu, keď každé pozorovanie (rôzne od ostatných) považujeme za realizáciu z iného pravdepodobnostného rozdelenia. Tento prípad obsahuje aj známy neparametrický jadrový odhad hustoty. Ten je pre náhodný výber $x_1, \ldots, x_n$ daný vzťahom

$$\widehat{f}(x_j) = \frac{1}{nh} \sum_{i=1}^{n} k((x_j - x_i)/h), \tag{1.4}$$

ktorý dostaneme z rovnice hustoty konečnej zmesi (1.1), ak položíme $g = n$, $\pi_i = \frac{1}{n}$ a

$$f_i(x_j) = h^{-1} k((x_j - x_i)/h),$$

kde $k(\cdot)$ je nejaká jadrová funkcia a h parameter (tzv. šírka pásma).

V sérii článkov (Simar, 1976; Laird, 1978; Lindsay, 1983) sa ukázalo, že vierohodnostná funkcia založená na náhodnom výbere z rozdelenia všeobecnej zmesi (1.3) dosahuje maximum cez všetky zmesové distribúcie F práve v distribúcii $\widehat{F}$ s konečným počtom komponentov. Konečné zmesi (zmesi s konečným počtom komponentov) teda tvoria akúsi podstatu modelov zmesí a v tejto práci bude práve im venovaná hlavná pozornosť.

1.3 Interpretácia zmesových modelov

Uvažujme hustotu konečnej zmesi (1.1) s počtom komponentov g. Nech Z_j je kategorická náhodná veličina nadobúdajúca hodnoty $1, \ldots, g$ s pravdepodobnosťami $\pi_1, \ldots, \pi_g$, a nech podmienená hustota vektora $\mathbf{X}_j$ za podmienky $Z_j = i$ je $f_i(\mathbf{x}_j)$, $i = 1, \ldots, g$. Potom (nepodmienená) hustota vektora $\mathbf{X}_j$ je daná práve hustotou zmesi $f(\mathbf{x}_j)$ z (1.1). Veličinu

[3] V prípade normálneho rozdelenia komponentov sa takýto model nazýva *sito normálenj zmesi* - toto je však len môj vlastný preklad originálneho termínu *Gaussian mixture sieve*.

[4] Aby sme sa vyhli nedorozumeniam, vzhľadom k tomu, že termín „distribučná funkcia zmesi" by mohol označovať aj celkovú distribučnú funkciu dát (v tomto prípade reprezentujúcich nejakú zmes), budem používať v tejto práci pre funkciu F termín zmesová distribúcia.

Z_j môžeme považovať za indikátor príslušnosti náhodného vektora $\mathbf{X}_j$ k určitému komponentu uvažovanej zmesi. Namiesto jednorozmernej kategorickej náhodnej veličiny Z_j je výhodnejšie pre rovnaký účel použiť g-rozmerný náhodný vektor $\mathbf{Z}_j$, ktorého i-ta zložka je definovaná ako 1 v prípade, že vektor $\mathbf{X}_j$ pochádza z i-teho komponentu zmesi, alebo nula inak. Tento vektor má teda multinomické rozdelenie

$$\mathbf{Z}_j \sim Mult_g(1,\boldsymbol{\pi}), \tag{1.5}$$

kde $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_g)^T$.

Z tejto konštrukcie vyplýva prvá interpretácia zmesového distribučného modelu s hustotou (1.1), a to *modelovanie populácie G pozostávajúcej z g skupín* $G_1,\ldots,G_g$ *v proporciách* $\pi_1,\ldots,\pi_g$. Jednotlivým skupinám G_i potom prislúchajú hustoty $f_i(\mathbf{x})$, $i = 1,\ldots,g$. Komponenty uvažovanej zmesi teda fyzicky identifikujeme s externe existujúcimi skupinami. V biometrike môžu byť zdrojom takejto heterogenity dát napríklad vek, pohlavie, druh, geografický pôvod a pod. Existuje veľa príkladov z praxe, keď je populácia zmesou g rôznych skupín, o ktorých *a priori* vieme, že existujú v istom fyzikálnom zmysle. Inokedy prítomnosť takýchto skupín vyplynie až ex post.

Komponenty zmesi však nemusíme vždy identifikovať s externe existujúcimi skupinami ako v predchádzajúcom prípade. Niekedy nás zaujíma len nájdenie vhodného modelu pre popísanie dát, popísanie ich rozdelenia (hustoty). Použitie modelu zmesi umožňuje dosiahnuť väčšiu *flexibilitu pri modelovaní dát*, ktoré nemôžeme adekvátne popísať jedným rozdelením. Dostávame sa tak k druhej interpretácii modelu so zmesou hustôt, ktorý označujeme termínom *semiparametrický* prístup modelovania dát. Názov plynie z toho, že modely zmesí okupujú zaujímavý priestor práve medzi *plne parametrickým modelom*, reprezentovaným jednou parametricky formulovanou zmesou s jedným komponentom ($g = 1$) a *neparametrickým modelom* reprezentovaným jadrovým odhadom hustoty, ktorý odpovedá zmesi s $g = n$ počtom komponentov (viď (1.4)).

Niekedy si po namodelovaní dát modelom zmesi kladieme otázku, či by sa nedali jednotlivé komponenty nájdeného modelu identifikovať s nejakými vopred nepozorovanými skupinami. Teda, či sa nám modelom náhodou nepodarilo v dátach odhaliť existenciu nami vopred nerozpoznaných a nedefinovaných subpopulácií.

1.4 Parmetrická formulácia

Často predpokladáme, že hustoty jednotlivých komponentov zmesi $f_i(\mathbf{x}_j)$ pochádzajú z nejakej parametrickej rodiny. V tomto prípade hustoty komponentov zmesi zapisujeme ako $f_i(\mathbf{x}_j;\ \boldsymbol{\theta}_i)$, kde $\boldsymbol{\theta}_i$ je vektor neznámych parametrov pre hustotu i-teho komponentu zmesi. Hustotu zmesi potom môžeme prepísať do tvaru

$$f(\mathbf{x}_j;\boldsymbol{\Psi}) = \sum_{i=1}^{g} \pi_i f_i(\mathbf{x}_j;\boldsymbol{\theta}_i), \tag{1.6}$$

kde vektor $\boldsymbol{\Psi}$ obsahuje všetky neznáme parametre v zmesovom modeli (g považujem za pevne dané), a teda

$$\boldsymbol{\Psi} = (\pi_1, \ldots, \pi_{g-1}, \boldsymbol{\xi}^T)^T, \tag{1.7}$$

kde $\boldsymbol{\xi}$ je vektor obsahujúci parametre $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_g$, o ktorých *a priori* vieme, že sú navzájom rôzne. Nech $\boldsymbol{\Omega}$ symbolizuje parametrický priestor pre $\boldsymbol{\Psi}$. Keďže proporcie zmesi π_i sa sčítajú na jednotku, jedna z nich je navyše. V neznámom parametrickom vektore (1.7) sa tak nachádza len $g - 1$ proporcií (náhodne sme vynechali proporciu π_g).

Ako príklad si uveďme zmes normálneho a Laplaceovho rozdelenia s rovnakou strednou hodnotou μ. Hustota zmesi pre tento model má tvar

$$f(x_j;\ \boldsymbol{\Psi}) = \pi_i \phi(x_j;\ \mu,\ \sigma^2) + \pi_2 (2\kappa)^{-1} exp(-\left|x_j - \mu\right| / \kappa),$$

kde

$$\boldsymbol{\Psi} = (\pi_1, \boldsymbol{\xi}^T)^T \quad \text{a} \quad \boldsymbol{\xi} = (\mu, \sigma^2, \kappa)^T.$$

Najčastejšie sa však stretávame s prípadom, keď hustoty komponentov zmesi pochádzajú z rovnakej parametrickej rodiny. Hustota zmesi má potom tvar

$$f(\mathbf{x}_j; \boldsymbol{\Psi}) = \sum_{i=1}^{g} \pi_i f(\mathbf{x}_j; \boldsymbol{\theta}_i), \tag{1.8}$$

kde funkcia $f(\cdot;\ \theta)$ je členom parametrickej rodiny $\{f(\mathbf{x}_j; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ a $\boldsymbol{\Theta}$ je parametrický priestor pre $\boldsymbol{\theta}$.

Typickým príkladom je zmes normálnych rozdelení, tzv. *normálna* alebo *gaussovská zmes*. Vo viacrozmernom prípade tak máme

$$f(\mathbf{x}_j; \boldsymbol{\theta}_i) = \phi(\mathbf{x}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \tag{1.9}$$

kde

$$\phi(\mathbf{x}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = (2\pi)^{-\frac{p}{2}} \left|\boldsymbol{\Sigma}_i\right|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\mathbf{x}_j - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\mathbf{x}_j - \boldsymbol{\mu}_i)\}$$

je hustota viacrozmerného normálneho rozdelenia so strednou hodnotou $\boldsymbol{\mu}_i$ a variančnou maticou $\boldsymbol{\Sigma}_i$. Vektorom neznámych pozorovaní je $\boldsymbol{\Psi} = (\pi_1, \ldots, \pi_{g-1}, \boldsymbol{\xi}^T)^T$, kde parameter $\boldsymbol{\xi}$ obsahuje stredné hodnoty $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_g$ a variančné matice jednotlivých komponentov $\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_g$.

1.5 Zmesová distribúcia

V prípade modelu so zmesou hustôt definovaného v (1.8) je každý z parameterov $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_g$ členom toho istého parametrického priestoru $\boldsymbol{\Theta}$. Pomocou proporcií jednotlivých komponentov $\pi_1, \ldots, \pi_g$ by sme teda mohli definovať diskrétne pravdepodobnostné rozdelenie $H(\boldsymbol{\theta})$ na tomto parametrickom priestore $\boldsymbol{\Theta}$, a to

$$H(\boldsymbol{\theta}_i) = P(\boldsymbol{\theta} = \boldsymbol{\theta}_i) = \pi_i, \quad i = 1, \ldots, g. \tag{1.10}$$

Ako alternatívu môžeme pre túto diskrétnu pravdepodobnostnú mieru na priestore $\boldsymbol{\Theta}$ písať

$$H(\boldsymbol{\theta}) = \sum_{i=1}^{g} \pi_i \mathbf{I}_{[\boldsymbol{\theta}_i \leq \boldsymbol{\theta}]}, \quad \boldsymbol{\theta} \in \boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_g\}, \tag{1.11}$$

kde nerovnosťou $\boldsymbol{\theta}_i \leq \boldsymbol{\theta}$ rozumieme $\theta_{il} \leq \theta_l$ pre každé $l = 1, \ldots, p$. Pomocou tejto pravdepodobnostnej miery môžeme potom formálne prepísať rovnicu parametricky formulovanej konečnej zmesi (1.8) na

$$f_H(\mathbf{x}_j) = f(\mathbf{x}_j; H) = \int_{\boldsymbol{\Theta}} f(\mathbf{x}_j; \boldsymbol{\theta}) dH(\boldsymbol{\theta}). \tag{1.12}$$

Funkcia H sa nazýva *zmesová distribúcia*, alebo aj *distribučná funkcia zmesi* a o jej všeobecnej forme som už hovoril v odstavci 1.2.

1.6 Flexibilná metóda modelovania

V tejto časti budem na určitých situáciách demonštrovať výhodu použitia zmesí. Najčastejšie sa v praxi stretávame s normálnym rozdelením, preto začnem práve ním.

1.6.1 Zmesi normálnych rozdelení

ZMES NORMÁLNYCH KOMPONENTOV S ROVNAKÝM ROZPTYLOM

Uvažujme zmes dvoch normálnych rozdelení s hustotou

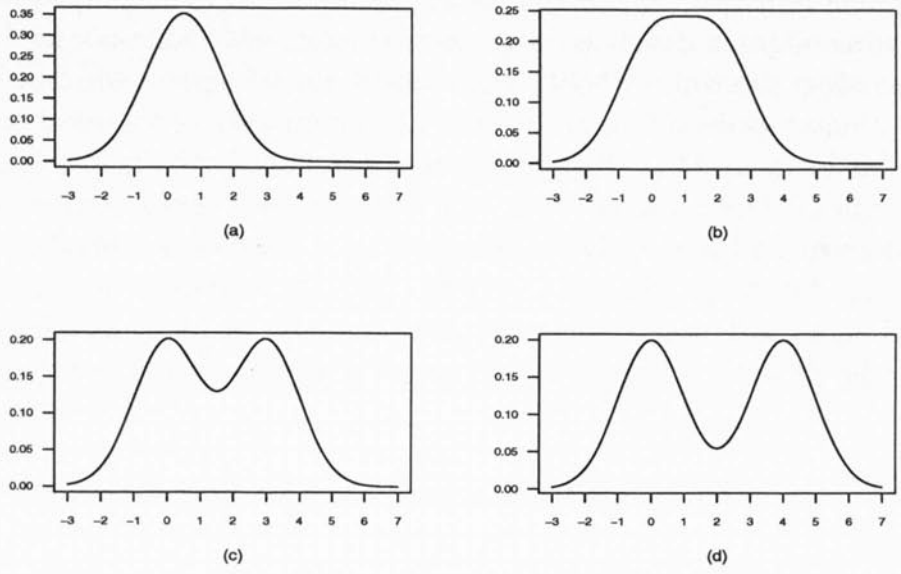
$$f(x_j) = \pi_1 \phi(x_j; \mu_1, \sigma^2) + \pi_2 \phi(x_j; \mu_2, \sigma^2), \tag{1.13}$$

kde $\phi(x_j; \mu, \sigma^2)$ je hustota normálneho rozdelenia so strednou hodnotou μ a rozptylom σ^2. Ak budú komponenty tejto zmesi od seba dostatočne vzdialené, budeme prirodzene očakávať, že ich rozdelenie bude bimodálne.

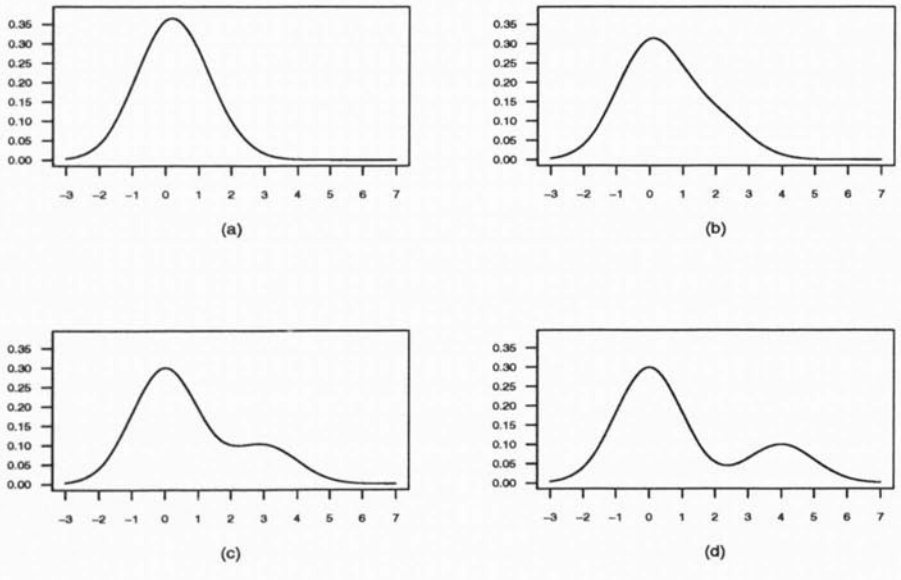
Nech najskôr zmes (1.13) má rovnako zastúpené komponenty ($\pi_1 = \pi_2 = 0.5$) s parametrami $\mu_1 = 0$, $\mu_2 = \Delta$, $\sigma^2 = 1$. Na obrázku 1.2 sú znázornené štyri hustoty takejto zmesi pre postupne sa zväčšujúcu hodnotu Δ. Vidíme, že ako Δ rastie, mení sa tvar hustoty zmesi z unimodálnej na bimodálnu. Hranicou zmeny je hodnota $\Delta = 2$, čo ostáva v platnosti aj pri zovšeobecnení Δ na $\Delta = |\mu_1 - \mu_2| / \sigma$[5]. Z obrázku môžeme ešte vidieť, aká jednoznačná môže byť niekedy prítomnosť dvoch (prípadne viacerých) populácií v dátach ($\Delta = 3$ a 4), a zároveň aké môže byť niekedy náročné takúto štruktúru odhaliť ($\Delta = 1$).

Uvažujme ďalej prípad normálnej zmesi (1.13) líšiacej sa od predchádzajúcej len inými proporciami komponentov, a to $\pi_1 = 0.75$, $\pi_2 = 0.25$. V tomto prípade sa postupne z bimodálnej hustoty ($\Delta = 4$) priblížením komponentov dostávame k zošikmenej hustote ($\Delta = 1$ a 2). Táto asymetria je spôsobená práve nerovnakým zastúpením oboch komponentov v zmesi.

[5]Táto hodnota sa nazýva *Mahalanobisova vzdialenosť* medzi homoskedastickými komponentami hustoty normálnej zmesi.



Obrázok 1.2: Hustota zmesi dvoch v rovnakom pomere zastúpených jednorozmerných normálnych komponentov s rovnakým rozptylom $\sigma^2 = 1$ a strednými hodnotami $\mu_1 = 0$, $\mu_2 = \Delta$ pre prípady (a) $\Delta = 1$, (b) $\Delta = 2$, (c) $\Delta = 3$, (d) $\Delta = 4$.



Obrázok 1.3: Hustota zmesi dvoch jednorozmerných normálnych komponentov s proporciami 0.75 a 0.25 s rovnakým rozptylom $\sigma^2 = 1$ a strednými hodnotami $\mu_1 = 0$, $\mu_2 = \Delta$ pre prípady (a) $\Delta = 1$, (b) $\Delta = 2$, (c) $\Delta = 3$, (d) $\Delta = 4$.

ZMES HETEROSKEDASTICKÝCH NORMÁLNYCH ROZDELENÍ

Väčšiu flexibilitu pri používaní normálnej zmesi získame pripustením heteroskedasticity jednotlivých komponentov. Modelom normálnej zmesi dvoch komponentov s rozličnými rozptylmi sa podrobe venuje článok Eisenberger (1964), v ktorom môžeme nájsť zaujímavé výsledky týkajúce sa charakteristiky rôznych tvarov hustoty takejto zmesi.

Peknú ukážku flexibility konečných zmesí predviedli aj Marron, Wand (1992) na 15 príkladoch normálnej zmesi (obsahujúcich 1 až deväť komponent). Grafy hustôt týchto zmesí sa nachádzajú na obrázku 1.4. Parametre týchto zmesí sú uvedené v dodatku A. Každá hustota je označená názvom[6], ktorý jej pridelili autori tohoto článku. Je to naozaj pekná ukážka toho, čo všetko dokážu normálne zmesi. Normálne rozdelenie tak v prípade konečných zmesí nadobúda úplne nové rozmery a stáva sa silnou zbraňou v rukách štatistika.

ZMESI VIACROZMERNÝCH NORMÁLNYCH KOMPONENTOV

V prípade viacrozmerných dát sme pri grafickej interpretácii odkázaní na rôzne projekcie do 1- resp. 2-dimenzionálneho priestoru.

Príkladom je *Fisherova lineárna diskriminačná funkcia*, ktorá zobrazuje viacrozmerné dáta pochádzajúce zo zmesi do jednej dimenzie tak, aby stredné hodnoty transformovaných dát boli čo najlepšie separovateľné vzhľadom k rozptylom komponentov. Nech $\mathbf{X} \sim \pi N(\boldsymbol{\theta}_1,\boldsymbol{\Sigma}) + (1-\pi)N(\boldsymbol{\theta}_1,\boldsymbol{\Sigma})$, potom lienárna diskriminačná funkcia má tvar

$$\mathbf{Y} = (\theta_1 - \theta_2)^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{X}. \tag{1.14}$$

K použitiu tejto projekcie potrebujeme odhady $\boldsymbol{\theta_1}$, $\boldsymbol{\theta_2}$ a $\boldsymbol{\Sigma}$, preto sa projekcia Y nazýva aj *náhodnou projekciou*. V jednoduchom prípade môže byť projekcia veľmi prospešná. Pokúsil som sa túto projekciu aplikovať na výber o rozsahu 150 pozorovaní z 10-rozmernej zmesi

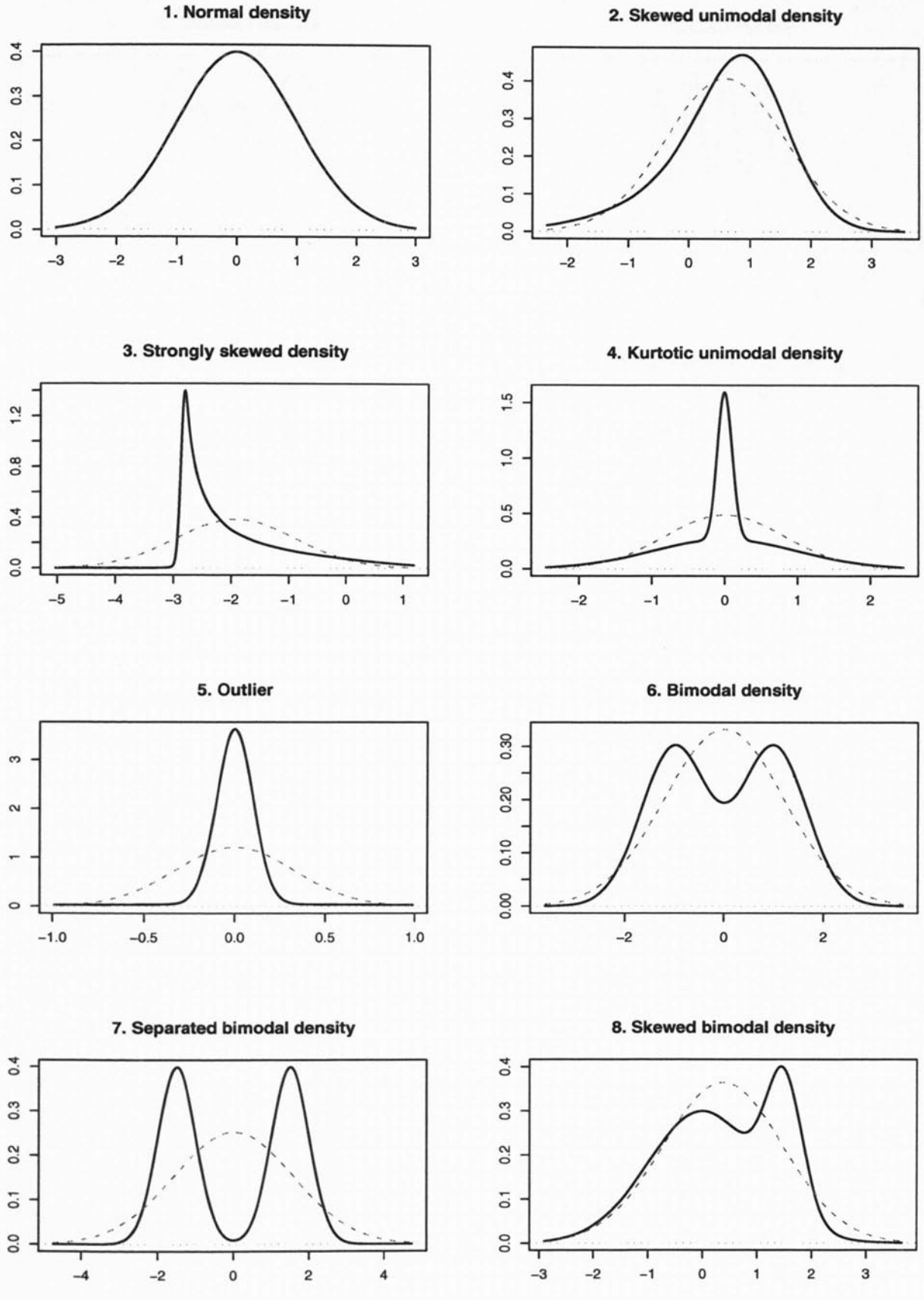
$$(2/3)N(-\mathbf{1},\,\mathbf{E}) + (1/3)N(\mathbf{1},\,\mathbf{E}),$$

kde $\mathbf{E}$ je jednotková diagonálna matica a $\mathbf{1}$, $-\mathbf{1}$ sú vektory samých jednotiek, resp. záporných jednotiek. Výsledok je v podobe histogramu dát transformovaných pomocou (1.14)[7] na obrázku 1.5. Na tom istom obrázku zároveň prezentujem aj úskalie tejto náhodnej projekcie na výber o rozsahu 50 zo sféricky symetrického 10-rozmerného normálneho rozdelenia. V tomto prípade sa mi podarilo dostať jasne bimodálny tvar histogramu transformovaných dát, ktorý mylne indikuje heterogenitu dát.

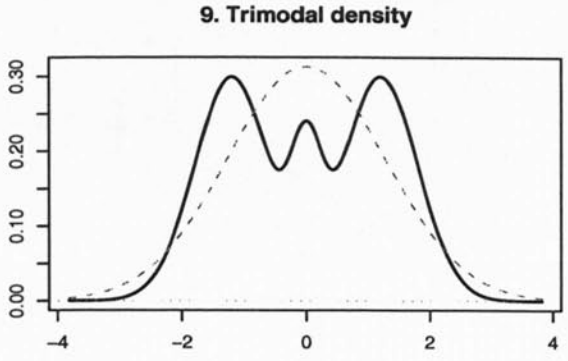
V prípade, že dáta sú dvojrozmerné, ešte stále dokážeme zobraziť hustotu, keďže táto tvorí povrch v trojrozmernom priestore. Často sa však používa aj *vrstevnicový graf* - 2D projekcia hustoty. Na obrázku 1.6 je znázornený príklad hustoty dvojrozmernej normálnej zmesi dvoch komponentov s príslušným vrstevnicovým grafom.

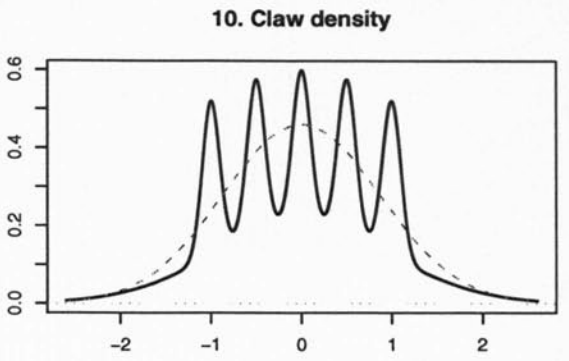
[6]Vzhľadom na to, že si netrúfam všetky názvy uspokojivo preložiť do slovenčiny, zanechal som ich v pôvodnom anglickom jazyku.

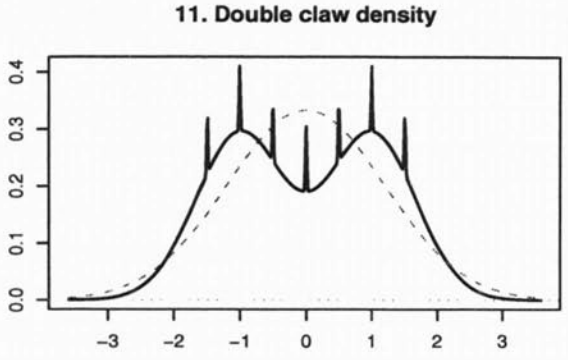
[7]Potrebné odhady boli nájdené metódou maximálnej vierohodnosti.

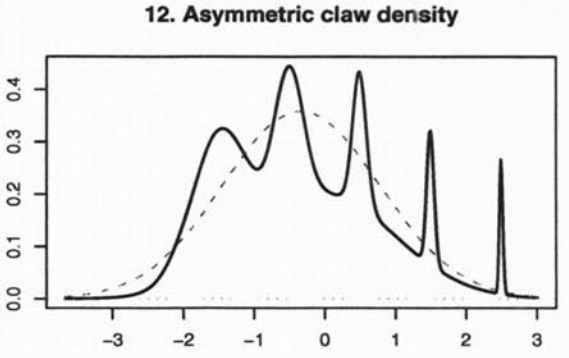


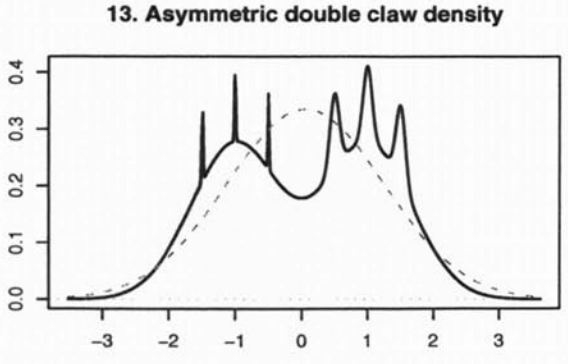
Obrázok 1.4: Hustoty normálnych zmesí z článku Marron, Wand (1992).

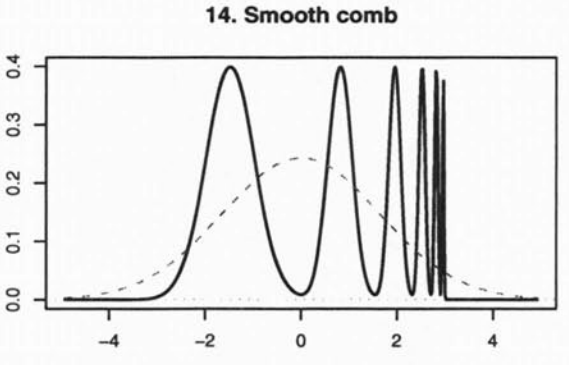

9. Trimodal density
0.00
0.10
0.20
0.30
−4
−2
0
2
4

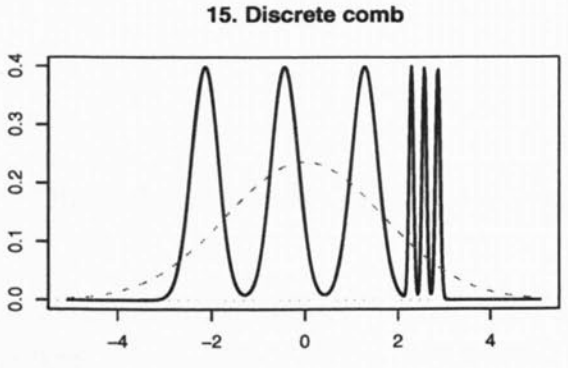

10. Claw density
0.0
0.2
0.4
0.6
−2
−1
0
1
2

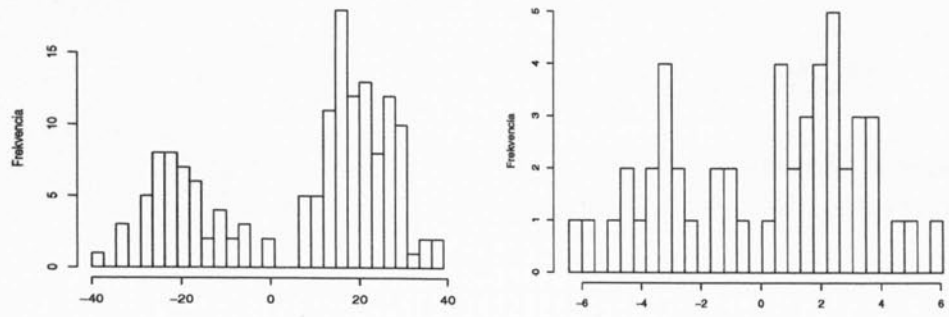

11. Double claw density
0.0
0.1
0.2
0.3
0.4
−3
−2
−1
0
1
2
3


12. Asymmetric claw density
0.0
0.1
0.2
0.3
0.4
−3
−2
−1
0
1
2
3


13. Asymmetric double claw density
0.0
0.1
0.2
0.3
0.4
−3
−2
−1
0
1
2
3


14. Smooth comb
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4

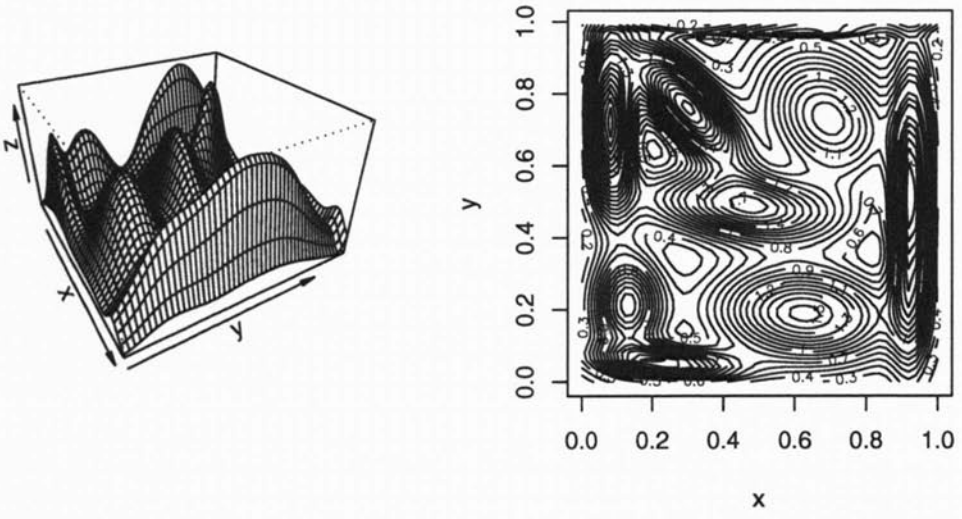

15. Discrete comb
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4



Obrázok 1.5: Histogram viacrozmerných dát transformovaných Fisherovou lineárnou diskriminačnou projekciou. Vľavo je graf pre výber o rozsahu 150 z jednoduchej 10-rozmernej zmesi s dvomi komponentami. Vpravo je graf pre výber o rozsahu 50 z 10 rozmerného normálneho rozdelenia $N(\mathbf{0},\mathbf{E})$.

1.6.2 Modelovanie asymetrických dát

Ako sme si už mohli všimnúť v úvodnom príklade o kraboch, ale aj v príklade zmesi dvoch jednorozmerných normálnych komponentov, zmesové distribučné modely zohrávajú dôležitú úlohu pri modelovaní asymetrických distribúcií. Iný spôsob, ako sa vysporiadať so šikmosťou dát, je použiť nejakú vhodnú transformáciu. Za týmto účelom sa



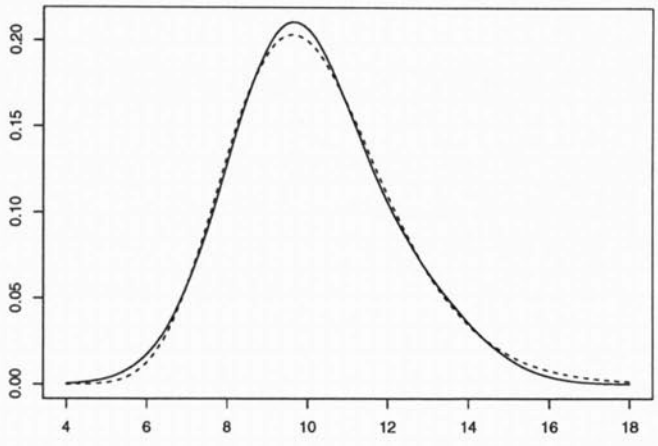
Obrázok 1.6: Príklad hustoty dvojrozmernej normálnej zmesi a jej vrstevnicového grafu.

častokrát používa napríklad logaritmická transformácia. Je známe, že parametre normálnej zmesi dvoch jednorozmerných normálnych homoskedastických komponentov možno voliť tak, že výsledná hustota zmesi je veľmi blízka hustote logaritmicko-normálneho

rozdelelnia, ktorého hustota je

$$f(x_j;\mu,\sigma^2) = (2\pi)^{-\frac{1}{2}} x_j^{-1}\sigma^{-1} \exp(-\frac{1}{2}(\log x_j - \mu)^2/\sigma^2). \tag{1.15}$$

Táto blízkosť, ktorej príklad je znázornený na obrázku 1.7 znamená, že v praxi je veľmi ťažké rozlišovať medzi týmito dvoma prípadmi. V 50-tych a 60-tych rokoch minulého



Obrázok 1.7: Graf hustoty logaritmicko normálneho rozdelenia (prerušovaná čiara) s parametrami $\mu = \log(10)$, $\sigma^2 = 0.04$ a hustoty zmesi dvoch normálnych homoskedastických komponentov (plná čiara) s parametrami $\pi_1 = 0.8$, $\mu_1 = 9.5$, $\mu_2 = 12.5$ a $\sigma^2 = 2.5$.

storočia dokonca tento problém vyústil do slávnej *Pickering/Platt rozpravy*[8] zaoberajúcej sa patofyziológiou hypertenzie. Veľa výskumníkov sa odvtedy snažilo vyriešiť tento spor aplikovaním normálnej zmesi na rozsiahle dáta o krvnom tlaku, ale výsledky boli nedostatočné. Problémom bolo dokázať opodstatnenie použitia zmesi, teda dokázať, že ide skutočne o zmes.

Problematiku „zošikmené rozdelenie verzus model zmesi" podrobne rozoberá článok Schork a kol. (1990) a referencie v ňom uvedené.

1.6.3 Robustnosť pomocou normálnej zmesi

Keďže flexibilita normálnych zmesí začala byť uplatňovaná pri odchýlení od normality, zmesi sa začali aplikovať aj na robustné modelovanie dát. Zaujímavým príkladom je Tukeyho nápad z roku 1960, ktorý použil zmes dvoch jednorozmerných normálnych

[8]Pickering a Platt boli dvaja anglický internisti s rôznymi pohľadmi na etiológiu zvýšeného krvného tlaku. Platt hlásal, že hypertenzia bola „choroba", a to že distribúcia krvného tlaku vykazovala šikmosť považoval za manifestáciu efektu Mendelianovho dominantného génu a rozdelenie krvného tlaku považoval za rozdelenie zmesi dvoch komponentov korešpondujúcich subpopulácii ľudí so zvýšeným krvným tlakom a subpopulácii ľudí s normálnym krvným tlakom. Pickering zase neochvejne zastával názor, že hypertenzia je úplne náhodný jav.

hustôt s rovnakými strednými hodnotami, ale rôznymi rozptylmi na popísanie populácie, ktorá vykazuje normálne rozdelenie až na niekoľko atypicky veľkých pozorovaní. Pracoval teda s hustotou

$$f(x_j) = \pi_1 \phi(x_j; \mu, \sigma^2) + \pi_2 \phi(x_j; \mu, k\sigma^2), \tag{1.16}$$

kde k je veľké a $\pi_2 = 1 - \pi_1$ je malé a reprezentuje malý pomer pozorovaní s relatívne veľkým rozptylom. Neskôr v roku 1964 Huber zovšeobecnil formu kontaminácie normálneho rozdelenia a odvodil známy *M-odhad parametra polohy*. Model (1.16) (*normal scale mixture model*) môže byť ešte prepísaný do všeobecnejšej podoby, a to

$$f(x_j) = \int \phi(x_j; \mu, \sigma^2/u) dH(u), \tag{1.17}$$

kde H je pravdepodobnostné rozdelenie s hodnotou π_1 v bode $u = 1$ a hodnotou $\pi_2 = 1 - \pi_1$ v bode $1/k$. Ak H nahradíme rozdelením chi-kvadrát s ν stupňami voľnosti, dostaneme z hustoty (1.17) *t-rozdelenie* o ν stupňoch voľnosti, ktoré je akousi robustnejšou alternatívou normálneho rozdelenia.

1.6.4 Ďalšie rodiny rozdelení

Okrem rodiny normálneho rozdelenia sa model so zmesou hustôt používa aj v prípade iných parametrických rozdelení. Tomu, v akých konkrétnych situáciách a za akých podmienok sa používajú zmesi iných ako normálnych rodín rozdelení, je venovaný až záver tejto kapitoly. Nateraz aspoň uvediem, že v prípade diskrétnych dát sa často môžeme stretnúť so zmesou *Poissonových* či *binomických* rozdelení, a v prípade spojitých napríklad so zmesou *t-*, *exponencionálnych*, *weibullových*, *Laplaceových*, *von-Misesových* a iných rozdelení.

1.7 Nekompletná datová štruktúra

V sekcii 1.3 bol zavedený vektor $\mathbf{Z}_j$ ako 0-1 indikátor definujúci príslušnosť náhodného vektora $\mathbf{X}_j$ k určitému komponentu zmesového modelu (1.1). Koncepcia takéhoto náhodného vektora asociovaného s náhodným vektorom $\mathbf{X}_j$ je síce užitočná, z fyzikálneho hľadiska však nie je vždy vhodné nahliadať na zmesový model prostredníctvom tejto formulácie. Zavedenie vektora $\mathbf{Z}_j$ je však kľúčové pre maximálne vierohodný odhad distribúcie zmesi, pretože umožňuje na jeho získanie priamočiaro aplikovať *EM algoritmus*, ktorý ponúka efektívne riešenie. Taktiež je táto koncepcia veľmi dôležitá pre implementáciu *MCMC metód* v bayesovskom prístupe hľadania distribúcie zmesi.

Vzhľadom k tomu, že k vstupným dátam $\mathbf{x}_1, \ldots, \mathbf{x}_n$ nemáme dostupné k nim asociované indikátory príslušnosti ku komponentu $\mathbf{z}_1, \ldots, \mathbf{z}_n$, môžeme vstupné dáta považovať za *nekompletné* a zaviesť *kompletnú datovú štruktúru* ako

$$\mathbf{x}_c = (\mathbf{x}^T, \mathbf{z}^T)^T, \tag{1.18}$$

kde $\mathbf{x} = (\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T)^T$ je vektor pozorovaných dát, alebo nekompletný datový vektor a $\mathbf{z} = (\mathbf{z}_1^T, \ldots, \mathbf{z}_n^T)^T$ je nepozorovaný vektor indikátorov príslušnosti ku komponentu.

Vzhľadom k (1.5) a vzhľadom k tomu, že náhodné vektory $\mathbf{X}_1,\ldots,\mathbf{X}_n$ sú nezávislé, považujeme vektory $\mathbf{z}_1,\ldots,\mathbf{z}_n$ za realizácie náhodných vektorov $\mathbf{Z}_1,\ldots,\mathbf{Z}_n$, pre ktoré platí

$$\mathbf{Z}_1,\ldots,\mathbf{Z}_n \overset{i.i.d}{\sim} Mult_g(1,\boldsymbol{\pi}). \tag{1.19}$$

Na i-tu zmesovú proporciu π_i potom môžeme nahliadať ako na *apriórnu pravdepodobnosť* príslušnosti entity do i-teho komponentu zmesi ($i = 1,\ldots,g$), pričom *aposteriórna pravdepodobnosť* príslušnosti entity do i-teho komponentu za podmienky, že pozorovaná hodnota entity je $\mathbf{x}_j$, potom je

$$\begin{aligned} \tau_i(\mathbf{x}_j) &= P(entita \in i-ty\ komponent \mid \mathbf{x}_j) \\ &= P(Z_{ij} = 1 \mid \mathbf{x}_j) \\ &= \pi_i f_i(\mathbf{x}_j)/f(\mathbf{x}_j) \quad (i = 1,\ldots,g;\ j = 1,\ldots,n). \end{aligned} \tag{1.20}$$

ODKUD MÁME TO?

Tieto aposteriórne pravdepodobnosti sa potom používajú ku *klasifikácii dát* v prípade, že model zmesi je aplikovaný za účelom zhlukovej analýzy. Dáta klasifikujeme tak, že pre výsledný odhad parametra spočítame pre každé pozorovanie aposteriórnu pravdepodobnosť príslušnosi k jednotlivým zhlukom modelu (1.20) a dané pozorovanie zaradíme do zhluku, pre ktorý je táto hodnota najväčšia.

Samozrejme, ak je kompletný dátový vektor $\mathbf{x}_c$ dostupný, je odhad distribúcie zmesi priamočiary. Odhady hustôt jednotlivých komponentov $f_i(\mathbf{x})$ dostaneme priamo z dát prislúchajúcich do príslušneho komponentu (teda $\mathbf{x}_j$ pre ktoré platí $z_{ij} = (\mathbf{z}_j)_i = 1$) a parametre zmesových proporcií odhadneme ako pomer dát prislúchajúcich konkrétnemu komponentu, teda

$$\hat{\pi}_i = \sum_{j=1}^{n} z_{ij}/n \quad (i = 1,\ldots,g).$$

V tomto prípade hovoríme, že dáta sú *úplne klasifikované*. Zaujímavejší je prípad *čiastočne klasifikovaných* dát, teda keď časť dát je klasifikovaná (poznáme pre ne hodnoty indikátorov príslušnosti ku komponentu) a časť klasifikovaná nie je.

V súvislosti so zavedením vektora $\mathbf{z}_j$ indikujúceho príslušnosť ku komponentu, ktorý úzko súvisí s interpretáciou zmesového distribučného modelu ako metódy zhlukovej analýzy, sa na chvíľu zastavím pri metóde zhlukovej analýzy nazývanej *metóda založená na maximálnej vierohodnosti* (maximum likelihood approach, model označíme MLA). Táto spolu s modelom so zmesou hustôt tvorí časť zhlukovej analýzy nazývanú *metódy založené na pravdepodobnostnom modeli*[9]. Dôvod, pre ktorý spomínam túto metódu je, že podobne ako model so zmesou hustôt používa veličinu z_j indikujúcu príslušnosť k danému zhluku[10], avšak na rozdiel od zmesového modelu považuje tento indikátor za súčasť parametra modelu (Nemeček, 2004).

[9]Tieto metódy sú častokrát považované za jediný správny matematický prístup k zhlukovaniu dát, keďže pri klasických zhlukovacích metódach zhluky nemáme presne definované a jedinou informáciou, ktorou do procedúry zhlukovania vstupujeme, je voľba vzdialenosti medzi dátami. Mariott(1974,str.70) o zmesiach píše: „*is about the only clustering technique that is entirely satisfactory from the mathematical point of view. It assumes a well-defined mathematical model, investigates it by well-established statistical technoques, and provides a test of significance for the results.*"

[10]V tomto prípade namiesto vektora $\mathbf{z}_j$ sa používa len náhodná veličina z_j. Teda ak j-te pozorovanie patrí do i-teho zhluku, tak $z_j = i$.

V roku 1990 bola pre zmesové modely, pre nami uvažovaný prípad nezávislých pozorovaní, použitá nomenklatúra *hidden multinomial model* (skrytý multinomický model), ktorá bola inšpirovaná práve koncepciou nekompletnej datovej štruktúry. Na základe tejto nomenklatúry vznikol potom názov pre pomenovanie zmesových distribučných modelov zavedených na všeobecnejšej štruktúre závislých dát. V prípade, že vektory $\mathbf{z}_1, \ldots, \mathbf{z}_n$ tvoria Markovov reťazec, zmesový model nazývame *hidden Markov chain model* (mohli by sme preložiť ako „model skrytého markovovho reťazca"), ktorý je veľmi populárny napríklad v oblasti *rozpoznávania zvukového či obrazového signálu.* V prípade, že vektory $\mathbf{z}_1, \ldots, \mathbf{z}_n$ korešpondujú nejakej dvojrozmernej mriežke Markovovho náhodného poľa, model zmesí nazývame *hidden Markov random field model* („model skrytého markovovho náhodného poľa"). Táto práca je však výhradne venovaná modelom zmesí s nezávislou štruktúrou dát.

1.8 Identifikovateľnosť

Odhad parametra $\boldsymbol{\Psi}$ založený na pozorovaniach $\mathbf{x}_j$ je zmysluplný, len ak je parameter $\boldsymbol{\Psi}$ identifikovateľný, resp. je identifikovateľná rodina hustôt, ktorá je určená rôznymi hodnotami parametra $\boldsymbol{\Psi}$.

Parametrická rodina hustôt $\{f(\mathbf{x}_j; \boldsymbol{\Psi}) : \boldsymbol{\Psi} \in \boldsymbol{\Omega}\}$, kde $\boldsymbol{\Omega}$ je špecifický parametrický priestor, je *identifikovateľná*, ak rôzne hodnoty parametra $\boldsymbol{\Psi}$ odpovedajú rôznym členom tejto parametrickej rodiny, teda, ak platí ekvivalencia

$$f(\mathbf{x}_j; \boldsymbol{\Psi}) = f(\mathbf{x}_j; \boldsymbol{\Psi}^*) \Leftrightarrow \boldsymbol{\Psi} = \boldsymbol{\Psi}^*, \quad \boldsymbol{\Psi}, \boldsymbol{\Psi}^* \in \boldsymbol{\Omega}. \tag{1.21}$$

1.8.1 Permutácia komponentov

Uvažujme teraz situáciu parametricky formulovanej zmesi (1.8). Pri zámene indexov jednotlivých komponentov tejto zmesi ostane prvá časť implikácie v definícii (1.21) v platnosti, ale druhá už bude porušená (vektor $\boldsymbol{\Psi}^*$ bude nejakou neidentickou permutáciou vektora $\boldsymbol{\Psi}$, a teda $\boldsymbol{\Psi} \neq \boldsymbol{\Psi}^*$). Zmesi hustôt sú totiž identifikovateľné len vzhľadom k permutácii indexov komponentov. V tomto prípade preto používame túto definíciu identifikovateľnosti:

(Id) *Nech* $f(\mathbf{x}_j; \boldsymbol{\Psi}) = \sum_{i=1}^{g} \pi_i f_i(\mathbf{x}_j; \boldsymbol{\theta}_i)$ *a* $f(\mathbf{x}_j; \boldsymbol{\Psi}^*) = \sum_{i=1}^{g^*} \pi_i^* f_i(\mathbf{x}_j; \boldsymbol{\theta}_i^*)$ *sú dva členy parametrickej rodiny hustôt konečných zmesí. Takýto systém hustôt pre* $\boldsymbol{\Psi} \in \Omega$ *je* identifikovateľný, *ak rovnosť* $f(\mathbf{x}_j; \boldsymbol{\Psi}) = f(\mathbf{x}_j; \boldsymbol{\Psi}^*)$ *s.i.* $d\boldsymbol{\mu}(\mathbf{x})$ *platí práve vtedy, keď* $g = g^*$ *a môžeme prepermutovať indexy jednotlivých komponentov tak, že* $\pi_i = \pi_i^*$ *a* $f_i(\mathbf{x}_j; \boldsymbol{\theta}_i) = f_i(\mathbf{x}_j; \boldsymbol{\theta}_i^*)$ $(i = 1, \ldots, g)$ (skratka *s.i.* $d\boldsymbol{\mu}(\mathbf{x})$ znamená rovnosť skoro isto vzhľadom k miere $\boldsymbol{\mu}$).

Pre všeobecne formulované zmesi pravdepodobnostných distribúcií (1.3) zase používame nasledovnú definíciu identifikovateľnosti:

(ID) *Systém hustôt zmesí pravdepodobnostných distribúcií*

$$\{p_F(\mathbf{x}) : p_F(\mathbf{x}) = \int_{\Lambda} f(\mathbf{x}; \lambda) dF(\lambda), \text{ F je pravdep. miera na} \Lambda\}$$

vzhľadom k nejakej σ-konečnej miere na $\mathbb{R}^p$ je identifikovateľný, *ak platí*

$$\int f(\mathbf{x};\boldsymbol{\lambda})dF_1(\boldsymbol{\lambda}) = \int f(\mathbf{x};\boldsymbol{\lambda})dF_2(\boldsymbol{\lambda}) \;\; s.i. \; d\boldsymbol{\mu}(\mathbf{x}) \Rightarrow F_1 = F_2.$$

Invariancia vierohodnosti vzhľadom k permutácii poradia komponentov v $\boldsymbol{\Psi}$ teda znamená, že vierohodnosť má $g!$ modusov. Takúto vlastnosť má potom aj aposteriórne rozdelenie parametra $\boldsymbol{\Psi}$ pri bayesovskom prístupe v prípade, že apriórne rozdelenie je symetrické vzhľadom k poradiu komponentov. Tento nedostatok identifikovateľnosti v prípade zmesí sa v niektorých prípadoch rieši zavedením určitých reštrikcií, napríklad pre proporcie zmesi

$$\pi_1 \leq \pi_2 \leq \ldots \leq \pi_g, \tag{1.22}$$

poprípade pre prvé zložky stredných hodnôt komponentov zmesi

$$(\boldsymbol{\mu}_1)_1 \leq (\boldsymbol{\mu}_2)_1 \leq \ldots \leq (\boldsymbol{\mu}_g)_1. \tag{1.23}$$

Zámena poradia komponentov však neovplyvňuje odhady zmesí pomocou maximálnej vierohodnosti, a preto sa reštrikcie v tomto prípade nepoužívajú. Značné problémy však poradie komponentov v $\boldsymbol{\Psi}$ spôsobuje pri bayesovskom odhade metódou MCMC, kde sa ukázalo, že použitie reštrikcií (1.22) a (1.23) v tomto prípade problém nerieši (Richardson, Green, 1997a).

1.8.2 Overfitting

Na neidentifikovateľnosť narážame aj v súvislosti s *overfittingom*, teda keď je zmes s $g-1$ počtom komponentov chybne modelovaná zmesou s počtom komponentov g. Toto môže nastať v dvoch prípadoch. Buď jedna z proporcií v zmesi s počtom komponentov g je nulová, alebo nejaké dve hustoty v zmesi s počtom komponentov g sú rovnaké. Predpokladajme, že skutočná hustota je $f(\mathbf{x}_j;\boldsymbol{\theta})$, ale mi použijeme model s dvoma komponentami s hustotami $f(\mathbf{x}_j;\boldsymbol{\theta}_1)$ a $f(\mathbf{x}_j;\boldsymbol{\theta}_2)$. Definujme

$$\boldsymbol{\Omega}_1 = \{\boldsymbol{\Psi} : \; \pi_1 = 1, \boldsymbol{\theta}_1 = \boldsymbol{\theta}\} \cup \{\boldsymbol{\Psi} : \; \pi_1 = 0, \boldsymbol{\theta}_2 = \boldsymbol{\theta}\}$$

a

$$\boldsymbol{\Omega}_2 = \{\boldsymbol{\Psi} : \; \pi_1 \in (0,1), \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 = \boldsymbol{\theta}\}.$$

Potom pre všetky $\boldsymbol{\Psi}$ patriace do $\boldsymbol{\Omega}_1 \cup \boldsymbol{\Omega}_2$ máme $f(\mathbf{x}_j;\boldsymbol{\Psi}) = f(\mathbf{x}_j;\boldsymbol{\theta})$ a rozšírenie na ľubovoľné veľké g je priamočiare. Tento problém odstraňujeme tak, že požadujeme, aby všetky proporcie modelu zmesi boli nenulové a parametre jednotlivých komponentov zmesi navzájom rôzne.

1.8.3 Všeobecná neidentifikovateľnosť

Bolo dokázané, že podmienku identifikovateľnosti **(Id)** splňujú jednorozmerné zmesi normálnych, gama, exponenciálnych, Cauchy a Poissonvých rozdelení, pričom zmesi

diskrétnych alebo spojitých rovnomerných rozdelení identifikovateľné nie sú (Titterington a kol., 1985). Zmesi binomických a multinomických rozdelení sú zase identifikovateľné len v určitých prípadoch, napr. zmes binomických rozdelení je identifikovateľná len ak $N \geq 2g - 1$, kde N je počet Bernoulliho pokusov v uvažovanom binomickom rozdelení.

1.9 Metódy odhadu

Za niečo viac ako storočie existencie modelu so zmesou hustôt bolo odvodených viacero spôsobov na odhadovanie parametrov tohoto modelu[11]. Za akési hlavné prúdy možno považovať *metódu momentov*, *grafické metódy*, *maximálnu vierohodnosť*, *metódy založené na minimálnej vzdialenosti* a *Bayesovské metódy*. Hlavným dôvodom veľkého zmnožstva literatúry venovanej metodológii odhadu pre model zmesi je zrejme fakt, že všeobecne pre odhad zmesi neexistuje žiadne riešenie v podobe nejakej explicitnej formule, a preto sa v prípade zmesí musíme vždy uspokojiť len s nejakým iteratívnym rešením.

1.9.1 Metóda momentov

Prvou metódou používanou na konštrukciu odhadu zmesovej distribúcie, ktorú aplikoval aj Pearson v príklade o kraboch z roku 1894, bola *metóda momentov*. Ide o klasický prístup, pri ktorom porovnávame empirické momenty odhadnuté z dát s ich teoretickými variantami vyjadrenými pomocou parametrov. V jednoduchom prípade zmesi dvoch normálnych komponentov s rôznymi rozptylmi je však už nutné za týmto účelom riešiť sústavu až piatich nelineárnych rovníc. Lindsay (1989) navrhol pre riešenie odhadu touto metódou šikovný a veľmi elegantný algoritmus založený na tzv. *momentových maticiach* (moment matrices), ktorý môže byť jednoducho implementovaný prostredníctvom softvéru umožňujucého prácu s maticami. V prípade, že riešenie existuje, je jediné a je nájdené veľmi rýchlo. Veľkou nevýhodou metódy momentov však je, že všeobecne neposkytuje pre model so zmesou hustôt konzistentný odhad, hoci v určitých špeciálnych prípadoch sa podarilo odvodiť taký systém momentových rovníc, ktorý ku konzistentnému odhadu už vedie (Lindsay, Basak, 1993; Heckman a kol., 1990). Lindsay, Furman (1994) poukazujú aj na to, že odhad metódou momentov môže byť veľmi efektívnou inicializáciou pre maximalizovanie vierohodnosti zmesí (napr. pomocou EM algoritmu).

1.9.2 Maximálne vierohodný odhad

Najčastejšie používaný a najefektívnejší spôsob odhadu parametrov zmesi je *metóda maximálnej vierohodnosti* (MLE). Podiel na tom má zaručene existencia *EM algoritmu*, pomocou ktorého sa vierohodnosť maximalizuje najčastejšie. Tento algoritmus je v štatistike skutočne pojem a nepoužíva sa len v prípade odhadu parametrov pre zmes hustôt. Typicky sa uplatňuje predovšetkým pri hľadaní MLE v modeli s chýbajúcimi

[11] Odhadovanie parametrov modelu je vlastne aj odhadovaním zmesovej distribúcie modelu. Pokiaľ je to z kontextu zrejmé používa sa v tejto súvislosti len termín „odhadovanie modelu".

dátami. V prípade zmesí ponúka najefektívnejšie riešenie spomedzi existujúcich metód. Vráťme sa ale ešte k všeobecnému odhadu metódou maximálnej vierohodnosti, ktorý bol nie vždy konštruovaný aplikáciou EM algoritmu.

VLASTNOSTI MLE PRE ZMESI HUSTÔT

Ako je dobre známe, odhad $\hat{\boldsymbol{\Psi}}$ parametra $\boldsymbol{\Psi}$ metódou maximálnej vierohodnosti je riešením vierohodnostnej rovnice

$$\partial L(\boldsymbol{\Psi})/\partial \boldsymbol{\Psi} = \mathbf{0},$$

alebo ekvivalentne rovnice

$$\partial \log L(\boldsymbol{\Psi})/\partial \boldsymbol{\Psi} = \mathbf{0}, \tag{1.24}$$

kde

$$L(\boldsymbol{\Psi}) = \prod_{j=1}^{n} f(\mathbf{x}_j; \boldsymbol{\Psi}) = \prod_{j=1}^{n} \left[\sum_{i=1}^{g} \pi_i f_i(\mathbf{x}_j; \boldsymbol{\theta}_i) \right]. \tag{1.25}$$

Zo všeobecnej teórie vieme, že ak existuje MLE parametra $\boldsymbol{\Psi}$ vierohodnostenej funkcie (1.25) na kompaktnom parametrickom priestore, potom za určitých podmienok (Wald, 1949) je tento odhad *konzistentný* (pripomínam, že predpokladáme identifikovateľnosť). V prípade zmesí sa však pomerne často stretávame so situáciou, keď vierohodnosť nie je ohraničená[12], a tak MLE nemôže existovať ako globálne maximum. V tejto situácii však MLE môže stále ešte existovať ako lokálne maximum vierohodnosti. Pre triedu identifikovateľných zmesí boli odvodené *podmienky regularity*[13] (Peters, Walker, 1978; Redner, Walker, 1984), za ktorých existuje postupnosť koreňov vierohodnostnej rovnice $\partial L(\boldsymbol{\Psi})/\partial \boldsymbol{\Psi} = \mathbf{0}$ s vlastnosťou konzistencie, eficiencie a asymptotickej normality. Samozrejme v prípade, že globálne maximum vierohodnsoti neexistuje, vzniká otázka, ktoré lokálne maximum vziať za hľadaný odhad, resp. či nájdené lokálne maximum možno považovať za hľadaný odhad (Gan, Jiang, 1999). Niekedy sa nevhodné lokálne extrémy dajú identifikovať na prvý pohľad, inokedy zase nie.

ŠTANDARDNÁ CHYBA

Pripomeňme si, že Fisherova informačná matica vektora parametrov $\boldsymbol{\Psi}$ je definovaná ako

$$J(\boldsymbol{\Psi}) = E_{\boldsymbol{\Psi}} \left\{ \mathbf{S}(\mathbf{X}; \boldsymbol{\Psi}) \mathbf{S}^{\mathbf{T}}(\mathbf{X}; \boldsymbol{\Psi}) \right\}, \tag{1.26}$$

kde

$$\mathbf{S}(\mathbf{X}; \boldsymbol{\Psi}) = \partial \log L(\boldsymbol{\Psi})/\partial \boldsymbol{\Psi}$$

je gradient logaritmicko vierohodnostnej funkcie a $\mathbf{x} = (\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T)^T$ je vektor pozorovaných dát. Za podmienok regularity potom $J(\boldsymbol{\Psi})$ môže byť vyjadrená ako

$$J(\boldsymbol{\Psi}) = E_{\boldsymbol{\Psi}} \left\{ \mathbf{I}(\Psi; X) \right\}, \tag{1.27}$$

[12]Príkladom môže byť napríklad zmes dvoch jednorozmerných normálnych komponentov, pre ktoré $\mu_1 = x_1$, $\sigma_1^2 \to 0$ a stredná hodnota a rozptyl druhého komponentu sú ľubovoľné.

[13]Platia pre väčšinu bežne používaných parametrických rodín. Ide o zovšeobecnenie známych Cramérových podmienok.

kde

$$\mathbf{I}(\Psi; X) = -\partial^2 \log L(\boldsymbol{\Psi}) / \partial \boldsymbol{\Psi} \partial \boldsymbol{\Psi}^T \tag{1.28}$$

je záporný hesián logaritmicko vierohodnostnej funkcie. Zo všeobecnej teórie je známe, že *asymptotická variančná matica* pre MLE $\hat{\boldsymbol{\Psi}}$ sa rovná inverzii informačnej matici $J(\boldsymbol{\Psi})$, ktorá môže byť aproximovaná pomocou $J(\hat{\boldsymbol{\Psi}})$. To znamená, že štandardná chyba parametra $\hat{\Psi}_r = (\hat{\boldsymbol{\Psi}})_r$ je

$$SE(\hat{\Psi}_r) \approx (J^{-1}(\hat{\boldsymbol{\Psi}}))_{rr}^{1/2} \quad (r = 1, \ldots, d), \;\; d = dim(\boldsymbol{\Psi})$$

V praxi sa však namiesto informačnej matice $J(\boldsymbol{\Psi})$ vyjadrenej v bode $\boldsymbol{\Psi} = \hat{\boldsymbol{\Psi}}$ používa výberová informačná matica $\mathbf{I}(\hat{\Psi}; x)$, počítaná pomocou (1.28). Potom teda pre štandardné chyby jednotlivých zložiek odhadovaného vektora parametrov $\boldsymbol{\Psi}$ máme

$$SE(\hat{\Psi}_r) \approx (I^{-1}(\hat{\boldsymbol{\Psi}}, \mathbf{x}))_{rr}^{1/2} \quad (r = 1, \ldots, d). \tag{1.29}$$

EM ALGORITMUS

Venujme sa teraz praktickému hľadaniu parametra $\boldsymbol{\Psi}$ metódou MLE (počet komponentov g je známy). Z tvaru logaritmicko vierohodnostnej funkcie parametra $\boldsymbol{\Psi}$

$$\log L(\boldsymbol{\Psi}) = \sum_{j=1}^{n} \log \left\{ \sum_{i=1}^{g} \pi_i f_i(\mathbf{x}_j; \boldsymbol{\theta}_i) \right\} \tag{1.30}$$

je zrejmé, že explicitný vzorec pre maximum, resp. lokálne maximá, sa nám vo všeobecnom prípade nájsť nepodarí. To je dôvod, prečo bolo potrebné vyvinúť nejaký iteratívny algoritmus na nájdenie maxima. Nakoniec sa to podarilo algoritmu EM (Dempster a kol., 1977). Uveďme si jeho základnú podobu pre model zmesi (1.6).

Nech $\mathbf{x}_c = (\mathbf{x}^T, \mathbf{z}^T)^T$ je podľa zavedenej notácie kompletný datový vektor a jemu prislúchajúcu logaritmickú vierohodnosť pre parameter $\boldsymbol{\Psi}$ označme $\log L_c(\boldsymbol{\Psi})$. Potom platí

$$\log L_c(\boldsymbol{\Psi}) = \sum_{i=1}^{g} \sum_{j=1}^{n} z_{ij} \left\{ \log \pi_i + \log f_i(\mathbf{x}_j; \boldsymbol{\theta}_i) \right\}. \tag{1.31}$$

EM algoritmus spočíva v maximalizácii strednej hodnoty tejto logaritmickej vierohodnosti $\log L_c(\boldsymbol{\Psi})$ podmienenej pri danom $\mathbf{x}$

$$E_{\boldsymbol{\Psi}^{(k)}} \left[\log L_c(\boldsymbol{\Psi}) \mid \mathbf{x} \right] =: Q(\boldsymbol{\Psi} | \boldsymbol{\Psi}^{(k)}), \tag{1.32}$$

kde $\boldsymbol{\Psi}^{(k)}$ je odhad parametra $\boldsymbol{\Psi}$ z predchádzajúcej iterácie.

Algoritmus treba naštartovať nejakou počiatočnou hodnotou parametra $\boldsymbol{\Psi}$, $\boldsymbol{\Psi}^{(0)}$. V $(k+1)$-vom kroku algoritmu potom z $\boldsymbol{\Psi}^{(k)}$ zostrojíme $\boldsymbol{\Psi}^{(k+1)}$ v dvoch fázach:

1. **E-step** - *spočítame* $Q(\Psi|\Psi^{(k)})$

$$\begin{aligned} Q(\Psi|\Psi^{(k)}) &= E_{\Psi^{(k)}}\left[\sum_{i=1}^{g}\sum_{j=1}^{n} z_{ij}[\log \pi_i + \log f_i(\mathbf{x}_j;\boldsymbol{\theta}_i)] \mid \mathbf{x}\right] \\ &= \sum_{i=1}^{g}\sum_{j=1}^{n} E_{\Psi^{(k)}}(z_{ij}|\mathbf{x})[\log \pi_i + \log f_i(\mathbf{x}_j;\boldsymbol{\theta}_i)]], \end{aligned}$$

kde

$$\begin{aligned} \tau_i(\mathbf{x}_j;\Psi^{(k)}) := E_{\Psi^{(k)}}(z_{ij}|\mathbf{x}) &= P_{\Psi^{(k)}}(z_{ij}=1|\mathbf{x}) \\ &= \pi_i^{(k)} f_i(\mathbf{x}_j;\boldsymbol{\theta}_i^{(k)}) / \sum_{h=1}^{g} \pi_h^{(k)} f_h(\mathbf{x}_j;\boldsymbol{\theta}_h^{(k)}) \end{aligned}$$

je aposteriórna pravdepodobnosť, že j-ty člen výberu pri pozorovanej hodnote $\mathbf{x}_j$ patrí do i-teho komponentu zmesi pri hodnote parametra $\Psi^{(k)}$.

2. **M-step** - *volíme* $\Psi^{(k+1)}$ *tak, aby sme maximalizovali* $Q(\Psi|\Psi^{(k)})$ *cez* Ψ

Maximalizácia $Q(\Psi|\Psi^{(k)})$ cez $\boldsymbol{\pi}$ vedie na jednoduchú úlohu o viazanom extréme, ktorej riešenie je

$$\pi_i^{(k+1)} = \frac{1}{n}\sum_{j=1}^{n} \tau_i(\mathbf{x}_j;\Psi^{(k)}).$$

Vektor $\boldsymbol{\xi}^{(k+1)}$ dostaneme riešením rovnice

$$\sum_{i=1}^{g}\sum_{j=1}^{n} \tau_i(\mathbf{x}_j;\Psi^{(k+1)}) \frac{\partial \log f_i(\mathbf{x}_j;\boldsymbol{\theta}_i)}{\partial \boldsymbol{\xi}} = \mathbf{0}. \tag{1.33}$$

Peknou vlastnosťou EM algoritmu je, že riešenie rovnice (1.33) pre väčšinu prípadov existuje v uzavrenej forme.

Fázy E-step a M-step dávajú EM algoritmu jeho meno. Ide o prvé písmená anglických slov *expectation* a *maximization*. V slávnom článku Dempster a kol. (1977) autori ukázali, že vierohodnosť (už pre nekompletný datový vektor) po EM iterácii neklesá, teda že

$$L(\Psi^{(k+1)}) \geq L(\Psi^{(k)}).$$

NEZNÁMY POČET KOMPONENTOV

V doterajších úvahach sme stále považovali počet komponentov g v zmesi za známy (viď zavedenie parametra Ψ (1.7)), teda g nebolo súčasťou odhadovaného parametra. Prípadu, keď je počet komponentov zmesi neznámy, je venovaná druhá kapitola. V prípade MLE odhadu však spravidla postupujeme tak, že pre množinu rôznych vhodných hodnôt g spočítame odhad vektora parametrov Ψ a tieto odhady prislúchajúce rôznym hodnotám g potom medzi sebou vzájomne porovnávame a podľa nejakého kľúča sa

rozhodneme pre výsledný model. Vedieť odhadnúť model zmesi pre známy počet komponentov je teda dôležité aj pre odhad počtu komponentov zmesi.

NAPÍSALO SA O EM ALGORITME

O EM algoritme sa toho publikovalo skutočne mnoho. Meng, Dyk (1997) uvádzajú, že v rokoch 1977 až 1997 to bolo viac než 1000 článkov a odvtedy ich počet ešte rástol. Čím všetkým sa teda autori článkov o EM algoritme v súvislosti so zmesami hustôt zaoberajú? Vo väčšine prípadov je to jeden z nasledovných okruhov:

- **Voľba počiatočnej inicializácie algoritmu** - algoritmus zvykne uviaznuť v lokálnych maximách a je značne pomalý, takže dobrá inicializácia môže výrazne zlepšiť kvalitu výsledku. Výborný prehľad používaných inicializácií možno nájsť v článku Karlis (2003). Vhodnou inicializáciou v prípade veľkého rozsahu dát sa zaoberá Coleman, Woodruff (2000).

- **Ukončenie iterácií** (stopping criterion) - väčšinou sa algoritmus ukončuje v okamihu, keď sa ďalšou iteráciou logaritmická vierohodnosť zväčší o menej ako je vopred stanovená hranica tolerancie, teda ak $\left|l^{k+1} - l^k\right| \leq tol$, kde $l^k = \log L(\Psi_k)$. Zaujímavou modifikáciou tohoto ukončovacieho kritéria je *Aitkenovo kritérium zrýchlujúce konvergenciu* (McLachlan, Peel, 2000, str. 52-53). Ďalším kritériom môže byť napríklad malá relatívna zmena odhadovaného parametra.

- **Rýchlosť konvergencie algoritmu** - bežnou kritikou EM algoritmu je jeho pomalá konvergencia v niektorých prípadoch. Za týmto účelom zrýchlenia konvergencie vzniklo niekoľko rôznych modifikácií EM algoritmu, ktoré sú však často vykúpené nestabilitou alebo komplikovanosťou algoritmu. Medzi pomerne úspešné modifikácie EM algoritmu patrí *Prírastkový EM* (*IEM*), *Riedky EM* (*SPEM*) a *Lenivý EM* (*Lazy EM*) algoritmus, no známe sú aj *ECME*, *AECM* a *PX-EM*. Prehľad k týmto algoritmom možno nájsť v monografii McLachlan, Peel (2000).

- **Výpočet štandardných chýb odhadu** - samotný EM algoritmus neposkytuje odhad variančnej matice odhadu $\hat{\boldsymbol{\Psi}}$, ako tomu je napr. v prípade metód Newtonovho typu. Vzhľadom k rozšírenosti EM algoritmu vznikli preto rozličné metódy pre odhad variančnej matice $\hat{\boldsymbol{\Psi}}$, resp. štandardných chýb $\hat{\boldsymbol{\Psi}}$. Väčšinou sú založené na výpočte výberovej informačnej matice $\mathbf{I}(\hat{\boldsymbol{\Psi}}; \mathbf{x})$ počítanej pomocou (1.28), ktorou v praxi odhadujeme inverziu variančnej matice $\hat{\boldsymbol{\Psi}}$[14].

- **Špeciálne prípady** - model so zmesou hustôt našiel uplatnenie v rôznych oblastiach, a preto nás častokrát zaujíma konkrétna podoba EM algoritmu pre rôzne

[14]Priamy výpočet $\mathbf{I}(\hat{\boldsymbol{\Psi}}; \mathbf{x})$, ktorý vyžaduje analytické vyčíslenie druhej derivácie logaritmickej vierohodnosti, je pre väčšinu zmesí obtiažny alebo prinajmenšom nudný a únavný. Boli preto odvodené aj iné spôsoby aproximácie $\mathbf{I}(\hat{\boldsymbol{\Psi}}; \mathbf{x})$ (Diebolt, Ip, 1996), ktoré nazývame súhrnne *metódy založené na informačnej matici*. Tieto vychádzajú z asymptotickej teórie. V prípade zmesí, musí ale rozsah výberu n byť veľmi veľký, aby asymptotická teória maximálnej vierohodnosti bola použiteľná. Pre prípady menšieho rozsahu dát sa preto často používa k aproximácii štandardných chýb $\hat{\boldsymbol{\Psi}}$ *bootstrap*.

konkrétne prípady, ako sú napr. zmesi t-rozdelení, zmesi dát so skupinovou štruktúrou, zmesi zobecnených lineárnych modelov a pod.

KLASIFIKAČNÁ VERZIA EM ALGORITMU

Niekedy môže byť klasické použitie EM algoritmu komplikované. Vtedy sa ponúka použiť zjednodušenie, ktorého podstata odpovedá modelu MLA. Vierohodnosť $L_c(\boldsymbol{\Psi})$ v tomto prípade maximalizujeme nielen vhodnou voľbou parametra $\boldsymbol{\Psi}$, ale aj zatriedením pozorovaní do zhlukov. Neznámy vektor $\mathbf{z} = (\mathbf{z}_1^T, \ldots, \mathbf{z}_n^T)^T$ tak odhadujeme spolu s parametrom $\boldsymbol{\Psi}$. Zjednodušenie spočíva v tom, že každé pozorovanie jednoznačne priradíme k nejakému zhluku (úplne klasifikujeme dáta do zhlukov) a na tomto potom priamo založíme výsledný odhad. Pri klasickom EM algoritme je však výsledný odhad priamo založený len na pravdepodobnosti príslušnosti pozorovaní k jednotlivým zhlukom (teda len na fuzzy klasifikácii).

Maximum $L_c(\boldsymbol{\Psi})$ môžeme v tomto prípade teoreticky nájsť vždy, a to prechádzaním všetkých prípustných hodnôt neznámeho vektora $\mathbf{z}$. Prakticky je to ale možné len v prípade malého n. Jednoduché riešenie ponúka nasledovná modifikácia EM algoritmu: v $(k+1)$-vej iterácii kroku E je namiesto aktuálneho odhadu aposteriórnej pravdepodobnosti $\tau_i(\mathbf{x}_j; \boldsymbol{\Psi}^{(k)})$, veličina z_{ij} nahradená jednotkou, ak platí

$$\pi_i^{(k)} f_i(\mathbf{x}_j; \boldsymbol{\theta}_i^{(k)}) \geq \pi_h^{(k)} f_h(\mathbf{x}_j; \boldsymbol{\theta}_h^{(k)}) \quad (h = 1, \ldots, g; h \neq i) \tag{1.34}$$

$(i = 1, \ldots, g; j = 1, \ldots, n)$, alebo nulou ak (1.34) neplatí. Takto v kroku $(k+1)$ získame odhady $\hat{\mathbf{z}}_1^{(k+1)}, \ldots, \hat{\mathbf{z}}_n^{(k+1)}$, na základe ktorých budú dáta úplne klasifikované. Odhady $\boldsymbol{\theta}_i^{(k+1)}$ a $\pi_i^{(k+1)}$ sú potom už priamočiare.

Táto alternatíva maximalizovania vierohodnosi však nevedie ku konzistentnému odhadu. V niektorých prípadoch môže síce poskytnúť rýchle a dobré riešenie, v situáciách keď komponenty nie sú dobre oddelené a proporcie komponentov sa líšia je odhad pomerne slabý. Vtedy ale tento odhad môže dobre poslúžiť ako inicializácia pre klasický EM algoritmus.

INÉ METÓDY VÝPOČTU MLE

Uveďme si ešte niektoré iné algoritmy používané na nájdenie MLE pre zmes hustôt. Pred objavením EM algoritmu sa používali a aj dnes sa ešte občas používajú v kombinácii s EM algoritmom predovšetkým numerické metódy optimalizácie, a to *metódy Newtonovho typu*. Jedna z nich, *Newton-Raphson-ova metóda*, používa na riešenie rovnice

$$\mathbf{S}(\mathbf{x}; \boldsymbol{\Psi}) = \partial \log L(\boldsymbol{\Psi}) / \partial \boldsymbol{\Psi} = \mathbf{0}$$

Taylorov rozvoj okolo aktuálneho odhadu $\boldsymbol{\Psi}^{(s)}$ parametra $\boldsymbol{\Psi}$. Toto vedie k aproximácii

$$\mathbf{S}(\mathbf{x}; \boldsymbol{\Psi}) \approx \mathbf{S}(\mathbf{x}; \boldsymbol{\Psi}^{(s)}) - \mathbf{I}(\boldsymbol{\Psi}^{(s)}; \mathbf{x})(\boldsymbol{\Psi} - \boldsymbol{\Psi}^{(s)}), \tag{1.35}$$

z ktorej potom obdržíme novú hodnotu parametra $\boldsymbol{\Psi}$, $\boldsymbol{\Psi}^{(s+1)}$, ako

$$\boldsymbol{\Psi}^{(s+1)} = \boldsymbol{\Psi}^{(s)} + \mathbf{I}^{-1}(\boldsymbol{\Psi}^{(s)}; \mathbf{x}) \mathbf{S}(\mathbf{x}; \boldsymbol{\Psi}^{(s)}). \tag{1.36}$$

Za platnosti určitých predpokladov o $L(\boldsymbol{\Psi})$ a v prípade dobrej počiatočnej inicializácie konverguje táto metóda k správnemu riešeniu veľmi rýchlo, čo je jej najväčšia výhoda. Problém je ale s voľbou vhodnej inicializácie. Nevýhodou metódy je aj nutnosť počítať v každom kroku algoritmu informačnú maticu $\mathbf{I}(\boldsymbol{\Psi}^{(s)};\mathbf{x})$[15].

1.9.3 Metódy založené na minimálnej vzdialenosti

Ďalší zaujímavý odhad parametrov modelu zmesi je odhad metódou minimálnej vzdialenosti (označím ho MDE). Pre tieto metódy síce zatiaľ nemáme tak dobre preskúmané algoritmy (alebo ich nemáme vôbec[16]), ale oproti metóde maximálnej vierohodnosti sú tu niektoré výhody. Tieto metódy sú robustnejšie (Cutler, Cordero-Braña, 1996) a umožňujú skonštruovať *konzistentný* odhad počtu komponentov v zmesi, čo je[17] hlavným nedostatkom metódy MLE. Odhad[18] hľadáme tak, že minimalizujeme (predom zvolenú) vzdialenosť empirického rozdelenia (alebo rozdelenia daného nejakým neparametrickým odhadom) od rozdelenia daného odhadnutou distribučnou funkciu zmesi. Formálne teda pre takýto odhad parametra $\boldsymbol{\Psi}$, $\hat{\boldsymbol{\Psi}}$, platí

$$\hat{\boldsymbol{\Psi}} \in \arg\min_{\boldsymbol{\Psi}} d(\hat{F}_n, F_{\boldsymbol{\Psi}}), \tag{1.37}$$

kde $\hat{F}_n$ je empirická distribučná funkcia, $F_{\boldsymbol{\Psi}}$ je distribučná funkcia odpovedajúca hustote zmesi (1.8) s neznámym parametrom $\boldsymbol{\Psi}$ a $d(F_1,F_2)$ je nejaká vzdialenosť distribučných funkcií F_1 a F_2. Najčastejšie sa používajú tieto vzdialenosti:

- *Kolmogorova-Smirnovova*

$$d(F_1,F_2) = \sup_{x\in\mathbb{R}^p} \|F_1(\mathbf{x}) - F_2(\mathbf{x})\|$$

- *Cramér-von-Misesova*

$$d(F_1,F_2) = \int_{\mathbb{R}^p} [F_1(\mathbf{x}) - F_2(\mathbf{x})]^2 \, d\{F_1(\mathbf{x}) + F_2(\mathbf{x})\}$$

- *Hellingerova*

$$d(F_1,F_2) = \sqrt{\int_{\mathbb{R}^p} \left(\sqrt{f_1} - \sqrt{f_2}\right)^2 d\boldsymbol{\mu}},$$

  kde F_1, F_2 sú distribučné funkcie s hustotami f_1, f_2 vzhľadom k $\boldsymbol{\mu}$.

Okrem týchto sa zvykne používať aj

[15] Náročnosť výpočtu tejto matice výrazne narastá s jej rastúcou dimenziou.

[16] Tým je myslené, že nemáme efektívny všeobecne použiteľný algoritmus. V niektorých jednoduchých prípadoch úlohu môžeme síce vyriešiť, ale v reálnych zložitejších príkladoch už nemusíme byť úspešní.

[17] S konzistenciou odhadu metódou MLE bol dlho problém a stále aj je, ale v niektorých dôležitých prípadoch sa konzistencia už podarila dokázať.

[18] Naďalej, ako v celej tejto kapitole, počet komponentov g považujeme za pevný, takže nie je súčasťou odhadovaného parametra modelu.

- *Kullback-Leiblerova vzdialenosť (divergencia)*

$$K(F_1,F_2) = \int_{\mathbb{R}^p} f_1(\mathbf{x}) \log \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \mu(d\mathbf{x}),$$

kde F_1, F_2 sú distribučné funkcie s hustotami f_1, f_2 vzhľadom k $\boldsymbol{\mu}$.

Pojem divergencia sa používa preto, lebo Kullback-Leibrerova vzdialenosť nie je vzdialenosť v pravom slova zmysle. Nie je totiž symetrická a nespĺňa trojuholníkovú nerovnosť. K jej základným vlastnostniam patrí, že $K(F_1,F_2) \geq 0$ a $K(F_1,F_2) = 0$ práve vtedy, ak $f_1 = f_2$ *s.i.* vzhľadom k $\boldsymbol{\mu}$.

Ako demonštráciu tejto metódy odhadu si uveďme **algoritmus HMIX** (Cutler, Cordero-Braña, 1996), ktorý slúži k hľadaniu odhadov parametrov zmesi metódou minimálnej Hellingerovej vzdialenosti. Nech $\hat{f}_n$ je neparametrický odhad hustoty založený na $\mathbf{x}_1, \ldots, \mathbf{x}_n$ (napr. jadrový). Potom odhad $\hat{\boldsymbol{\Psi}}$ zvolíme tak, aby

$$\hat{\boldsymbol{\Psi}} \in \arg\min_{\boldsymbol{\Psi}\in\boldsymbol{\Omega}} \sqrt{\int_{\mathbb{R}^p} \left(\sqrt{\hat{f}_n(\mathbf{x})} - \sqrt{\sum_{i=1}^{g} \pi_i f_i(\mathbf{x};\boldsymbol{\theta}_i)} \right)^2 \mu(d\mathbf{x})},$$

čo je zrejme ekvivalentné s hľadaním maxima funkcie

$$G(\boldsymbol{\Psi}) = \int_{\mathbb{R}^p} \sqrt{\hat{f}_n(\mathbf{x}) \sum_{i=1}^{k} \pi_i f(\mathbf{x};\boldsymbol{\theta}_i)} \mu(d\mathbf{x}).$$

Ďalej platí

$$G(\boldsymbol{\Psi}) = \sum_{i=1}^{k} \sqrt{\pi_i} \cdot \int_{\mathbb{R}^p} \sqrt{\frac{\pi_i f(\mathbf{x};\boldsymbol{\theta}_i)}{\sum_{l=1}^{k} \pi_l f(\mathbf{x};\boldsymbol{\theta}_l)}} \cdot \sqrt{f(\mathbf{x};\boldsymbol{\theta}_i)\hat{f}_n(\mathbf{x})} \mu(d\mathbf{x}). \tag{1.38}$$

Ak by prvá odmocnina integrandu (1.38) nezávisela na $\boldsymbol{\Psi}$, bolo by hľadanie maxima oveľa jednoduchšie, pretože by sme mohli najskôr maximalizovať každý sčítanec zvlášť cez $\boldsymbol{\theta}_i$ a potom súčet cez $\boldsymbol{\pi}$. Toto je vlastne základná myšlienka HMIX algoritmu, podrobnosti ku ktorému možno nájsť v Nemeček (2004).

1.9.4 Bayesovský prístup

Bayesovské metódy dnes už v štatistike majú svoje pevné miesto a často sú pri riešení praktických a komplikovaných problémov uprednostňované pred metódami klasickými. Ich význam výrazne narástol pomerne nedávnym objavom *metód MCMC* (Markov Chain Monte Carlo), ktoré umožňujú efektívne prakticky implementovať Bayesovské metódy aj v náročnejších prípadoch ako sú aj zmesové modely pravdepodobnostných distribúcií. V prípade zmesí sa dokonca metódam MCMC pri konštrukcii Bayesovského odhadu nevyhneme.

BAYESOVSKÝ ODHAD PRE ZMESI HUSTÔT

Uvažujme model so zmesou hustôt (1.6) a jemu odpovedajúci vektor neznámych parametrov $\boldsymbol{\Psi}$ definovaný v (1.7). Nech $p(\boldsymbol{\Psi})$ je predpokladaná *apriórna* hustota vektora parametrov $\boldsymbol{\Psi}$. Potom pre *aposteriórnu* hustotu tohoto vektora pri pozorovanom $\mathbf{x} = (\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T)^T$ platí

$$\begin{aligned} p(\boldsymbol{\Psi}|\mathbf{x}) &= C^{-1} L(\boldsymbol{\Psi}) p(\boldsymbol{\Psi}) \\ &= C^{-1} \sum_{\mathbf{z}} L_c(\boldsymbol{\Psi}) p(\mathbf{z}|\boldsymbol{\Psi}) p(\boldsymbol{\Psi}), \end{aligned} \tag{1.39}$$

kde $p(\mathbf{z}|\boldsymbol{\Psi})$ označuje podmienenú hustotu $\mathbf{Z}$ pri danom $\boldsymbol{\Psi}$ a normalizačná konštanta C je daná vzťahom

$$C = \int \sum_{\mathbf{z}} L_c(\boldsymbol{\Psi}) p(\mathbf{z}|\boldsymbol{\Psi}) p(\boldsymbol{\Psi}) d\boldsymbol{\Psi},$$

kde $L_c(\boldsymbol{\Psi})$ je logaritmická vierohodnosť pre kompletný datový vektor podľa zavedenej notácie. V (1.39) sčítame cez všetky možné hodnoty $\mathbf{z}$, ktoré definujú príslušnosť ku komponentu pre $\mathbf{x}_j$ $(j = 1, \ldots, n)$. Pri vhodnej voľbe apriórnej hustoty $p(\boldsymbol{\Psi})$ (napr. prirodzený konjugovaný systém rozdelení, ak existuje) je možné potom aposteriórnu hustotu pre $\boldsymbol{\Psi}$ nájsť v uzavretej forme. Jej priame použitie je ale uskutočniteľné len v prípade malého rozsahu výberu, a preto sa používajú metódy MCMC.

VOĽBA KONJUGOVANÝCH APRIORNÝCH ROZDELENÍ

Predpokladajme, že hustoty komponentov uvažovanej zmesi (1.6) patria do tej istej exponenciálnej rodiny rozdelení, teda $f_i(\mathbf{x}_j; \boldsymbol{\theta}_i) = f(\mathbf{x}_j; \boldsymbol{\theta}_i)$, kde

$$f(\mathbf{x}_j; \boldsymbol{\theta}_i) = \exp\left\{\boldsymbol{\theta}_i^T \mathbf{x}_j - b(\boldsymbol{\theta}_i) + c(\mathbf{x}_j)\right\}. \tag{1.40}$$

K tejto rodine existuje konjugovaný systém hustôt pre apriórne rozdelenie $\boldsymbol{\theta}_i$, ktoré má tvar

$$p(\boldsymbol{\theta}_i; \boldsymbol{\omega}_i, \gamma_i) \propto \exp\left\{\boldsymbol{\theta}_i^T \boldsymbol{\omega}_i - \gamma_i b(\boldsymbol{\theta}_i)\right\}, \tag{1.41}$$

kde $\boldsymbol{\omega}_i$ vektor reálnych konštánt a γ_i je skalárna konštanta $(i = 1, \ldots, g)$. Konjugované apriórne rozdelenie pre vektor proporcií $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_g)^T$ je Dirichletovo rozdelenie $D(\alpha_1, \ldots, \alpha_g)$ s hustotou

$$p_D(\boldsymbol{\pi}) = \Gamma\left(\sum_{i=1}^{g} \alpha_i - g\right) \prod_{i=1}^{g} \pi_i^{\alpha_i - 1} / \Gamma(\alpha_i), \tag{1.42}$$

ktoré je zobecnením beta rozdelenia $B(\alpha_1, \alpha_2)$. Pre takéto apriórne rozdelenie vektora parametrov $\boldsymbol{\Psi}$ je aposteriórna hustota

$$p(\boldsymbol{\Psi}|\mathbf{x}) \propto \sum_{\mathbf{z}} p_D(\alpha_1 + n_1, \ldots, \alpha_g + n_g) \prod_{i=1}^{g} p(\boldsymbol{\theta}_i; \boldsymbol{\omega}_i + n_i \bar{\mathbf{x}}_i, \gamma_i + n_i), \tag{1.43}$$

kde $n_i = \sum_{j=1}^{n} z_{ij}$ a $\bar{\mathbf{x}}_i = \sum_{j=1}^{n} z_{ij}\mathbf{x}_j/n_i$.

Hoci aposteriórnu pravdepodobnosť pre odhadovaný parameter $\boldsymbol{\Psi}$ sa nám podarilo vyjadriť v uzavretej forme (1.43), čas potrebný k vyčísleniu (1.43) je príliš veľký na praktické použitie, a to aj pre umiernený rozsah výberu.

MCMC

Riešenie metódou MCMC je založené na konštrukcii ergodického Markovovho reťazca, ktorého stacionárne rozdelenie sa rovná práve aposteriórnemu rozdelniu parametra, ktorý odhadujeme (v našom prípade $\boldsymbol{\Psi}$). Existujú dva hlavné prístupy, ako zostrojiť takýto Markovov reťazec:

- **Gibbsov výberový plán** - prvky Markovovho reťazca simulujeme priamo z podmieneného rozdelenia podvektora $\boldsymbol{\Psi}$ pri daných zvyšných parametroch v $\boldsymbol{\Psi}$ (a pozorovaných dátach $\mathbf{x}$),
- **Metropolis-Hastingsov algoritmus** - prvky reťazca simulujeme z vhodne navrhnutého rozdelenia, pričom každú vygenerovanú hodnotu prijmeme, resp. zamietneme do Markovovho reťazca s vopred definovanou pravdepodobnosťou.

Získame tak výber z hľadaného Markovovho reťazca $\boldsymbol{\Psi}^{(1)},\ldots,\boldsymbol{\Psi}^{(N)}$ o rozsahu N. Od určitého (dostatočne veľkého) N_1 potom prvky tohoto reťazca $\boldsymbol{\Psi}^{(k)}$ $(k = N_1+1,\ldots,N)$ aproximujú výber z hľadaného aposteriórneho rozdelenia pre $\boldsymbol{\Psi}$.

V prípade modelu so zmesou hustôt pri implementácii MCMC metód generujeme vektor $\mathbf{z}$, indikujúci príslušnosť ku komponentom zmesi, a na základe neho potom upravujeme hodnotu $\boldsymbol{\Psi}$ (takže v tomto prípade produkujeme vlastne dva reťazce, reťazec chýbajúcich dát $\mathbf{z}$ a reťazec parametrov $\boldsymbol{\Psi}$).

Simulovaný výber $\boldsymbol{\Psi}^{(N_1+1)},\ldots,\boldsymbol{\Psi}^{(N)}$ môže byť samozrejme použitý na aproximáciu ľubovoľnej dobre definovanej aposteriórnej veličiny, ako napr. nejakej funkcie $\boldsymbol{\Psi}$. Bodový odhad funkcie a získame pomocou ergodického priemeru

$$E\{a(\boldsymbol{\Psi})|\mathbf{x}\} \approx \sum_{k=N_1+1}^{N} \frac{a(\boldsymbol{\Psi}^{(k)})}{N-N_1}$$

a intervalový odhad získame zase pomocou výberových kvantilov pre $a(\boldsymbol{\Psi}^{(k)})$ $(k = N_1+1,\ldots,N)$.

MCMC PRE EXPONENCIÁLNU RODINU ROZDELENÍ

Teraz popíšem v krokoch MCMC metódu s Gibbsovým výberovým plánom v prípade konjugovaných apriórnych rozdelení (1.41) a (1.42) pre komponenty zo všeobecnej exponenciálnej rodiny (1.40).

1. Simuluj

$$\boldsymbol{\pi} \sim D(\alpha_1+n_1,\ldots,\alpha_g+n_g) \text{ a } \boldsymbol{\theta}_i \sim p(\boldsymbol{\theta};\boldsymbol{\omega}_i+n_i\bar{\mathbf{x}}_i,\gamma_i+n_i) \quad (i=1,\ldots,g).$$

2. Simuluj

$$\mathbf{Z}_j \sim Mult_g(1,\boldsymbol{\tau}_j) \quad (j = 1,\ldots,n),$$

kde $\boldsymbol{\tau}_j = (\tau_1(\mathbf{x}_j;\boldsymbol{\Psi}),\ldots,\tau_g(\mathbf{x}_g;\boldsymbol{\Psi}))^T$, $\tau_i(\mathbf{x}_j;\boldsymbol{\Psi}) = \pi_i f(\mathbf{x}_j;\boldsymbol{\theta}_i)/\sum_{h=1}^{g} \pi_h f(\mathbf{x}_j;\boldsymbol{\theta}_h)$.

3. Aktualizuj n_i a $\bar{\mathbf{x}}_i$ $(i = 1,\ldots,g)$.

VOĽBA INÝCH APRIÓRNYCH ROZDELENÍ

Pri voľbe apriórneho rozdelenia sa snažíme nepoužívať príliš silnú apriórnu informáciu, aby sme ňou príliš neovplyvnili výsledný odhad. Niektorí autori sa preto zaoberali apriormi s malou informáciou, ktoré by poprípade mohli byť aj úple nezávislé od dát (Richardson, Green, 1997b). Ukázalo sa však, že použitie úplne neinformatívnych apriórnych rozdelení (napr. *princíp neurčitosti*) neumožňuje konštrukciu vhodného aposteriórneho rozdelenia. To je spôsobené tým, že vždy existuje pravdepodobnosť, že pre jednu alebo viac komponentov nebude alokované ani jedno pozorovanie, a teda dáta nebudú poskytovať dostatočnú informáciu pre odhad všetkých parametrov. Taktiež sa ukázalo, že v prípade zmesí je aposteriórna hustota nevhodná aj pre *Jeffreysovu apriórnu hustotu.*

Otázke vhodného apriorného rozdelenia pre aplikáciu Bayesovských metód v prípade modelu zmesi bola venovaná veľká pozornosť, napr. Roeder, Wasserman (1997).

1.10 Dimenzia robí problémy

Je známe, že rastúca dimenzia pozorovaní úlohu štatistika neúmerne komplikuje, hovoríme o tzv. *prekliatí dimenzie* (curse of dimensionality). Uvažujme zmes viacrozmerných normálnych rozdelení s rôznymi variančnými maticami. Každá variančná matica $\boldsymbol{\Sigma}_i$ $(i = 1,\ldots,g)$ potom obsahuje $\frac{1}{2}p(p+1)$ parametrov. Počet parametrov pre každý komponent takéhoto modelu je teda rádu $O(p^2)$, čo už je pri trochu väčšej dimenzii veľký problém.

Uvažujme parametrizáciu tohoto modelu založenú na spektrálnom rozklade jednotlivých variančných matíc

$$\boldsymbol{\Sigma}_i = \sum_{v=1}^{p} \lambda_{iv}\mathbf{a}_{iv}\mathbf{a}_{iv}^T, \tag{1.44}$$

kde $\mathbf{a}_{iv},\ldots,\mathbf{a}_{ip}$ sú vlastné vektory prislúchjúce vlastným číslam $\lambda_{i1} \geq \lambda_{i2}, \geq \ldots \lambda_{ip} > 0$ matice $\boldsymbol{\Sigma}_i$ $(i = 1,\ldots,g)$. Rovnosť (1.44) môžeme vyjadriť aj ako

$$\boldsymbol{\Sigma}_i = \lambda_i\mathbf{A}_i\boldsymbol{\Lambda}_i\mathbf{A}_i^T \tag{1.45}$$

kde $\mathbf{A}_i = (\mathbf{a}_{i1},\ldots,\mathbf{a}_{ip})$ je (ortogonálna) matica vlastných vektorov matice $\boldsymbol{\Sigma}_i$. Konvencie pre normalizáciu λ_i a $\boldsymbol{\Lambda}_i$ zahrňujú buď $\lambda_i = \lambda_{i1}$ (najväčšie vlastné číslo $\boldsymbol{\Sigma}_i$), pre ktoré $\boldsymbol{\Lambda}_i = diag(1,\lambda_{i2}/\lambda_{i1},\ldots,\lambda_{ip}/\lambda_{i1})$, alebo $|\boldsymbol{\Lambda}_i| = 1$ pre ktoré $\lambda_i = |\boldsymbol{\Sigma}_i|^{1/p}$ a $\boldsymbol{\Lambda}_i = diag(\lambda_{i1}/\lambda_i,\ldots,\lambda_{i1}/\lambda_i)$. Parameter λ_i riadi potom objem i-teho zhluku (zhluk prislúchajúci i-temu komponentu), $\boldsymbol{\Lambda}_i$ jeho tvar a $\mathbf{A}_i$ jeho orientáciu.

Tento rozklad sa potom používa k redukcii počtu parametrov vo viacrozmernom prípade, a to zavedením určitých reštrikcií na $\mathbf{A}_i$,$\mathbf{\Lambda}_i$ a λ_i. Podmienka $\mathbf{A}_i = \mathbf{A}$ $(i = 1,\ldots,g)$ napríklad znamená, že všetky zhluky v populácii majú rovnakú orientáciu.

1.11 Uplatnenie modelu so zmesou hustôt

Zmesové distribučné modely našli obrovské uplatnenie v prírodných, ale aj sociálnych vedách. Boli úspešne aplikované napr. v biológii, genetike, medicíne, psychiatrii, astronómii, ekonomike, marketingu, strojárstve a iných ďalších oblastiach. V tejto časti by som chcel trochu predstaviť niektoré zaujímavé prípady použitia zmesi.

DÁTA OBSAHUJÚCE SPOJITÉ AJ KATEGORICKÉ VELIČINY

Uvažujme model (1.6) pre dáta $\mathbf{x} = (\mathbf{x}_1^T,\ldots,\mathbf{x}_n^T)^T$, ktorých niektoré zložky $\mathbf{x}_j$ $(j = 1,\ldots,n)$ sú kategorické a niektoré spojité (*mixed-mode data*). Najjednoduchší prípad nastáva, keď predpokladáme, že kategorické zložky sú navzájom nezávislé a že zložky reprezentujúce spojité náhodné veličiny pochádzajú z viacrozmerného normálneho rozdelenia. V takomto prípade pre model zmesi používame tzv. *model polohy* (Jorgensen, Hunt, 1996; Lawrence, Krzanowski, 1996).

Predpokladajme, že p_1 z p zložiek náhodnej veličiny $\mathbf{X}_j$ je kategorických, pričom q-ta kategorická veličina nadobúda m_q rôznych hodnôt. Potom existuje $m = \prod_{q=1}^{p1} m_q$ rôznych možností realizácie všetkých p_1 kategirických veličín. V modeli polohy sa potom nahradia všetky tieto kategorické veličiny jednou náhodnou veličinou $\mathbf{X}_j^{(1)}$ s multinomickým rozdelením $Mult(1,m)$. Predpokladajme ďalej, že za podmienky $(\mathbf{x}_j^{(1)})_s = 1$ a príslušnosti j-teho pozorovania do i-teho komponentu je rozdelenie zvyšných $p - p_1$ spojitých veličín (označujeme ich spolu $\mathbf{X}_j^{(2)}$) normálne so strednou hodnotou $\boldsymbol{\mu}_{is}$ a variančnou maticou $\mathbf{\Sigma}_i$, ktorá je rovnaká pre všetky s (teda pre všetky možné kombinácie realizácie p_1 kategorických náhodných veličín). Podmienené normálne rozdelenie navádza k priamej implementácii algoritmu EM. Nech p_{is} označuje pravdepodobnosť, že $\mathbf{X}_j^{(1)})_s = 1$ za podmienky, že $\mathbf{X}_j$ patrí do i-teho komponentu, a nech δ_{js} je jedna alebo nula, vzhľadom k tomu či $(\mathbf{x}_j^{(1)})_s$ je jedna alebo nula. Potom iterácia EM algoritmu pre krok $(k+1)$ vyzará nasledovne

$$\begin{aligned}
\pi_i^{(k+1)} &= \sum_{s=1}^{m}\sum_{j=1}^{n} \delta_{js}\tau_{ijs}^{(k)}/n, \\
p_{is}^{(k+1)} &= \sum_{j=1}^{n} \delta_{js}\tau_{ijs}^{(k)} / \sum_{s=1}^{m}\sum_{j=1}^{n} \delta_{js}\tau_{ijs}^{(k)}, \\
\boldsymbol{\mu}_{is}^{(k+1)} &= \sum_{j=1}^{n} \delta_{js}\tau_{js}^{(k)}\mathbf{x}_j^{(2)} / \sum_{j=1}^{n} \delta_{js}\tau_{ijs}^{(k)}
\end{aligned}$$

a

$$\mathbf{\Sigma}_i^{(k+1)} = \sum_{s=1}^{m}\sum_{j=1}^{n} \delta_{js}\tau_{ijs}^{(k)}(\mathbf{x}_j^{(2)} - \boldsymbol{\mu}_{is}^{(k+1)})(\mathbf{x}_j^{(2)} - \boldsymbol{\mu}_{is}^{(k+1)})^T / \sum_{s=1}^{m}\sum_{j=1}^{n} \delta_{js}\tau_{ijs}^{(k)},$$

kde

$$\tau_{ijs}^{(k)} = \frac{\pi_i^{(k)} p_{is}^{(k)} \phi(\mathbf{x}_j^{(2)}; \boldsymbol{\mu}_{is}^{(k)}, \boldsymbol{\Sigma}_i^{(k)})}{\sum_{h=1}^{g} \pi_h^{(k)} p_{hs}^{(k)} \phi(\mathbf{x}_j^{(2)}; \boldsymbol{\mu}_{hs}^{(k)}, \boldsymbol{\Sigma}_h^{(k)})}$$

pre $s = 1,\ldots,m;\ i = 1,\ldots,g$.

OVERDISPERSSION

Zaoberajme sa teraz prípadom diskrétnych dát modelovaných Poissonovým rozdelením. Silným obmedzením v tomto prípade je fakt, že rozptyl a stredná hodnota Poissonovho rozdelenia sú rovnaké. V praxi je však častokrát pozorovaný rozptyl u dát väčší ako pozorovaná stredná hodnota. Hovoríme, že dáta sú *prerozptýlené* (*overdispersed*) a tento problém označujeme termínom *overdispersion*[19]. Jednoduché Poissonove rozdelenie v tomto prípade popisuje dáta nedostatočne, a pokiaľ sa pri hľadaní lepšieho modelu chceme vyhnúť neparametrickému odhadu, ponúka sa ako vhodná alternatíva použiť práve zmes Poissonových rozdelení.

Niekedy sa používa spojitá zmes Poissonových rozdelení, kde poissonova stredá hodnota λ je modelovaná nejakou spojitou distribúciou $H(\lambda)$. Potom hustota pozorovanej náhodnej veličiny X (uvažujeme jednorozmerný prípad) je modelovaná ako

$$f(x) = \int_0^\infty \frac{e^{-\lambda}\lambda^x}{x!} I_A(x) dH(\lambda), \tag{1.46}$$

kde $A = 0,1,2,\ldots$ je množina nezáporných celých čísel. Najčastejšou voľbou $H(\lambda)$ v (1.46) je gama rozdelenie $Ga(\alpha,\beta)$ s hustotou

$$h(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} I_{[0,\infty)}(\lambda) \quad \lambda,\beta > 0,$$

kde $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx$ je gama funkcia. Toto potom vedie k hustote

$$f(x;\alpha,\beta) = \binom{x+\alpha-1}{x} \left(\frac{\beta}{\beta+1}\right)^\alpha \left(\frac{1}{\beta+1}\right)^x I_A(x),$$

čo je vlastne hustota *negatívne binomického rozdelenia* $NBi(\alpha,\beta/(\beta+1))$. Keď označíme strednú hodnotu tohoto rozdelenia μ, potom

$$\mu = \frac{\alpha}{\beta},$$

pričom jeho rozptyl je

$$var(X) = \frac{\alpha}{\beta}\left(\frac{\beta+1}{\beta}\right) = \mu + k\mu^2,$$

kde $k = 1/\alpha$. Z tohoto vidíme, že tento dvojparametrický model už umožňuje, aby rozptyl modelovaných dát bol väčší ako ich stredná hodnota.

[19] Analogicky samozrejme existuje prípad *underdispersion*, ktorý sa však vyskytuje zriedkavo.

V súvislosti s Poissonovým rozdelení sa stretávame ešte s jedným zaujímavým prípadom, pri ktorom sa dobre uplatnilo použitie modelu zmesi. Ide o prípad, keď v diskrétnych dátach modelovaných Poissonovým rozdelením prevažuje výskyt núl. Tento prípad sa v odborej literatúre vyskytuje pod názvom *Zero-inflated Poisson model* (ZIP) a má už pomerne dlhú históriu (Johnson, Kotz, 1969). Situácia ZIP modelu sa dá pekne demonštrovať na príklade výrobného procesu, počas ktorého dochádza k určitým defektom, ktorých výskyt nás zaujíma. Predpokladajme, že nejaké nepatrné, nepozorované zmeny v prostredí zapríčiňujú, že výrobný proces sa náhodne pohybuje medzi akýmsi „perfektným stavom", kedy k defektom vo výrobe dochádza len extrémne zriedkavo, a „stavom neperfektným", počas ktorého sú defekty vo výrobe možné, ale nie nevyhnutné. Zvýšený výskyt núl (teda, že k defektu v danom časovom úseku nedošlo) v dátach potom interpretujeme existenciou perfektného stavu. Výskyt defektov vo výrobe tak môžeme modelovať zmesou degenerovaného rozdelenia s hodnotou 1 pre $x = 0$ (prípad perfektného stavu) a Poissonovho rozdelenia (prípad neperfektného stavu), takže potom platí

$$\begin{aligned} P(x=0) &= 1-p+pe^{-\lambda} \\ P(x=q) &= p\,\frac{e^{-\lambda}\lambda^q}{q!}, \quad q=1,2,\ldots . \end{aligned}$$

Zmesi Poissonových rozdelení zohrávajú dôležitú úlohu napríklad aj pri zostrojovaní chorobových máp (*disease mapping*), čo je dôležitá oblasť epidemiológie.

ZOBECNENÉ LINEÁRNE MODELY (GLM)

V jednorozmernom prípade zobecnených lineárnych modelov má logaritmus hustoty (log hustota) vysvetľovanej premennej $\mathbf{Y}_j$ tvar

$$\log f(y_j;\theta_j,\kappa) = m_j\kappa^{-1}(\theta_j y_j - b(\theta_j)) + c(y_j;\kappa), \tag{1.47}$$

kde θ_j je prirodzený alebo kanonický parameter, κ je disperzný parameter a m_j je apriórna váha. Stredná hodnota a rozptyl sú $E(Y_j) = \mu_j = b'(\theta_j)$, $var(Y_j) = \kappa b''(\theta_j)$. GLM potom predpokladá

$$\eta_j = h(\mu_j) = \mathbf{x}_j^T\boldsymbol{\beta}, \tag{1.48}$$

kde $\mathbf{x}_j$ je vektor vysvetľujúcich premenných pre j-tu vysvetľovanú premennú y_j, $\boldsymbol{\beta}$ je vektor neznámych parametrov a $h(\cdot)$ je monotónna funkcia nazývaná *linkovacia funkcia*. V prípade, že κ je známe, (1.47) patrí medzi hustoty (regulárneho) exponenciálneho rozdelenia. Poissonove a binomické rozdelenie majú $\kappa = 1$.

Zmes g komponentov rozdelení GLM v proporciách $\pi_1,\ldots,\pi_g$ je definovaná ako

$$f(y_j;\boldsymbol{\Psi}) = \sum_{i=1}^{g}\pi_i f(y_j;\theta_{ij},\kappa_i), \tag{1.49}$$

kde pre fixný disperzný parameter κ_i máme

$$\log f(y_j;\theta_{ij},\kappa_i) = m_j\kappa_i^{-1}(\theta_{ij}y_j - b_i(\theta_{ij})) + c_i(y_j;\kappa_i)$$

pre $i = 1, \ldots, g$. Pre i-ty komponent GLM je μ_{ij} stredná hodnota Y_j, $h_i(\mu_{ij})$ je linkovacia funkcia a $\eta_i = h_i(\mu_{ij}) = \boldsymbol{\beta}_i^T \mathbf{x}_j$ je lineárny prediktor $(i = 1, \ldots, g)$. Tieto modely sa používajú v oblasti *machine learning*, kde sú známe pod menom *mixtures-of-experts* modely.

V prípadoch, kde získanie odhadu pre model so zmesou hustôt nie je priamočiare, sa niekedy snažíme daný problém formulovať ako zmes GLM, ktorú by sme už vedeli vyriešiť pomocou EM algoritmu pre zmesi GLM. Takým príkladom sú dáta, ktoré obsahujú len binárne premenné (pozorovanie $\mathbf{x}_j = (x_{1j}, \ldots, x_{pj})^T$ obsahuje p binárnych premenných). Základný predpoklad pre analýzu zmesovej štruktúry takýchto dát je ich vzájomná podmienená nezávislosť vzhľadom k príslušnosti ku komponentu zmesi. Hustota i-teho komponentu zmesi potom je

$$f(\mathbf{x}_j; \boldsymbol{\theta}_i) = \prod_{v=1}^{p} \theta_{vi}^{x_{vj}} (1 - \theta_{vi})^{1 - x_{vj}},$$

kde $\boldsymbol{\theta}_i = (\theta_{1i}, \ldots, \theta pi)^T$ a θ_{vi} je podmienená pravdepodobnosť, že $X_{vj} = 1$ $(v = 1, \ldots, p)$ za podmienky, že patrí do i-teho komponentu zmesi. S týmto modelom sa často stretávame v sociálnych vedách a bioštatistike (Everitt, 1984).

ZMESI T-ROZDELENÍ

Ako som už v sekcii 1.6.3 naznačil, t-rozdelenie možno považovať za akúsi alternatívu normálneho rozdelenia s ťažšími chvostami. Chvosty normálneho rozdelenia totiž v praxi veľakrát nie sú postačujúce (príliš rýchlo konvergujú k nule). Ako robustnejší prístup modelovania normálnych zmesí bolo preto navrhnuté použitie zmesi (aj viacrozmerných) t-rozdelení (McLachlan, Peel, 1998).

Problém odľahlých pozorovaní je v prípade zmesí však obšírnejší a jeho náročnosť rastie s rastúcou dimenzou dát. Niekedy sa odľahlé pozorovania modelujú rovnomerným rozdelením na zvolenej množine A. Proporcie zmesi $\boldsymbol{\pi}$ sa rozšíria o π_0 (pravdepodobnosť, že pozorovanie je odľahlé). Vierohodnosť pre model (1.6) má potom tvar

$$L(\boldsymbol{\theta}) = \prod_{j=1}^{n} \left[\frac{\pi_0}{\lambda(A)} + \sum_{i=1}^{g} \pi_i f_i(\mathbf{x}_j; \boldsymbol{\theta}_i) \right]. \tag{1.50}$$

Ako robustná varianta odhadu v modeli so zmesou hustôt sa používal M-odhad strednej hodnoty a variančnej matice v M-kroku EM algoritmu (Campbell, 1984). Nedávno bola pre robustné odhady v modeloch zmesí vypracovaná metodológia váženej vierohodnosti (Markatou, 1998).

Viacrozmerné t-rozdelenie patrí medzi širšiu triedu *elipticky symetrických rozdelení* s dodatočným parametrom nazývaným *stupeň voľnosti* ν. Keď sa tento blíži k nekonečnu, t-rozdelnie sa blíži k normálnemu rozdeleniu. Môžeme teda na neho nahliadať ako na parameter robustnosti. Hustota (viacrozmerného) t-rozdelenia má tvar

$$f(\mathbf{x}_j; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = \frac{\Gamma(\frac{\nu+p}{2}) |\boldsymbol{\Sigma}|^{-\frac{1}{2}}}{(\pi\nu)^{\frac{1}{2}} \Gamma(\frac{\nu}{2}) (1 + \delta(\mathbf{x}_j, \boldsymbol{\nu}; \boldsymbol{\Sigma})/\nu)^{\frac{1}{2}(\nu+p)}},$$

kde

$$\delta(\mathbf{x}_j,\boldsymbol{\nu};\boldsymbol{\Sigma}) = (\mathbf{x}_j - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_j - \boldsymbol{\mu})$$

označuje Mahalanobisovu vzdialenosť medzi $\mathbf{x}_j$ a $\boldsymbol{\mu}$ (s variančnou maticou $\boldsymbol{\Sigma}$). Pri aplikácii EM algoritmu narazíme však na jednu komplikáciu. Pre odhad parametra robustnosti totiž neexistuje v kroku M explicitné riešenie rovnice (1.33), takže je nutné použiť nejakú iteratívnu metódu. Týmto vlastne v rámci EM algoritmu používame ďalší iteratívny algoritmus, čím dostávame pomerne komplikovanú procedúru. Je preto rozumnejšie stupeň voľnosti predom nejak zvoliť a neodhadovať ho spolu s ostatnými parametrami.

ZMESI V ANALÝZE PORÚCH

Je to len nedávno, čo sa zmesové distribučné modely začali aplikovať v *analýze prežitia* a *analýze spoľahlivosti*. Pekný príklad ponúka napr. článok Blackstone a kol. (1986), ktorý sa zaoberá operáciami s otvoreným srdcom. Autori uvádzajú, že riziko úmrtia prechádza troma fázami: ranná fáza, počas ktorej je riziko relatívne vysoké, stredná fáza s konštantným rizikom a posledná fáza, pri ktorej sa riziko začína zväčšovať s vekom pacienta. Tieto fázy sa prekrývajú v čase, a preto nemôžu byť dostatočne modelované pomocou troch oddelených modelov (oddelené časové úseky). Model so zmesou troch rozdelení v tomto prípade ponúka šikovný spôsob modelovania prežitia. Najviac používaným rozdelením v tejto oblasti (tzv. životné rozdelenie) je *Weibullovo rozdelenie* $W(\alpha,\kappa)$ s hustotou

$$f(x;\alpha,\kappa) = \alpha\kappa x^{\kappa-1}\exp\{-\alpha x^{\kappa}\}\, I_{(0,\infty)}(x), \quad \alpha,\, \kappa > 0.$$

Je to veľmi flexibilné rozdelenie s jednoduchým tvarom hustoty, *krivkou prežitia*

$$S(t,\boldsymbol{\theta}) = \exp\{-(\alpha t)^{\kappa}\}\, I_{(0,\infty)}(t),$$

kde $\boldsymbol{\theta} = (\alpha,\kappa)^T$ a jednoduchou jemu príslušnou hazardnou funkciou

$$h(t,\boldsymbol{\theta}) = \alpha^{\kappa}\kappa t^{\kappa-1} I_{(0,\infty)}(t).$$

Vzhľadom na hodnotu parametera κ (parameter tvaru) môže byť hazardná funkcia monotónne rastúca, klesajúca alebo konštantná. Ak tento parametr položíme rovný jednej, dostaneme hustotu *exponenciálneho rozdelenia*, ktoré je takisto veľmi rozšírené a používa sa napr. v prípade, keď empiricky pozorovaná hazardná funkcia klesá v čase.

ZMESI V PRÍPADE SMEROVÝCH DÁT

Zmesi našli uplatnenie aj v zaujímavej oblasti tzv. smerových dát. Ide o dáta, ktoré vyjadrujú nejaký smer. Príkladom môže byť napr. analýza štruktúry kamennej hmoty alebo jednoducho skaly. Tá zväčša obsahuje rôzne trhliny a fraktúry, ktorých orientácia zohráva dôležitú úlohu pri ďalšom vývoji tejto hmoty. Uplatnenie nachádzame napr. v baníckom priemysle, kde je extrémne dôležitá štruktúra povalu tunela. Vznik fraktúr je spôsobený nejakou vonkajšou silou, preto sa často tieto fraktúry formujú približne v

rovnobežnom smere. Takýchto rodín rovnobežných trhlín môže byť na jednej kamennej mase samozrejme prítomných viac. Dáta sú zaznamenané v tvare smerových vektorov jednotlivých fraktúr, $\mathbf{x} = (x_1,x_2,x_3)^T$. Autor používa k modelovaniu dát zmes *Kentových rozdelení.* Hustota tohoto rozdelenia má tvar

$$f_K(\mathbf{x};\boldsymbol{\theta}) = C_K \exp\left[\kappa(\mathbf{x}^T\boldsymbol{\omega_1}) + \beta\left\{(\mathbf{x}^T\boldsymbol{\omega}_2)^2 - (\mathbf{x}^T\boldsymbol{\omega}_3)^2\right\}\right],$$

kde

$$C_K = (2\pi)^{-1}\exp(-\kappa)(\kappa^2 - 4\beta^2)^{1/2} \tag{1.51}$$

a $\boldsymbol{\theta} = (\kappa,\beta,\boldsymbol{\omega}_1,\boldsymbol{\omega}_2,\boldsymbol{\omega}_3)$ je vektor parametrov. Podrobnejší popis tohoto rozdelenia, aj popis EM algortimu pre tento prípad možno nájsť v Peel a kol. (2001). Niektorí autori k analýze smerových dát používajú zmes *Fisherových* alebo *von-Misesových* rozdelení[20].

[20]Fisherovo rozdelenie je zobecnením von-Misesovho rozdelenia z kruhu na sféru a špeciálnym prípadom Kentovho rozdelenia pre prípad $\beta = 0$.

Kapitola 2

Počet komponentov

V predchádzajúcej kapitole sme sa pri odhadovaní modelu zmesi zamerali na prípad, keď bol počet komponentov g zmesi a priori známy, a teda nebol súčasťou odhadovaného parametra. Toto je však značne obmedzujúca podmienka. V praxi totiž počet komponentov väčšinou známy nie je, hoci zväčša máme aspoň predstavu o tom, v akom rozsahu sa jeho skutočný počet pohybuje. Niekedy napríklad ani nevieme, či analyzované dáta majú heterogénnu štruktúru. Vtedy hovoríme o testovaní homogenity a testujeme hypotézu, že počet komponentov modelu je jedna, proti hypotéze, že počet komponentov je väčší. Odhadnúť zložitosť modelu zmesi, ako niekedy nazývame počet jeho komponentov, je teda úloha veľmi dôležitá. Zároveň je to však náročný problém, ktorý doposiaľ nebol uspokojivo vyriešený. Veľkou výzvou je zhotovenie *konzistentného* odhadu a predovšetkým algoritmu, ktorý by takýto odhad umožňil efektívne aplikovať v praxi.

V tejto kapitole sa pokúsim predstaviť hlavné existujúce metódy pre odhad zložitosti zmesi a zorientovať sa v spleti týchto rozličných techník. Pokúsim sa zachytiť ich historický vývoj, smerovanie a napokon sa budem venovať aj zaujímavým moderným výsledkom.

Pri odhadovaní počtu komponentov treba striktne rozlišovať medzi prípadom, keď je model zmesi použitý len ako vhodný *model pre popísanie dát*, alebo keď reprezentuje *metódu zhlukovej analýzy*. Ako vysvetlenie si uveďme nasledujúci príklad. V prvej kapitole som poukázal na to, že pomocou zmesi normálnych rozdelení možno vhodne modelovať asymetrické distribúcie. Ak normálnu zmes použijeme k detekcii zhlukov v dátach, nemusí byť každý zhluk reprezentovaný práve jedným komponentom zmesi. Niektoré zhluky totiž môžu odpovedať zošikmenému rozdeleniu, na popísanie ktorého model normálnej zmesi spotrebuje vždy viac ako jeden normálny komponent[21]. V kontexte zhlukovej analýzy je teda v tomto prípade skutočný počet komponentov (zhlukov) menší ako počet komponentov odhadnutej normálnej zmesi. Typickým príkladom, kedy sa zmesi používajú k analýze zhlukov, sú napríklad úlohy z genetiky. Zhluky v tomto prípade reprezentujú rôzne genotypy a častou úlohou v tomto prípade je práve zistenie

[21] S efektom zošikmeného rozdelenia zhlukov sa možno niekedy vyrovnať tak, že dáta sa pokúsime vhodne transformovať, napr. *Box-Coxovou transformáciou.*

počtu rôznych genotypov v sledovanej populácii.

V prípade, že *nemáme* žiadnu *apriórnu informáciu* o dátach (parametrickú špecifikáciu komponentov), je otázka počtu komponentov hodnotená cez počet modusov. Procedúry na odhad počtu modusov rozdelenia popisuje napr. Titterington a kol. (1985), ktorý využíva metódy jadrového odhadu hustoty. Bolo tiež odvodených aj niekoľko testov pre počet modusov daného rozdelenia (Hartigan, Hartigan, 1985; Fisher a kol., 1994). Nevýhodou tohoto prístupu je, že funguje len v prípadoch, keď sú komponenty zmesi od seba navzájom dostatočne oddelené.

V ďalšom už budem pracovať s *parametricky formulovanou zmesou* s hustotou

$$f(\mathbf{x}; \boldsymbol{\Psi}) = \sum_{i=1}^{g} \pi_i f(\mathbf{x}; \boldsymbol{\theta}_i), \tag{2.1}$$

kde $f(\cdot;\ \theta)$ je členom nejakej parametrickej rodiny $\{f(\mathbf{x}_j; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ a $\boldsymbol{\Theta}$ je parametrický priestor pre $\boldsymbol{\theta}$. Budem sa zaoberať odhadom zmesovej distribúcie

$$H(\boldsymbol{\theta}) = \sum_{i=1}^{g} \pi_i \mathbf{I}_{[\boldsymbol{\theta}_i \leq \boldsymbol{\theta}]}, \quad \boldsymbol{\theta} \in \boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_g\}, \tag{2.2}$$

kde g bude neznáme.

2.1 Definícia počtu komponentov v zmesi

Počet komponentov v modeli so zmesou hustôt sa niekedy označuje aj termínom *stupeň* (ang. order), poprípade *zložitosť* (ang. complexity) modelu so zmesou hustôt. Predtým, ako sa pustíme ďalej, venujme trochu pozornosti jeho definícii:

TO JE MATOU DEFINICE?

Skutočný počet komponentov g_0 *zmesi* $f(\mathbf{x}; \boldsymbol{\Psi}) = \sum_{i=1}^{g} \pi_i f_i(\mathbf{x}; \theta_i)$ *je najmenšia hodnota* g, *pre ktorú všetky komponenty* $f_i(\mathbf{x}; \theta_i)$ *sú navzájom rôzne a všetky k nim asociované proporcie* π_i *sú nenulové.*

2.2 Základné prístupy riešenia

Metódy odhadu počtu komponentov zmesi (2.1) možno rozdeliť na štytri základné skupiny:

1. **Test hypotézy o počte komponentov metódou pomeru vierohodností** - Pre model so zmesou hustôt nie sú splnené klasické podmienky regularity, v prípade platnosti ktorých by testová štatistika založená na pomere vierohodností (LRTS) mala klasické χ^2 rozdelenie. Výskum v tejto oblasti pozostáva z hľadania vhodnej aproximácie LRTS, poprípade jej teoretického rozdelenia v rôznych špeciálnych prípadoch.

2. **Metódy založené na penalizovanej vierohodnosti** - Vierohodnosť pri modeloch zmesí s rastúcim počtom komponentov narastá (čím väčší počet komponentov, tým lepšie popísanie dát). Po presiahnutí určitej hranice počtu komponentov už ale vierohodnosť narastá len veľmi pomaly. Je teda adektvátne za odhad počtu komponentov brať hodnotu, po zväčšení ktorej sa vierohodnosť už „výrazne" nezvyšuje. Vierohodnosť preto penalizujeme členom, ktorý rastie s počtom parametrov modelu alebo vyjadruje nejakým spôsobom nevhodnosť, resp. chybu tohoto modelu. Za odhad potom berieme maximum tejto penalizovanej vierohodnosti. Tento odhad označíme MPLE.

3. **Metódy založené na minimálnej vzdialenosti** - Vzdialenosť medzi empirickou distribúciou (resp. iného „overfitu" dát) a distribúciou modelu zmesi sa zmenšuje s rastúcim počtom komponentov zmesi. Podobne, ako v predchádzajúcom príklade, teda vhodnou penalizáciou (nepriamo) penalizujeme počet parametrov (a teda aj komponentov) modelu a za odhad berieme minimum penalizovanej vzdialenosti alebo stanovíme hranicu dostatočnej blízkosti a za výsledný počet komponentov berieme najmenšiu hodnotu g, v prípade ktorej došlo k prekročeniu tejto hranice. Tento odhad označíme MPDE[22].

4. **Bayesovský prístup pre odhad počtu komponentov** - V tomto prípade pôjde buď o MCMC metódy na priestore s premenlivou dimenziou, alebo metódy založené na aposteriórnych modusoch vierohodností modelov s počtom komponetov $g = 1, \ldots, g_{max}$, kde g_{max} je nejaká horná hranica pre zložitosť zmesi. Predstavím aj bayesovské testovanie hypotézy o počte komponentov modelu zmesi pomocou Bayesovho faktora.

2.3 Test pomerom vierohodností

Predpokladajme, že dáta pochádzajú z rozdelenia daného hustotou (2.1). Pre voľbu počtu komponentov g sa náramne ponúka test metódou pomeru vierohodností (LRT) pre hypotézu $H_0 : g = g_0$ proti alternatíve $H_1 : g = g_1$, pre nejaké $g_1 > g_0$. Zvyčajne $g_1 = g_0 + 1$. V praxi väčšinou postupujeme tak, že po jednom zvyšujeme počet komponentov v zmesi, až kým sa prírastok maximálnej vierohodnosti po určitej hranici g nezačne prepadať. Túto hranicu potom často volíme za g_0 v H_0.

Nech $\hat{\boldsymbol{\Psi}}_i$ je MLE pre $\boldsymbol{\Psi}$ za podmienky H_i $(i = 0,1)$. Potom testová štatistika v tomto prípade má tvar

$$\text{LRTS} = 2\{\log L(\hat{\boldsymbol{\Psi}}_1) - \log L(\hat{\boldsymbol{\Psi}}_0)\} \tag{2.3}$$

a jej veľké hodnoty svedčia proti hypotéze H_0. Nanešťastie však modely so zmesou hustôt nespĺňajú podmienky regularity(Cramér, 1946), za ktorých má štatistika (2.3) štandardne asymptotické chí-kvadrát rozdelenie so stupňami voľnosti danými rozdielom počtu parametrov modelu v prípade alternatívy a parametrov modelu v prípade nulovej hypotézy.

[22]Takto budem v práci označovať všetky metódy pre odhad počtu komponentov založené na minimálnej vzdialenosti, hoci nie vždy pôjde o penalizovanú vzdialenosť.

2.3.1 Porušenie podmienok regularity

Porušenie podmienok regularity v prípade modelu zmesi ukážem na príklade. Majme zmes dvoch normálnych rozdelení v proporcii π_1 a $\pi_2 = 1 - \pi_1$ so strednou hodnotou $\mu_1 = 0$ a μ_2 a s rovnakým jednotkovým rozptylom. Testujme hypotézu

$$H_0 : f(x_j; \boldsymbol{\Psi}) = \phi(x_j; 0,1) \tag{2.4}$$

proti alternatíve

$$H_1 : f(x_j; \boldsymbol{\Psi}) = \pi_1 \phi(x_j; 0,1) + \pi_2 \phi(x_j; \mu_2, 1). \tag{2.5}$$

Parametrický priestor v tomto prípade teda je

$$\boldsymbol{\Omega} = \{\boldsymbol{\Psi} = (\pi_1, \mu_2)^T : \ [0,1] \times (-\infty, \infty)\}$$

a podpriestor $\boldsymbol{\Omega}_0$ priestoru $\boldsymbol{\Omega}$, ktorý špecifikuje H_0 je

$$\boldsymbol{\Omega}_0 = \{\boldsymbol{\Psi} = (\pi_1, \mu_2)^T : \ ([1] \times (-\infty, \infty)) \cup ([0,1] \times [0])\}.$$

V prípade platnosti H_0 leží teda skutočná hodnota parametra na hranici parametrického priestoru (porušenie podmienok regularity) a zároveň v neidentifikovateľnom podpriestore $\boldsymbol{\Omega}_0$ parametrického priestoru $\boldsymbol{\Omega}$. V prípade neidentifikovateľnosti je však asymptotické rozdelenie pre MLE neznáme.

Ak platí H_0, platí, že $\hat{\boldsymbol{\Psi}}_1$ konverguje v pravdepodobnosti k $\boldsymbol{\Omega}_0$ (Feng, McCulloch, 1996), takže by sme očakávali, že LRTS sa bude pre $n \to \infty$ správať rozumne. Hartigan (1985) ale ukázal, že LRTS je v prípade platnosti H_0 v pravdepodobnosti zhora asymptoticky neobmedzená (hoci veľmi pomalou mierou $\frac{1}{2}\log(\log n)$). Pre konečný rozsah výberu n však rozdelenie LRTS existuje a môžeme ho aproximovať napríklad pomocou bootstrapu.

2.3.2 Pravdepodobnostné rozdelenie pre LRTS

Rozdelenie štatistiky (2.3) najčastejšie aproximujeme technikou bootstrap, avšak v rôznych špeciálnych prípadoch sa ho podarilo odvodiť aj **teoreticky**. Napríklad Ghosh, Sen (1985) odvodili pre prípad zmesi dvoch normálnych komponentov s neznámymi proporciami, s neznámymi ale identifikovateľnými strednými hodnotami, kde μ_2 je z kompaktnej množiny, a s rovnakým a známym rozptylom, že LRTS má rozdelenie ako funkcionál

$$\left[\max\left\{0, \sup_{\mu_2} W(\mu_2)\right\}\right], \tag{2.6}$$

kde $W(\cdot)$ je Gaussovský proces s nulovou strednou hodnotou a variančným jadrom závisiacim na skutočnej hodnote μ_1 za H_0, a rozptyl $W(\mu_2)$ je jednotkový pre všetky μ_2. Podobný výsledok ako (2.6) bol dosiahnutý aj pre prípad všetkých neznámych parametrov (teda aj σ^2 neznáme), ale s podmienkou $0 < a_1 \leq \Delta \leq a_2$ pre nejaké pevné a_1 a a_2, kde $\Delta = |\mu_1 - \mu_2|/\sigma$ (mahalanobisova vzdialenosť). Ukázalo sa, že pre nie príliš veľké hodnoty a_2 sú asymptotické percentily a percentily rozdelenia χ^2_2 veľmi podobné.

Taktiež boli uskutočnené aj nejaké **simulačné štúdie** za účelom aproximácie rozdelenia LRTS. Pre viacrozmerný prípad normálnej zmesi dvoch komponentov (test H_0 : $g_0 = 1$ proti H_1 : $g_1 = 2$) so spoločnou variančnou maticou bola odvodená aproximácia

$$C(\text{LRTS}) \sim \chi^2_{d_1}, \tag{2.7}$$

kde stupeň voľnosti d_1 je dvojnásobok diferencie počtu parametrov za H_0 a H_1 bez parametrov proporcií. Doporučená hodnota pre C je $(n - 1 - p - \frac{1}{2}g_1)/n$, kde p je dimenzia pozorovaní. Neskôr sa ukázalo, že pomer n/p by však mal byť aspoň 5, aby mohla byť aproximácia použiteľná pre stanovenie p-hodnôt. Aj potom je ale sila testu pomerne malá, pokiaľ mahalanobisova vzdialenosť nie je väčšia ako dva. Ďalším dôležitým faktom je, že rozdelenie LRTS určené pomocou simulácií závisí na počiatočnej inicializácii, ukončovacom kritériu a spôsobe vysporadúvania sa s falošnými lokálnymi maximami pri procese iteratívneho výpočtu MLE (napr. EM algoritmus) v prípade jednotlivých hypotéz.

McLachlan (1987) k odhadnutiu *p-hodnoty* pre LRTS používa **bootstrap**. Bootstrapové výbery sú generované zo zmesi s vektorom parametrov $\boldsymbol{\Psi}$, ktorý nahradíme maximálne vierohodným odhadom $\hat{\boldsymbol{\Psi}}_{g_0}$ spočítaným v prípade platnosti nulovej hypotézy z pôvodných dát. Počet týchto bootstrapových výberov označme B. Pre každý bootstrapový výber spočítame LRTS, čím dostaneme aproximáciu jej skutočného rozdelenia. Treba si uvedomiť náročnosť tejto procedúry, pretože pre každý bootstrapový výber musíme spočítať MLE pre tento výber za platnosti oboch hypotéz, keďže

$$\text{LRTS}^{(j)} = 2\left\{L(\hat{\boldsymbol{\Psi}}^{(j)}_{g_1}) - L(\hat{\boldsymbol{\Psi}}^{(j)}_{g_0})\right\} \quad (j = 1, \ldots, B). \tag{2.8}$$

Pre odhad LRTS je ešte dôležité, aby pri výpočtoch MLE z pôvodných dát a bootstrapových výberov bola použitá rovnaká procedúra vzhľadom k tomu, že iný typ počiatočnej inicializácie, ukončovacieho kritéria, atď. má dopad na simulované rozdelenie LRTS. Ak chceme dosiahnuť veľkú presnosť odhadu p-hodnoty pre LRTS touto bootstrapovou technikou, musí byť B veľmi veľké. Niekedy sa stretávame aj s použitím *dvojitého bootstrapu*. Bootstrapové výbery sú vtedy generované zo zmesi s parametrom $\hat{\boldsymbol{\Psi}}^*_{g_0}$, kde $\hat{\boldsymbol{\Psi}}^*_{g_0}$ je MLE počítaný z bootstrapového výberu generovaného zo zmesi s parametrom $\hat{\boldsymbol{\Psi}}_{g_0}$.

Teoretické rozdelenie pre LRTS sa nakoniec podarilo odvodiť pomocou zavedenia špeciálnej reparametrizácie pre model so zmesou hustôt (1.8) (Dacunha-Castelle, Gassiat, 1999), ale jeho podoba[23] je skôr teoretická ako prakticky použiteľná. Tomuto výsledku predchádzal článok od tých istých autorov (Dacunha-Castelle, Gassiat, 1997), v ktorom bola myšlienka reparametrizácie použitá len pre špeciálny prípad testovania $H_0 : g = 1$ proti $H_1 : g > 1$, ktorý sa v prípade modelov zmesí často označuje ako **testovanie homogenity**.

Na záver teda môžno konštatovať, že testovanie hypotézy o počte komponentov zmesi metódou pomeru vierohodností nie je moc vhodné pre praktické použitie. Teoretické

[23] Ide o Gaussovský proces na špeciálnej množine, pričom odvodená asympotická teória predpokladá niekoľko zložito vyzerajúcich podmienok na rodinu rozdelení zmesi (pre normálne a poissonove zmesi už tieto boli overené).

rozdelenie LRTS je totiž príliš komplikované a simulačné štúdie alebo bootstrap zase príliš náročné na výpočtový čas.

2.4 Testovanie homogenity

Tento špeciálny prípad testovania zložitosti modelu zmesi sa niekedy oddeľuje od klasického testovania počtu komponentov. Pri testovaní homogenity totiž rozhodujeme, či vôbec ide o zmes, resp. či dáta majú heterogénnu štruktúru. Naproti tomu o testovaní počtu komponentov zmesi hovoríme väčšinou v prípade, keď je prítomnosť heterogenity už známa, teda $g > 1$.

Ak teda platí nulová hypotéza, náhodný výber v tomto prípade pochádza z parametrického rozdelenia s hustotou $f(\mathbf{x},\theta)$. Klasickým príkladom testu homogenity je $C(\alpha)$ test Neymana a Scotta z roku 1966, ktorý zamieta hypotézu ak

$$\sum_{i=1}^{n} t(\mathbf{x}_i,\boldsymbol{\theta}) > c, \tag{2.9}$$

kde $t(\mathbf{x},\boldsymbol{\theta})$ je vhodne zvolená štatistika, ktorá má v prípade nulovej hypotézy vzhľadom k hustote $f(\mathbf{x},\theta)$ nulovú strednú hodnotu (napr. $t(\mathbf{x},\boldsymbol{\theta}) = \frac{\partial^2}{\partial\theta^2} f(\mathbf{x},\theta)/f(\mathbf{x},\theta)$).

2.4.1 Vážený test homogenity

Susko (2003) odvodil *vážený test homogenity* j-teho komponentu zmesi, ktorý testuje hypotézu $H_0 : g = g_0$ proti alternatíve $H_1 : g > g_0$ tak, že testuje či je aktuálny j-ty komponent zmesi homogénny alebo nie. Ak testom zamietneme homogenitu tohoto komponentu, tak zamietame aj hypotézu H_0. Testom zamietame H_0 ak

$$\sum_{i=1}^{n} p(j|\mathbf{x}_i;\hat{\boldsymbol{\Psi}}) t(\mathbf{x}_i,\hat{\boldsymbol{\theta}}_j) > c,$$

kde $p(j|\mathbf{x}_i;\hat{\boldsymbol{\Psi}}) = \boldsymbol{\pi}_j f(\mathbf{x}_j;\hat{\boldsymbol{\theta}}_j)/\sum_{k=1}^{g_0} \hat{\pi}_k f(\mathbf{x}_j;\hat{\boldsymbol{\theta}}_k)$ sú váhy j-teho komponentu (pravdepodobnosť, že pozorovanie i padne do j-teho komponentu). Pracujme so štatistikou

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} p(j|\mathbf{x}_i;\hat{\boldsymbol{\Psi}}) t(\mathbf{x}_i,\hat{\boldsymbol{\theta}}_j), \tag{2.10}$$

kde $\hat{\boldsymbol{\Psi}}$ je MLE parametra $\boldsymbol{\theta}$ za H_0. Stredná hodnota tejto štatistiky v prípade nulovej hypotézy je vzhľadom k hustote $f(\mathbf{x};\boldsymbol{\Psi}) = \sum_{k=1}^{g_0} \pi_k f(\mathbf{x};\theta_k)$ nulová, pretože štatistika $t(\mathbf{x},\boldsymbol{\theta})$ je volená tak, aby mala nulovú strednú hodnotu a dá sa odvodiť, že jej rozptyl má tvar

$$E\left[\{p(j|\mathbf{X};\boldsymbol{\Psi}) t(\mathbf{X},\boldsymbol{\theta}_j)\}^2\right] - r_j(\boldsymbol{\Psi})^T J^{-1}(\boldsymbol{\Psi}) r_j(\boldsymbol{\Psi}), \tag{2.11}$$

kde $r_j(\boldsymbol{\Psi}) = E\left[u(\mathbf{X},\boldsymbol{\Psi}) t(\mathbf{X},\boldsymbol{\Psi}) p(j|\mathbf{X};\boldsymbol{\Psi})\right]$, $u(\mathbf{x},\boldsymbol{\Psi})$ je gradient $\log\left[f(\mathbf{x},\boldsymbol{\Psi})\right]$ a $J(\boldsymbol{\Psi})$ je Fisherova informačná matica. Vzhľadom k tomu, že štatistika (2.10) má (asymptoticky) normálne rozdelenie, pre test homogenity j-teho komponentu zmesi ju stačí už len

znormovať (podeliť odmocninou odvodeného rozptylu) a porovnať s kvantilmi štandardizovaného normálneho rozdelenia. Odhad rozptylu (2.11) však obsahuje odhad Fisherovej Informačnej matice, čo komplikuje praktickú implementáciu tohoto testu.

Ak nás zaujíma len test hypotézy $H_0 : g = g_0$ proti $H_1 : g > g_0$ a nie jednotlivé komponenty, môžeme použiť *agregovaný vážený test homogenity.* Štatistika potom vyzerá takto

$$\frac{1}{\sqrt{n}} \sum_{j=1}^{g_0} \sum_{i=1}^{n} p(j|\mathbf{x}_i; \hat{\boldsymbol{\Psi}}) t(\mathbf{x}_i, \boldsymbol{\theta}_j). \tag{2.12}$$

Použitie agregácie významne nezníži silu testu, pričom sila testu je podľa autora porovnateľná s použitím oveľa viac výpočtovo náročnejšej bootstrapovej aproximácie štatistiky LRTS.

2.4.2 Grafický prístup

V prvej kapitole sme mohli vidieť, že hustota zmesi dvoch normálnych komponentov nie je vždy bimodálna. Roeder (1994) si ale všimla, že pomer hustoty zmesi dvoch normálnych komponentov a hustoty vhodného normálneho rozdelenia je bimodálny vždy a na tomto fakte založila svoj prístup testovania homogenity.

Predpokladajme, že chceme testovať hypotézu

$$H_0 : f(x) = f(x; H_1, \sigma^2) = \varphi(x; \mu, \sigma^2)$$

proti alternatíve

$$H_1 : f(x) = f(x; H_2, \sigma^2) = \pi\varphi(x; \mu_1, \sigma^2) + (1 - \pi)\varphi(x; \mu_2, \sigma^2),$$

kde φ je hustota normálneho rozdelenia. Nech $\mu = \pi\theta_1 + (1 - \pi)\theta_2$ a $\tau^2 = \pi(\theta_1 - \mu)^2 + (1 - \pi)(\theta_2 - \mu)^2$. Roeder vo svojom článku ukázala, že podiel funkcií definovaný ako

$$r(x) = \frac{f(x; H_2, \sigma^2)}{\varphi(x; \mu, \sigma^2 + \tau^2)} \tag{2.13}$$

je funkcia bimodálna práve vtedy, ak $\theta_2 \neq \theta_1$ a $0 < \pi < 1$. Funkcia $r(x) - 1$ teda v prípade hypotézy H_1 štyrikrát mení znamienko, a to v tomto poradí $(-, +, -, +, -)$. Vzhľadom k tomu, že odhad $r(x)$ vykazuje nestabilitu v chvostoch rozdelenia, bola odvodená modifikácia tejto diagnostiky

$$\begin{aligned} t(x) &= \varphi^{1/2}(x; \mu, \sigma^2 + \tau^2)[r(x) - 1] \\ &= \frac{f(x; H_2, \sigma^2) - \varphi(x; \mu, \sigma^2 + \tau^2)}{\varphi^{1/2}(x; \mu, \sigma^2 + \tau^2)}, \end{aligned} \tag{2.14}$$

ktorá taktiež v prípade hypotézy H_1 štyrikrát mení znamienko.

K **aplikácii** tohoto výsledku použijeme neparametrický odhad $t(x)$,

$$[f_n(x; h_n) - \varphi(x; \overline{x}, s^2 + h_n^2)] / \varphi^{1/2}(x, \overline{x}, s^2 + h_n^2),$$

kde $f_n(x;h_n) = \frac{1}{n}\sum_{j=1}^{n}\varphi(x;x_j,h_n^2)$ je jadrový odhad $f(x;H_2,\sigma^2)$, $\overline{x}$ je výberový priemer a s^2 výberový rozptyl. Tento odhad môžeme ešte znormovať do záverečnej podoby tak, aby mal jednotkový rozptyl

$$T_n(x;h_n) = (2\pi^{1/2}nh_n)^{1/2}\left[\frac{f_n(x;h_n) - \varphi(x;\overline{x},s^2+h_n^2)}{\varphi^{1/2}(x;\overline{x},s^2+h_n^2)}\right]. \tag{2.15}$$

V prípade alternatívnej hypotézy potom odhad $T_n(x;h_n)$ aproximuje bimodálnu funkciu premennej x so štyrmi zmenami znamienka, a zároveň sa dá ukázať, že v prípade nulovej hypotézy je približne rozdelená ako stacionárny Gaussovský proces. Graf $T_n(x;h_n)$ (*mixture detection plot*, MDP) teda diagnostikuje homogenitu skúmaného náhodného výberu. Vykresľujeme ho zväčša pre viac vyhladzovacích parametrov h_n. Odporúča sa začať s hodnotou $h_n = sn^{-1/5}$ a tú potom vhodne upravovať tak, aby sme z grafu odstránili prudké oscilácie.

U MDP je dôležité rozlišovať medzi náhodným šumom a modusmi, ktoré majú svoj pôvod v komponentoch zmesi. V prípade H_1 MDP typicky vykazuje relatívne veľké modusy a jasne viditeľné dno. Za týmto účelom boli odvodené aj kritické hranice, v ktorých by mal v prípade nulovej hypotézy MDP oscilovať okolo nuly. Na testovanie však nie sú tieto hranice úplne vhodné, lebo závisia na náhodne zvolenej šírke vyhladzovacieho parametra h_n a sú odvodené pomocou asymptotickej teórie, ktorá si vyžaduje veľký rozsah náhodného výberu. Treba ešte dodať, že metóda je citlivá na predpoklad normality.

Táto technika sa dá rozšíriť aj na prípad testovania väčšieho počtu komponentov, pričom nemusí už ísť o homoskedastické komponenty. Diagnostika má však v tomto prípade menšiu silu, preto sa táto metóda hodí predovšetkým na testovanie homogenity.

2.5 MPLE

Odhad zložitosti modelu zmesi, alebo jenoducho výber modelu, pomocou metódy MPLE realizujeme tak, že maximalizujeme penalizovanú vierohodnosť modelu cez nejakú množinu uvažovaných hodnôt zložitosti modelu G (väčšinou $G = 1,\ldots,g_{max}$), teda

$$\hat{g} = \sup_{g\in G}\left\{\log L(\hat{\boldsymbol{\Psi}}_g) - C(g)\right\}, \tag{2.16}$$

kde $C(g)$ je penalizačný člen pre model s počtom komponentov g a $\hat{\boldsymbol{\Psi}}_g$ je MLE odhad parametra $\hat{\boldsymbol{\Psi}}$ pre tento model. Opodstatnenie tohoto prístupu som naznačil už v úvode kapitoly a v nasledujúcej časti ho vysvetlím ešte z iného pohľadu.

2.5.1 Minimalizácia Kullback-Leiblerovej divergencie

Nech $f(\mathbf{u})$ je nejaká hustota zmesi a $f(\mathbf{u},\hat{\boldsymbol{\Psi}})$ nech označuje jej odhad. Pokúsme sa zvoliť odhad modelu tak, aby sme minimalizovali Kullback-Leiblerovu (KL) divergenciu $f(\mathbf{u})$ vzhľadom k $f(\mathbf{u},\hat{\boldsymbol{\Psi}})$, ktorá je definovaná ako

$$KL(f(\mathbf{u}),f(\mathbf{u},\hat{\boldsymbol{\Psi}})) = \int f(\mathbf{u})\log f(\mathbf{u})d\mathbf{u} - \int f(\mathbf{u})\log f(\mathbf{u},\hat{\boldsymbol{\Psi}})d\mathbf{u}. \tag{2.17}$$

Keďže prvý člen na pravej strane (2.17) nezávisí na modeli, zaujíma nás len člen druhý, ktorý môžeme vyjadriť ako

$$\eta(\mathbf{x}; F) = \int f(\mathbf{u}) \log f(\mathbf{u}; \hat{\mathbf{\Psi}}) d\mathbf{u} = \int \log f(\mathbf{u}; \hat{\mathbf{\Psi}}) dF(\mathbf{u}), \tag{2.18}$$

kde F je skutočná distribučná funckia a $\mathbf{x} = (\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T)^T$ je vektor pozorovaní. Jednoduchý odhad tohoto členu potom získame tak, že za F dosadíme empirickú distribučnú funkciu (EDF) $\hat{F}_n$. Dostaneme tak násobok logaritmickej vierohodnosti pre parameter $\hat{\mathbf{\Psi}}$

$$\eta(\mathbf{x}; \hat{F}_n) = \frac{1}{n} \sum_{j=1}^{n} \log f(\mathbf{x}_j; \hat{\mathbf{\Psi}}) = \frac{1}{n} \log L(\hat{\mathbf{\Psi}}). \tag{2.19}$$

Minimalizácia (2.17) odpovedá maximalizácii $\eta(\mathbf{x}; \hat{F}_n)$, a teda maximalizácii logaritmickej vierohodnosti. Vzhľadom k tomu, že EDF $\hat{F}_n$ je vo všeobecnosti bližšie k odhadu distribučnej funkcie $F_{\hat{\mathbf{\Psi}}}$ ako ku skutočnej distribučnej funckii F, odhad $\eta(\mathbf{x}; \hat{F}_n)$ dáva nadhodnotenie pre $\int \log f(\mathbf{x}) dF(\mathbf{x})$. Vychýlením tohoto odhadu je funkcionál

$$\begin{aligned} b(F) &= E_F \left\{ \eta(\mathbf{X}; \hat{F}_n) - \eta(\mathbf{X}; F) \right\} \\ &= E_F \left\{ \frac{1}{n} \sum_{j=1}^{n} \log f(\mathbf{X}_j; \hat{\mathbf{\Psi}}) - \int \log f(\mathbf{u}; \hat{\mathbf{\Psi}}) dF(\mathbf{x}) \right\}, \end{aligned} \tag{2.20}$$

kde E_F je stredná hodnota vzhľadom k spoločnej distribučnej funkcii $\mathbf{X}_1, \ldots, \mathbf{X}_n$. Pri odhade modelu založenom na minimalizácii KL divergencie - maximalizácii vierohodnosti - je teda potrebné zahrnúť aj toto vychýlenie. Odhad volíme tak, že maximalizujeme

$$\log L(\hat{\mathbf{\Psi}}) - b(F), \tag{2.21}$$

kde $b(F)$ nahradíme vhodným odhadom. Súvis s (2.16) je už zrejmý.

2.5.2 Klasické informačné kritériá

Použitím konkrétnej voľby penalizácie v (2.16) dostaneme kritérium pre výber modelu. Takéto kritérium sa nazýva *informačné kritérium pre výber modelu so zmesou hustôt* alebo len *informačné kritérium*. Väčšinou sa ale tieto kritériá uvádzajú v tejto podobe

$$-2 \log L(\mathbf{\Psi}) + 2C, \tag{2.22}$$

a namiesto maximalizácie (2.21) sa používa minimalizácia (2.22).

AIC

Toto kritérium je veľmi známe a nepoužíva sa len v prípade modelu zmesi. Akaike ukázal, že $b(F)$ definované v (2.20) je asymptoticky rovné d, kde d je celkový počet parametrov modelu. *Akaikeho informačné kritérium* (AIC) teda vyberá model na základe minimalizácie výrazu

$$-2 \log L(\hat{\mathbf{\Psi}}) + 2d. \tag{2.23}$$

Toto kritérium patrí k tým používanejším, hoci častokrát vedie k nadhodnoteniu skutočného počtu komponentov. Okrem klasického AIC kritéria sa niekedy vyskytuje aj jeho varianta AIC3, ktorá má tvar $-2\log L(\hat{\Psi}) + 3d$.

EIC

Ishiguro a kol. (1997) odhadujú vychýlenie $b(F)$ v (2.21) bootstrapom. Autorom tejto resamplovej techniky je Efron, preto odvodené kritérium pomenovali po ňom ako *Efronovo informačné kritérium* (EIC). Kritérium má tvar

$$-2\log L(\hat{\Psi}) + 2b(\hat{F}_n) \tag{2.24}$$

kde

$$b(\hat{F}_n) = E_{\hat{F}_n}\left\{\frac{1}{n}\sum_{j=1}^{n}\log f(\mathbf{X}_j^*;\hat{\Psi}^*) - \frac{1}{n}\sum_{j=1}^{n}\log f(\mathbf{X}_j;\hat{\Psi}^*)\right\} \tag{2.25}$$

je vychýlenie založené na (neparametrickom) bootstrapovom výbere. V (2.25) teda $\mathbf{X}_1^*,\ldots,\mathbf{X}_n^* \overset{i.i.d.}{\sim} \hat{F}_n$ a $\hat{\Psi}^*$ je MLE založený na tomto bootstrapovom výbere. Ak teda $\mathbf{X}_1^{(b)},\ldots,\mathbf{X}_n^{(b)} \overset{i.i.d.}{\sim} \hat{F}_n$ $(b = 1,\ldots,B)$ je B nezávislých bootstrapových výberov a $\hat{\Psi}^{(b)}$ $(b = 1,\ldots,B)$ sú k nim príslušné MLE, potom

$$b(\hat{F}_n) \approx \frac{1}{B}\sum_{b=1}^{B}\left\{\frac{1}{n}\sum_{j=1}^{n}\log f(\mathbf{x}_j^{(b)};\hat{\Psi}^{(b)}) - \frac{1}{n}\sum_{j=1}^{n}\log f(\mathbf{x}_j;\hat{\Psi}^{(b)})\right\}. \tag{2.26}$$

Počet bootstrapových výberov možno ešte zredukovať *technikou redukcie rozptylu* navrhnutou v Konishi, Kitagawa (1996). I napriek tomu je však táto metóda výpočtovo veľmi náročná.

ICOMP

Kritérium informačnej zložitosti (informational complexity) vzniklo v roku 1990 za účelom vylepšenia kritéria AIC. Jeho forma je

$$-2\log L(\hat{\Psi}) + C_1 - C_2, \tag{2.27}$$

kde

$$\begin{aligned} C_1 &= d\log\left\{d^{-1}\mathrm{tr}J^{-1}(\hat{\Psi})\right\}, \\ C_2 &= \log\left\{\det[J^{-1}(\hat{\Psi})]\right\}, \end{aligned} \tag{2.28}$$

d je počet parametrov modelu, „tr" značí stopu a „det" determinant matice. Znepokojujúce u tohoto kritéria je, že nie je invariantné k reparamterizácii modelu. Keďže odhad matice $J(\Psi)$ nevhodný pre použitie v praxi, používa sa aproximácia tejto matice, ktorá však vedie k pomerne komplikovanému výsledku, preto ho neuvádzam. Celeux, Soromenho uvádzajú, že toto kritérium nadhodnocuje viac ako kritérium AIC.

CVIC

Korekcii logaritmickej vierohodnosti o vychýlenie (2.20) sa podľa Smyth (2000) dá vyhnúť použitím *cross-validácie*. Počet komponentov zmesi tak volíme na základe cross-validovanej logaritmickej vierohodnosti (*informačné kritérium založené na cross-validácii*, CVIC), ktorej základná forma je

$$\sum_{j=1}^{n} \log f(\mathbf{x}_j; \hat{\boldsymbol{\Psi}}_{(j)}), \tag{2.29}$$

kde $\hat{\boldsymbol{\Psi}}_{(j)}$ je MLE založený na $\mathbf{x}_1, \ldots, \mathbf{x}_{j-1}, \mathbf{x}_{j+1}, \ldots, \mathbf{x}_n$, teda na všetkých pozorovaniach s výnimkou j-teho. Táto podoba kritéria CVIC je však veľmi časovo náročná, preto sa používa k-násobná cross-validácia ($k > 1$) alebo *Monte Carlo cross-validácia*[24].

MIR

Mieru konvergencie EM algoritmu určuje najmenšie vlastné číslo tzv. *informačnej matice pomeru* (McLachlan, Peel, 2000)

$$J_c^{-1}(\hat{\boldsymbol{\Psi}}; \mathbf{x}) \mathbf{I}(\hat{\boldsymbol{\Psi}}; \mathbf{x}), \tag{2.30}$$

kde $\mathbf{I}(\hat{\boldsymbol{\Psi}}; \mathbf{x})$ je výberová informačná matica počítaná podľa (1.28) a $J_c(\hat{\boldsymbol{\Psi}}; \mathbf{x})$ je informačná matica (viď (1.27)) pre prípad kompletného datového vektora. Kritérium MIR (minimum information ratio) je založené na tomto vlastnom čísle. Heuristicky totiž veľká hodnota tohoto najmenšieho vlastného čísla matice (2.30) odpovedá dobrému popísaniu dát, pričom malá naopak zlému. Nech $\hat{\boldsymbol{\Psi}}_g$ je MLE neznámeho parametra $\boldsymbol{\Psi}$ pre model zmesi s počtom komponentov g a e_g je najmenšie vlastné číslo matice (2.30) odpovedajúcej odhadu $\hat{\boldsymbol{\Psi}}_g$. V pôvodnej verzii kritéria MIR z roku 1992 je počet komponentov volený tak, aby maximalizoval e_g cez g (Windham, Cutler, 1992). Hodnota e_g sa dá jednoducho spočítať vďaka súvisu s konvergenciou EM algoritmu (McLachlan, Peel, 2000, str. 206).

V roku 1998 bola odvodená modifikácia tohoto kritéria, ktorej konkrétnu podobu možno nájsť v Polynemis, Titterington (1998).

2.5.3 Informačné kritériá odvodené v rámci bayesovského prístupu

Niekoľko kritérií odvodených v rámci bayesovského prístupu pre výber modelu sa používa aj v „nebayesovskom„ prístupe. Hlavné z nich používajú aproximáciu integrovanej vierohodnosti, ktorá bola použitá pri odvodení známeho Bayesovského informačného kritéria.

[24] Dáta sa B-krát nezávisle rozdelia na testovaciu množinu o rozsahu γn a trénovaciu množinu o rozsahu $(1-\gamma)n$, z ktorej sa počíta odhad $\boldsymbol{\Psi}$ (γ je volená konštanta).

Laplaceova aproximačná metóda

Nech $p(\boldsymbol{\Psi})$ je hustota apriórneho rozdelenia parametra $\boldsymbol{\Psi}$. Integrovaná vierohodnsoť $p(\mathbf{x})$ je definovaná ako

$$p(\mathbf{x}) = \int p(\boldsymbol{\Psi}, \mathbf{x}) d\boldsymbol{\Psi} = \int \exp\{\log p(\boldsymbol{\Psi}, \mathbf{x}) d\boldsymbol{\Psi}\} \tag{2.31}$$

kde

$$p(\boldsymbol{\Psi}, \mathbf{x}) = p(\boldsymbol{\Psi}) L(\boldsymbol{\Psi}).$$

V skutočnosti sú výsledky podmienené konkrétnym modelom (hustotami komponentov a počtom komponentov), ale zatiaľ tento fakt z notácie vynechám. Nech $\tilde{\boldsymbol{\Psi}}$ je modus aposteriórneho rozdelenia, pre ktorý

$$\partial \log p(\tilde{\boldsymbol{\Psi}}, \mathbf{x}) / \partial \boldsymbol{\Psi} = \mathbf{0}. \tag{2.32}$$

Integrál (2.31) aproximujeme pomocou Taylorovho rozvoja do druhého rádu okolo bodu $\boldsymbol{\Psi} = \tilde{\boldsymbol{\Psi}}$

$$\log p(\boldsymbol{\Psi}, \mathbf{x}) \approx \log p(\tilde{\boldsymbol{\Psi}}, \mathbf{x}) - \frac{1}{2}(\boldsymbol{\Psi} - \tilde{\boldsymbol{\Psi}})^T \mathbf{H}(\tilde{\boldsymbol{\Psi}})(\boldsymbol{\Psi} - \tilde{\boldsymbol{\Psi}}), \tag{2.33}$$

kde $\mathbf{H}(\tilde{\boldsymbol{\Psi}})$ je záporný hesián pre $\log p(\boldsymbol{\Psi}, \mathbf{x})$. Člen prvého rádu vzhľadom k (2.32) z aproximácie vypadol. Ak integrand v (2.31) nahradíme touto aproximáciou, tak po úprave dostaneme

$$\begin{aligned} p(\mathbf{x}) &= \exp\{\log p(\tilde{\boldsymbol{\Psi}},\mathbf{x})\} \int \exp\{-\frac{1}{2}(\boldsymbol{\Psi} - \tilde{\boldsymbol{\Psi}})^T \mathbf{H}(\tilde{\boldsymbol{\Psi}})(\boldsymbol{\Psi} - \tilde{\boldsymbol{\Psi}})\} d\boldsymbol{\Psi} \\ &= p(\tilde{\boldsymbol{\Psi}},\mathbf{x})(2\pi)^{\frac{1}{2}d} |\mathbf{H}(\tilde{\boldsymbol{\Psi}})|^{-1/2}. \end{aligned} \tag{2.34}$$

Logaritmickú integrovanú vierohodnosť potom môžeme aproximovať ako

$$\log p(\mathbf{x}) \approx \log L(\tilde{\boldsymbol{\Psi}}) + \log p(\tilde{\boldsymbol{\Psi}}) - \frac{1}{2} \log |\mathbf{H}(\tilde{\boldsymbol{\Psi}})| + \frac{1}{2} d \log(2\pi). \tag{2.35}$$

Táto aproximácia sa nazýva *Lapleceova metóda* a jej dôležitou variantou je

$$\log p(\mathbf{x}) \approx \log L(\hat{\boldsymbol{\Psi}}) + \log p(\hat{\boldsymbol{\Psi}}) - \frac{1}{2} \log |\mathbf{I}(\tilde{\boldsymbol{\Psi}}; \mathbf{x})| + \frac{1}{2} d \log(2\pi), \tag{2.36}$$

kde sme modus aposteriórneho rozdelenia nahradili s MLE $\hat{\boldsymbol{\Psi}}$ a člen $\mathbf{H}(\tilde{\boldsymbol{\Psi}})$ výberovou informačnou maticou. Táto aproximácia predpokladá, že apriórne rozdelenie je veľmi rozptýlené, a preto jeho efekt môžeme ignorovať.

BIC

Bayesovské informačné kritérium (Schwarz, 1978) patrí medzi jedno z najlepších a najčastejšie používaných informačných kritérií pre výber modelu zmesi. Je založené na maximalizácii logaritmickej integrovanej vierohodnosti (2.36), pričom ignoruje člen rádu $O(1)$. Vzhľadom k

$$|\mathbf{I}(\hat{\boldsymbol{\Psi}}; \mathbf{x})| = O(n^d) \tag{2.37}$$

má toto kritérium podobu

$$-2\log L(\hat{\Psi}) + d\log n. \tag{2.38}$$

Pre $n > 8$ ($\log n > 2$) toto kritérium penalizuje zložitosť modelu viac ako AIC, ktorého penalizačný člen nezávisí na rozsahu výberu.

LEC

Empirické Laplaceovo kritérium (LEC) sa na rozdiel od BIC s aproximáciou (2.36) vysporadúva priamo. Výbérovú informačnú maticu $\mathbf{I}(\hat{\Psi}; \mathbf{x})$ v (2.36) nahradíme empirickou informačnou maticou

$$\mathbf{I}_e(\hat{\Psi}; \mathbf{x}) = \sum_{j=1}^{n} \mathbf{s}(\mathbf{x_j}; \hat{\Psi})\mathbf{s}^{\mathrm{T}}(\mathbf{x_j}; \hat{\Psi}), \text{ kde } \mathbf{s}(\mathbf{x_j}; \Psi) = \partial \log L_j(\Psi)/\partial\Psi$$

a kde $L_j(\Psi) = f(\mathbf{x}_j; \Psi)$ je vierohodnostná funkcia v bode Ψ pre jedno pozorovanie $\mathbf{x}_j$ ($j = 1, \ldots, g$). Kritérium LEC má teda nasledovný tvar

$$-2\log L(\hat{\Psi}) - 2\log p(\hat{\Psi}) + \log|\mathbf{I}_e(\hat{\Psi})| - d\log(2\pi). \tag{2.39}$$

Vzhľadom k

$$\mathbf{s}(\mathbf{x_j}; \Psi) = E_{\Psi}\{\partial \log L_{cj}(\Psi)/\partial\Psi |\ \mathbf{x}\}, \tag{2.40}$$

kde $L_{cj}(\Psi)$ je vierohodnosť v prípade kompletného datového vektora pre jedno pozorovanie $\mathbf{x}_j$ ($j = 1, \ldots, n$), sa nedávno (1997) začal pre výpočet empirickej informačnej matice používať vzťah

$$\mathbf{s}(\mathbf{x_j}; \Psi) = \sum_{i=1}^{g} \tau_i(\mathbf{x}_j; \Psi)\partial\{\log \pi_i + \log f_i(\mathbf{x}_j; \boldsymbol{\theta}_i)\}/\partial\Psi \quad (j = 1, \ldots, n). \tag{2.41}$$

Nevýhodou tohoto kritéria je jeho komplikovanosť.

2.5.4 Informačné kritériá založené na klasifikácii

Ďalšou zaujímavou skupinou kritérií pre odhad počtu komponentov v modeli so zmesou hustôt sú kritériá založené na tzv. *klasifikačnej vierohodnosti* $L_c(\Psi)$, čo je vlastne iný názov pre vierohodnosť založenú na kompletnom dátovom vektore (1.31).

CLC

V prípade konečnej zmesi môžeme logaritmickú vierohodnosť vyjadriť pomocou klasifikačnej vierohodnosti ako

$$\log L(\Psi) = \log L_c(\Psi) - \log k(\Psi), \tag{2.42}$$

kde

$$\log k(\Psi) = \sum_{i=1}^{g}\sum_{j=1}^{n} z_{ij}\log \tau_{ij} \tag{2.43}$$

a kde $\tau_{ij} = \tau_i(\mathbf{x}_j; \boldsymbol{\Psi})$ je aposteriórna pravdepodobnosť toho, že j-te pozorovanie patrí do i-teho komponentu. Takže $k(\boldsymbol{\Psi})$ je podmienené rozdelenie vektora $\mathbf{z} = (\mathbf{z}_1^T, \ldots, \mathbf{z}_n^T)^T$ pri dátach $\mathbf{x} = (\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T)^T$. Vzhľadom k

$$E(Z_{ij}|\ \mathbf{x}) = P(Z_{ij} = 1|\ \mathbf{x}) = \tau_{ij},$$

sa podmienená stredná hodnota $\log k(\boldsymbol{\Psi})$ pri pozorovaných dátach $\mathbf{x}$ rovná $-EN(\boldsymbol{\tau})$, kde

$$EN(\boldsymbol{\tau}) = -\sum_{i=1}^{g}\sum_{j=1}^{n} \tau_{ij} \log \tau_{ij} \tag{2.44}$$

je entropia fuzzy klasifikačnej matice $\mathbf{C} = (\tau_{ij})_{ij}$ a kde $\boldsymbol{\tau} = (\boldsymbol{\tau}_1^T, \ldots, \boldsymbol{\tau}_n^T)^T$ a $\boldsymbol{\tau}_j = (\tau_1(\mathbf{x}_j; \boldsymbol{\Psi}), \ldots, \tau_g(\mathbf{x}_j; \boldsymbol{\Psi}))^T$ $(j = 1, \ldots, n)$. Ak v (2.42) použijeme MLE parametra $\boldsymbol{\Psi}$, $\hat{\boldsymbol{\Psi}}$, a v $L_c(\boldsymbol{\Psi}, \boldsymbol{z})$ nahradíme $\mathbf{z} = \hat{\boldsymbol{\tau}}$ dostávame

$$\log L_c(\hat{\boldsymbol{\Psi}}; \hat{\boldsymbol{\tau}}) = \log L(\hat{\boldsymbol{\Psi}}) - EN(\hat{\boldsymbol{\tau}}), \tag{2.45}$$

kde $\hat{\boldsymbol{\tau}}$ je MLE pre $\boldsymbol{\tau}$, ktoré dostaneme, ak namiesto τ_{ij} použijeme

$$\hat{\tau}_{ij} = \tau_i(\mathbf{x}_j; \hat{\boldsymbol{\Psi}}) \quad (i = 1, \ldots, g; j = 1, \ldots, n). \tag{2.46}$$

Kritérium založené na klasifikačnej vierohodnosti (CLC) z roku 1997 má teda tento tvar

$$-2 \log L(\hat{\boldsymbol{\Psi}}) + 2EN(\hat{\boldsymbol{\tau}}). \tag{2.47}$$

Ide o kritérium, ktoré penalizuje zložitosť modelu zmesi odhadnutou entropiou $EN(\hat{\boldsymbol{\tau}})$.

Treba si uvedomiť, že $EN(\hat{\boldsymbol{\tau}})$ je blízko nuly, ak sú komponenty zmesi výrazne oddelené, a naopak, ak sú komponenty „blízko pri sebe", $EN(\hat{\boldsymbol{\tau}})$ dosahuje veľkých hodnôt. Ako silno toto kritérium penalizuje logaritmickú vierohodnosť závisí teda od toho, ako veľmi sú komponenty zmesi navzájom oddelené. Ukázalo sa tiež, že kritérium CLC funguje dobre, ak sú proporcie zmesi približne rovnaké, ale v prípade, že neexistuje reštrikcia na proporcie zmesi, CLC zložitosť modelu nadhodnocuje.

NEC

Kritérium založené na normalizovanej entropii (NEC) z roku 1996 používa pre voľbu počtu komponentov priamo odhadnutú entropiu $EN(\hat{\boldsymbol{\tau}})$, ktorá je však normalizovaná v dôsledku toho, že $\log L(\hat{\boldsymbol{\Psi}})$ s počtom komponentov g rastie. Kritérium NEC má tvar

$$NEC(g) = \frac{EN(\hat{\boldsymbol{\tau}})}{\log L(\hat{\boldsymbol{\Psi}}) - \log L(\hat{\boldsymbol{\Psi}}^*)}, \tag{2.48}$$

kde $\hat{\boldsymbol{\Psi}}^*$ je MLE parametra $\boldsymbol{\Psi}$ v prípade jedného komponentu $(g = 1)$. Počet komponentov g volíme tak, aby sme minimalizovali $NEC(g)$. Keďže entropia pre $g = 1$ je nula, kritérium nedokáže dobre rozhodnúť medzi $g = 1$ a hodnotou g väčšou ako jedna.

V roku 1999 bola odvodená modifikácia tohoto kritéria, ktorá pre $g = 1$ definuje NEC kritérium jednotkou, teda $NEC(1) = 1$. Odôvodnenie je nasledovné. Keď porovnáme

dve hodnoty g, g_0 a g_1 ($g_0 < g_1$), je zrejmé, že ak $L_c(\hat{\boldsymbol{\Psi}}_1; \hat{\boldsymbol{\tau}}_1) < L_c(\hat{\boldsymbol{\Psi}}_0; \hat{\boldsymbol{\tau}}_0)$, kde $\hat{\boldsymbol{\Psi}}_i$ a $\hat{\boldsymbol{\tau}}_i$ sú MLE pre $g = g_i$ ($i = 0,1$), tak preferujeme g_0. Pre uprednostnenie $g > 1$ pred $g = 1$ teda prirodzene požadujeme aby $L_c(\hat{\boldsymbol{\Psi}}; \hat{\boldsymbol{\tau}})$ bola pre $g > 1$ väčšia ako pre $g = 1$. Ak toto platí, tak vzhľadom k (2.42) potom platí, že $0 \leq NEC(g) \leq 1$. Ak teda neexistuje žiadne g, pre ktoré $NEC(g) < 1$, potom nie je žiaden dôvod na voľbu $g > 1$.

ICL

Kritérium založené na integrovanej klasifikačnej vierohodnosti (ICL) bolo v roku 1998 odvodené za účelom odstránenia nedostatkov kritérií BIC a CLC.

Nech apriórne rozdelenie parametra $\boldsymbol{\Psi} = (\boldsymbol{\pi}^T, \boldsymbol{\xi}^T)^T$ má hustotou $p(\boldsymbol{\Psi})$, ktorá sa dá faktorizovať ako $p(\boldsymbol{\Psi}) = p(\boldsymbol{\pi})p(\boldsymbol{\xi})$, kde $\boldsymbol{\xi} = (\boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_g^T)^T$ a nech parameter $\boldsymbol{\pi}$ má Dirichletovo rozdelenie $D(\alpha_1, \ldots, \alpha_g)$ s parametrami $\alpha_i = \alpha$ ($i = 1, \ldots, g$). Kritérium je založené na aproximácii integrovanej klasifikačnej vierohodnosti

$$p(\mathbf{x}, \mathbf{z}) = \int L_c(\boldsymbol{\Psi}) p(\boldsymbol{\Psi}) d\boldsymbol{\Psi}, \tag{2.49}$$

presnejšie jej logaritmu $\log p(\mathbf{x}, \mathbf{z})$. Samotné kritérium má tvar

$$-2 \log L(\hat{\boldsymbol{\Psi}}) + 2EN(\hat{\boldsymbol{\tau}}) + 2n \sum_{i=1}^{g} \hat{\pi}_i \log \hat{\pi}_i + d_1 \log n - 2K(n\hat{\pi}_1, \ldots, n\hat{\pi}_g), \tag{2.50}$$

kde $K(n_1, \ldots, n_g) = \sum_{i=1}^{g} \log \Gamma(n_i + \alpha) - \log \Gamma(n + g\alpha) - g \log \Gamma(\alpha) + \log \Gamma(g\alpha)$, $n_i = \sum_{j=1}^{n} z_{ij}$ ($i = 1, \ldots, g$), $\mathbf{z} = \hat{\boldsymbol{\tau}}$, d_1 je počet neznámych parametrov v $\boldsymbol{\xi}$ a $\hat{\boldsymbol{\pi}}_i = \frac{1}{n} \tau_i(\mathbf{x}_j; \hat{\boldsymbol{\Psi}})$.

Častokrát sa však ICL kritérium nedá použiť priamo, kvôli veľkým hodnotám funkcie $K(\cdot)$ v prípade veľkých $n\hat{\pi}_i$. Vtedy sa používa modifikovaná verzia ICL kritéria, označovaná ako ICL-BIC kritérium (niekedy len BIC). Táto pre veľké hodnoty $n\hat{\pi}_i$ aproximuje gama funkciu v $K(n\hat{\pi}_1, \ldots, n\hat{\pi}_g)$ použitím Stirlingovej formule

$$\Gamma(u) \approx \sqrt{2\pi} u^{u+\frac{1}{2}} e^{-u}.$$

V prípade, že $\alpha = 1$ a po vynechaní členu rádu $O(1)$ dostávame

$$K(n\hat{\pi}_1, \ldots, n\hat{\pi}_g) \approx n \sum_{i=1}^{g} \hat{\pi}_i \log \hat{\pi}_i - \frac{1}{2}(g-1) \log n \tag{2.51}$$

Záverečná podoba ICL-BIC kritéria potom je

$$-2 \log L(\hat{\boldsymbol{\Psi}}) + 2EN(\hat{\boldsymbol{\tau}}) + d \log n, \tag{2.52}$$

kde $d = d_1 + (g-1)$ je počet neznámych parametrov v $\boldsymbol{\Psi}$.

2.5.5 Konzistencia v prípade odhadov MLE a MPLE

Konzistencia štatistického odhadu je jednou z jeho základných vlastností, bez ktorej je jeho zmysluplnosť vždy vážne ohrozená, alebo prinajmenšom teoreticky nepodložená. V tejto časti sa budem venovať konzistencii zmesovej distribúcie v prípade MLE a MPLE odhadu.

V prípade tejto konzistencie treba upresniť, že pracujeme s *topológiou slabej konvergencie na priestore distribučných funkcií.* Teda odhad $\hat{H}$ je konzistentný, ak s pravdepodobnosťou jedna $\hat{H}$ slabo konverguje ku skutočnej H, H^*, pre $n \to \infty$, kde n je počet pozorovaní vo výbere (píšeme $\hat{H} \xrightarrow{w} H^*$). V prípade platnosti miernych podpokladov potom platí $\hat{H} \xrightarrow{w} H^* \Rightarrow f_{\hat{H}} \to f_{H^*}$, kde f_H je hustota zmesi prislúchajúca zmesovej distribúcii H (Leroux, 1992).

Konzistentný odhad zmesovej distribúcie v prípade MLE je relatívne jednoduchá úloha. Medzi **prvé práce venované asymptotike** tohoto odhadu zmesovej distribúcie, ktoré sa venovali všeobecnému modelu zmesi, patrí Kiefer, Wolfowitz (1956), Simar (1976), Laird (1978), Lindsay (1983), pričom v poslednom uvedenom článku autor ukázal, že MLE zmesovej distribúcie má najviac K komponentov, kde K je počet navzájom rôznych pozorovaní vo výbere. V tomto článku bola zavedená aj existencia odhadu metódou obmedzenej maximálnej vierohodnosti (*constraint* MLE), čo je odhad s vierohodnosťou maximalizovanou len cez modely s K alebo menším počtom komponentov.

Asymptotikou MLE odhadu zmesovej distribúcie **pre prípad parametricky definovanej konečenej zmesi** (1.8) sa venoval Sundberg (1974), Redner (1981), Redner, Walker (1984), Hathaway (1985). Najdôležitejšie výsledky však pochádzajú z článku Leroux (1992).

Ak však pre odhad zmesovej distribúcie platí $\hat{H} \xrightarrow{w} H^*$, nie je ešte zrejmé, že sú konzistentné aj odhady jednotlivých paramterov zmesi $\hat{\pi}_1, \ldots, \hat{\pi}_{\hat{g}}$, $\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_{\hat{g}}$, alebo aj $\hat{g}$ v prípade MPLE, teda že $\hat{\pi}_i \xrightarrow{n\to\infty} \pi_i^*$, $\hat{\boldsymbol{\theta}}_i \xrightarrow{n\to\infty} \boldsymbol{\theta}_i^*$, $\hat{g} \xrightarrow{n\to\infty} g^*$. Spomínaný článok Leroux (1992) sa venuje aj **konzistencii** týchto **paramterov**, a to aj pre odhad **MPLE** (kritérium AIC a BIC). Výsledky môžeme zhrnúť asi takto:

- ak poznáme počet zhlukov zmesi, je odhad parametra konzistentný, teda $\hat{\pi}_i \xrightarrow{n\to\infty} \pi_i^*$ *s.i.* a $\hat{\theta}_i \xrightarrow{n\to\infty} \theta_i^*$ *s.i.*
- ak počet zhlukov nepodceníme, je odhad zmesovej distribúcie konzistentný
- ak počet zhlukov podceníme, bude odhad zmesovej distribúcie tiež v istom zmysle konzistentný - pôjde o odhad distribučnej funkcie zmesi s týmto počtom komponentov, ktorá je najbližšie skutočnej distribučnej funkcii zmesi (v zmysle minimálnej divergencie)
- ak k odhadu počtu zhlukov použijeme kritérium AIC a BIC[25], dostaneme odhad, ktorý nepodceňuje počet zhlukov

[25]Leroux vo svojom článku použil trochu obecnejšiu penalizáciu, ktorej vyhovuje kritérium AIC aj BIC.

Veta je síce dokázaná pre jednorozmerný prípad parametra θ ($\theta \in \mathbb{R}$), ale zrejme by sa jej výsledky dali bez problémov zobecniť aj pre jeho viacrozmernú variantu ($\theta \in \mathbb{R}^d$). Uveďme si ešte predpoklady, za ktorých je tento výsledok dokázaný:

1. $f(x,\theta)$ je spojitá na $E \times \Theta$, kde E aj Θ sú borelovské podmnožniny euklidovského priestoru.
2. Pre každý kompakt $C \in E$ a $\epsilon > 0$ existujú také $a,b \in \Theta$, že $f(x,\theta) < \epsilon$, $\theta \in \Theta \backslash [a,b]$, $x \in C$.
3. $f(x,\theta) > 0$ pre každé $x \in E$, $\theta \in \Theta$.
4. Existuje funkcia $h(x)$ spojitá na E, pre ktorú platí $f(x,\theta) \leq h(x)$, $\theta \in \Theta$, $x \in E$ a $\int f_{H^*} |\log h| d\mu < \infty$.
5. $\int f_{H^*} [\log f(x,\theta)]^- d\mu(x) < \infty$, $\theta \in \Omega$ ($x^- = \max\{-x,0\}$).

Podmienky 4 a 5 sú trochu silnejšie ako predpoklad konečnej entropie ($\int f_{H^*} |\log f_{H^*}| d\mu < \infty$), ktorý sa často používa v súvislosti s MLE. Pre exponenciálne, poissonove a normálne (parameter $\theta = \mu$) zmesi je splnenie predpokladov 1-5 ľahko overiteľné. Okrem týchto podmienok musí byť ešte samozrejme splnený predpoklad identifikovateľnosti.

Zaujímavý výsledok dosiahol ešte Chen (1995), ktorý ukázal, že najlepšia možná rýchlosť konvergencie parametrov zmesi pri nadhodnotení počtu komponentov je rádu $n^{-1/4}$, pričom v prípade, že je počet komponentov správny je táto rýchlosť rádu $n^{-1/2}$.

2.6 MPDE

Metódy založené na minimálnej vzdialenosti boli predstavené v prvej kapitole. V tejto časti sa budeme zaoberať tým, ako pomocou tohoto prístupu možno odhadnúť aj počet komponentov v zmesi. Hoci vo väčšine prípadov vedie táto metóda teoreticky ku konzistentnému odhadu zložitosti, nevýhodou je, že (vo všeobecnosti) je problematické metódu efektívne prakticky implementovať .

2.6.1 Henna

Prvé zaujímavé riešenie konzistentného odhadu počtu komponentov poskytol Henna (1985).

Uvažujeme prípad, keď komponenty parametricky formulovanej zmesi (1.8) pochádzajú z jednorozmerného spojitého rozdelenia. Definujme množinu všetkých zmesových distribúcií konečnej zmesi s počtom komponentov g

$$\mathcal{H}_g = \{H_g : H_g = H_g(\boldsymbol{\Psi});\ \boldsymbol{\Psi} = (\pi_1,\ldots,\pi_g,\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_g),$$
$$\sum_{i=1}^{g} \pi_i = 1,\ 0 \leq \pi_i \leq 1,\ \boldsymbol{\theta}_i \in \mathbb{R}^p,\ i = 1,\ldots,g\},$$

kde $H_g(\pi_1,\ldots,\pi_g,\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_g)$ je diskrétna distribučná funkcia so skokmi π_i v $\boldsymbol{\theta}_i$ ($i =$

$1,\ldots,g$). Definujme odhad zmesovej distribúcie zmesi s počtom komponentov g ako

$$\hat{H}_{g,n} = \min_{H_g \in \mathcal{H}_g} D_n(H_g), \tag{2.53}$$

kde

$$\begin{aligned} D_n(H_g) &= \int \{F_{H_g}(x) - F_n(x)\}^2 dF_n(x) \\ &= \frac{1}{n}\sum_{j=1}^{n}\left\{\sum_{i=1}^{g}\pi_i F_{\boldsymbol{\theta}_i}(X_{(j)}) - \frac{j}{n}\right\}^2 \end{aligned} \tag{2.54}$$

je Cramér-von-Misesova vzdialenosť distribučnej funkcie pre zmes so zmesovou distribúciou H_g a empirickej distribučnej funkcie $F_n(x)$. V (2.54) $X_{(j)}$ značí i-tu poradovú štatistiku $\mathbf{X} = (X_1,\ldots,X_n)$ a index n zdôrazňuje, že odhad je skonštruovaný na základe pozorovaní o rozsahu n. Existencia $\hat{H}_{g,n}$ je garantovaná tým, že $D_n(H_g)$ je spojitá funckia v H_g na kompakte $\mathbb{R}^{g(p+1)}$.

Henna ukázal, že potom

- pre ľubovoľné n platí $D_n(\hat{H}_{g,n}) \geq D_n(\hat{H}_{g+1,n})$ a
- v prípade identifikovateľnosti **(Id)** odhad počtu komponentov definovaný ako

$$\hat{g}_n = \min\{g : D_n(\hat{H}_{g,n}) < \lambda^2(n)/n\}, \tag{2.55}$$

kde $\lambda(n) \uparrow \infty$, $\lambda^2(n)/n \to 0, n \to \infty$ a $\sum\{\lambda^2(n)/n\}e^{-2\lambda^2(n)} < \infty$, je odhadom konzistentným (podmienky uložené na funkciu $\lambda(n)$ spĺňa napr. funkcia $\log n$).

Počet komponentov zmesi teda volíme ako najmenšiu hodnotu g, pre ktorú je Cramér-von-Misesova vzdialenosť distribučnej funkcie odhadnutej zmesi (s počtom komponentov g) a EDF menšia ako stanovená hranica (ktorá s rastúcim n klesá k nule). V reálnom prípade je rozsah pozorovaní n konštantný, a teda aj hranica $\lambda^2(n)/n$ je konštanta. Pre praktické použitie je teda často potrebné vynásobiť hranicu vhodnou konštantou $\delta > 0$, aby hranica aspoň rádovo odpovedala minimalizovanej vzdialenosti[26]. V praxi je teda kvalita odhadu značne poškodená.

2.6.2 Chen

Chen, Kalbfleisch (1996) namiesto stanovenia hranice „dostatočnej blízkosti", volia počet komponentov tak, že minimalizujú penalizovanú vzdialenosť EDF a distribučnej funkcie odhadnutej zmesi. Namiesto klasickej penalizácie založenej na počte parametrov v modeli však penalizujú zmesové distribúcie H s malými hodnotami π_i. Pri nadhodnotení počtu komponentov sú niektoré komponenty v zmesi zastúpené len minimálne, teda niektoré π_i sú potom veľmi malé, čo je odôvodnením tejto penalizácie. Okrem toho, že

[26]Napríklad pre rozsa $n = 100$ je hranica $\log^2(100)/100$ približne 0.2 pričom uvažovaná vzdialenosť môže byť pri tomto rozsahu rádovo v tisícinách. Pre praktické použitie by sme teda mali použiť aspoň hodnotu $0.001 * 0.2$ resp. menšiu.

odhad zmesovej distribúcie $\hat{H}$ je v tomto prípade konzistentný, je navyše konzistentný aj odhad počtu komponentov $\hat{g}$. Autori podobne ako tomu bolo v predchádzajúcom prípade pracujú so zmesou spojitých rozdelení.

Penalizovaná vzdialenosť je definovaná ako

$$D(F_n,F_{H_n}) = d(F_n,F_{H_n}) - c_n \sum_{i=1}^{k} \log p_i \tag{2.56}$$

kde $H_n(\boldsymbol{\theta}) = \sum_{i=1}^{g} \pi_i \mathbf{I}_{[\boldsymbol{\theta}_i \leq \boldsymbol{\theta}]}$, $d(F_n,F_{H_n})$ je vzdialenosť na priestore pravdepodobnostných rozdelení, taká že $d(H_n,H) \to 0 \Rightarrow H_n \xrightarrow{w} H$, parametre $\boldsymbol{\theta}_i$ $(i = 1,\ldots,g)$ sú navzájom rôzne a c_n je vhodne zvolená postupnosť kladných konštánt. Za modifikovaného predpokladu identifikovateľnosti, ktorý má v tomto prípade tvar

$$d(F_{H_n},F_H) \to 0 \Rightarrow d(H_n,H) \to 0 \tag{2.57}$$

a za predpokladu $c_n = o(1)$ platí

1. ak $d(F_n,F_{H^*}) = O(c_n)$, potom $\hat{H}_n \xrightarrow{w} H^*$,
2. ak $d(F_n,F_{H^*}) = o(c_n)$, potom $\hat{g}_n \xrightarrow{n\to\infty} g^*$,

kde H^* je skutočná distribučná funkcia zmesi (teda $F^* = F_{H^*}$ je skutočná distribučná funkcia z ktorej pochádzajú vyšetrované dáta), g^* je skutočný počet komponentov a notácia malé o a veľké O je v zmysle skoro isto. Podmienky kladené na vzdialenosť d spĺňa väčšina bežných vzdialeností, ako napr. Kolmogorov-Smirnovova (KS) vzdialenosť, Cramér-von-Misesova (CvM) vzdialenosť, Kullback-Leiblerova (KL) vzdialenosť a iné.

Voľba penalizácie v (2.56) môže byť volená aj inak, podstatné pre zachovanie konzistencie je, aby penalizácia rástla do nekonečna, ak sa $\min \hat{p}_i$ blíži k nule. Alternatívou tak môže byť penalizácia $\sum p_i^{-1}$.

Aby sme však túto metódu mohli efektívne implementovať, je nutné vhodne zvoliť postupnosť konštánt c_n. Nie pre každú vzdialenosť d však poznáme rád konvergencie EDF ku skutočnej distribučnej funkcii, hoci napr. pre suprémovú vzdialenosť je známe, že $KS(F_n,F^*) = O(n^{-\frac{1}{2}}\sqrt{\log\log n})$. Väčším problémom je však fakt, že ak môžeme voliť $\{c_n\}$, je vhodná aj voľba $\{\delta c_n\}$. V oboch prípadoch totiž metóda bude konzistentná, ale môže sa stať, že pri nevhodnej voľbe $\{c_n\}$ bude konvergencia $\hat{g}_n \to g^*$, $n \to \infty$ príliš pomalá, a teda prakticky nepoužiteľná. Ide vlastne o podobný problém ako v predchádzajúcom prípade. Dá sa povedať, že tento problém je charakteristický pre odhady MPDE.

2.6.3 AKM algoritmus

Idea algoritmu AKM (Alternate kernel and mixture density estimate) je založená na striedaní parametrického odhadu hustoty, v tomto prípade reprezentovaného zmesou hustôt (parametricky formulovanou) a neparametrického odhadu hustoty, na ktorý sa použije *filtrovaný jadrový odhad* (FKE).

FKE (Marchette a kol., 1996) kombinuje klasický jadrový odhad (s normálnym jadrom) $\hat{f}(x,h) = 1/nh \sum_{j=1}^{n} \varphi_h((x - X_j)/h)$ s odhadom zmesi (v tomto prípade zmes

normálnych rozdelení) $f(x,\hat{g},\hat{\Psi}) = \sum_{i=1}^{g} \hat{\pi}_i \varphi(x;\hat{\boldsymbol{\theta}}_i)$, kde $\varphi(\cdot,\boldsymbol{\theta})$ je hustota normálneho rozdelenia s parametrami $\boldsymbol{\theta} = (\mu,\sigma^2)$ a φ_h je hustota $N(0,h)$. Všeobecne je FKE definovaný ako

$$\hat{f}(x) = 1/n \sum_{j=1}^{n} \sum_{i=1}^{m} \frac{\rho_i(x)}{h_i} K(\frac{x - X_j}{h_i}),$$

kde $(\rho_i)_{i=1}^{m}$ je filter, $\sum \rho_j = 1$. Hlavnou odlišnosťou od klasického jadrového odhadu je použitie viacerých vyhladzovacích parametrov (šírky pásma). Filtrovaný jadrový odhad založený na odhade zmesi $f(x,\hat{g},\hat{\Psi}) = \sum_{i=1}^{g} \hat{\pi}_i \varphi(x;\hat{\mu}_i,\hat{\sigma}_i^2)$ má tvar

$$\hat{f}(x;h,\hat{g},\hat{\Psi}) = \frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{\hat{g}} \frac{\hat{\pi}_i \varphi(X_j;\hat{\mu}_i,\hat{\sigma}_i^2)}{h\hat{\sigma}_i f(X_j;\hat{g},\hat{\Psi})} \varphi((x - X_j)/(h\hat{\sigma}_i)). \tag{2.58}$$

Tento jadrový odhad sa pre každý komponent zmesi snaží používať inú (v istom zmysle optimálnu) šírku pásma podľa rozptylu toho ktorého komponentu. Voľba optimálneho h je založená na minimalizácii asymptotickej strednej integrovanej štvorcovej chyby (AMISE) za podmienky, že filtrovacia zmes je skutočná.

Popíšme si teraz AKM algoritmus (Priebe, Marchette, 2000). Začneme predpokladmi a notáciou. Nech $d(\cdot,\cdot)$ reprezentuje nejakú vzdialenosť na priestore hustôt pravdepodobnostných rozdelení a nech $\mathcal{F}_g$ označuje rodinu normálnych rozdelení s počtom komponentov g a s obmedzeným rozptylom $\sigma_i^2 \in [l_g,u_g]$ $(i = 1,\ldots,g)$. Skratkou FKE budem označovať filtrovaný jadrový odhad a skratkou FMM zase odhad konečnej zmesi. Pseudokód algoritmu AKM je potom definovaný ako:

1. spočítaj štandardný normálny odhad $\hat{f}_{FMM}^{(1)}(x) \equiv \varphi(x;\overline{X},S^2)$ a štandardný jadrový odhad $\hat{f}_{FKE}^{(1)}(x) \equiv \hat{f}(x;h_{opt},1,\hat{\Psi}^{(1)})$
2. $\hat{f}_{FMM}^{(2)} \equiv \arg\min_{f\in\mathcal{F}_2} d(f,\hat{f}_{FKE}^{(1)})$ a $g \leftarrow 1$
3. Ak $d(\hat{f}_{FMM}^{(g)},\hat{f}_{FMM}^{(g+1)}) \geq c > 0$ choď na (4) inak choď na (8)
4. $g \leftarrow g + 1$
5. $\hat{f}_{FKE}^{(g)} \equiv \hat{f}(x;h_{opt},g,\hat{\Psi}^{(g)})$
6. $\hat{f}_{FMM}^{(g+1)} \equiv \arg\min_{f\in\mathcal{F}_{g+1}} d(f,\hat{f}_{FKE}^{(g)})$
7. choď na (3)
8. $\hat{g}_n \leftarrow g$

Algoritmus v každom kroku teda minimalizuje vdialenosť hustoty pre normálnu zmes s počtom komponentov g od filtrovaného jadrového odhadu založeného na hustote normálnej zmesi s počtom komponentov $g - 1$. Okrem odhadu zložitosti zmesi $\hat{g}_n$ dáva algoritmus aj odhad hustoty príslušnej zmesi $\hat{f}_{FMM}^{(\hat{g}_n)} \in \mathcal{F}_{\hat{g}_n}$ (a tiež filtrovaný jadrový odhad $\hat{f}_{FKE}^{(\hat{g}_n)}$ založený na zmesi $\hat{f}_{FMM}^{(\hat{g}_n)}$). Pre spustenie algorimtu treba zvoliť hranice obmedzujúce rozptyl jednotlivých komponentov $[l_g,u_g]$ $g = 1,2,\ldots$, ďalej konštantu c, ktorá určuje koniec iterovania a napokon ešte vzdialenosť $d(\cdot,\cdot)$, ktorou môže byť napr. integrovaná štvorcová vzdialenosť (L_2 vzdialenosť) $d(f,g) = \int (f - g)^2$.

V prípade, že konštanta c_n s rastúcim n konverguje k nule dostatočne pomaly, je odhad zmesi algoritmom AKM konzistentný, a naviac, ak skutočná hustota je konečná

zmes normálnych rozdelení, je konzistentý aj odhad počtu komponentov a jednotlivých zvyšných parametrov zmesi. Podobne ako pri iných MPDE odhadoch je ale problém s riešením úlohy $\arg\min_{f\in\mathcal{F}_{g+1}} d(f,\hat{f}^{(g)}_{FKE})$, ktorá predstavuje ťažkú úlohu nelineárnej optimalizácie. Taktiež je otázna voľba konštanty c_n vzhľadom k tomu, že jedniná podmienka $c_n \to 0$, $n \to \infty$ nevytvára v prípade konečného n na c_n žiadne obmedzenie.

2.6.4 MPDE v prípade jadrového odhadu hustoty

James a kol. (2001) navrhli ďalší prístup pre odhad počtu komponentov jednorozmernej normálnej zmesi založený na minimalizácii hustoty zmesi od jadrového odhadu hustoty. Vzhľadom k tomu, že klasický jadrový odhad hustoty je vychýlený, pričom vychýlenie je rádu $O(h^2)$ (h je šírka pásma), autori volili odhad hustoty normálnej zmesi na priestore normálnych zmesí s počtom komponentov g ($\mathcal{F}_g$) ako

$$\hat{f}^{(g)} = \arg\min_{f\in\mathcal{F}_g} KL(\tilde{f}_h,\varphi_h * f), \tag{2.59}$$

kde $\tilde{f}_h$ je štandardný jadrový odhad so šírkou pásma h, φ_h je hustota pre rozdelenie $N(0,h)$, $*$ symbolizuje konvolúciu a KL je označenie pre Kullback-Leiblerovu vzdialenosť. Ak $f \in \mathcal{F}_g$, tak $\varphi_h * f$ je hustota normálnej zmesi s rozptylmi jednotlivých komponentov $\sigma_i^2 + h^2$ $(i = 1,\ldots,g)$. Motivácia k zavedeniu konvolúcie prvkov z $\mathcal{F}_m$ s φ_h je jednoduchá, vzhľadom k tomu, že $\tilde{f}_h$ je nevychýlený odhad skôr pre $\varphi_h * f_0$ ako pre f_0, kde f_0 je skutočná hustota. Odhad počtu komponentov potom volíme ako

$$\hat{g}_n = \min\{m : KL(\tilde{f}_h,\varphi_h * \tilde{f}^{(g)} \leq) KL(\tilde{f}_h,\varphi_h * \tilde{f}^{(g+1)} + a_{n,m+1}\}, \tag{2.60}$$

kde $\{a_n,j : j \geq 1\}$ je postupnosť pozitívnych konštánt, pre ktorú platí $a_{n,j} \overset{n\to\infty}{\to} 0$. Autori navrhujú $a_{n,g} = 3/n$.

Procedúra vedie ku konzistentnému odhadu počtu komponentov. Opäť však ide o teoretický poznatok. Podobne ako v predchádzajúcich prípadoch totiž rýchlosť konvergencie závisí na voľbe $a_{n,g}$. Dôležitá je aj presnosť jadrového odhadu, teda voľba šírky pásma. Autori navrhujú pri každej iterácii voliť optimálnu šírku pásma na základe aktuálneho odhadu zmesi.

2.7 Bayesovský prístup

2.7.1 Bayesov faktor

Uvažujme hypotézu H_0, v prípade ktorej dáta pochádzajú zo zmesi s parametrom $\boldsymbol{\Psi}_0$ a H_1, v prípade ktorej dáta pochádzajú zo zmesi s parametrom $\boldsymbol{\Psi}_1$. Bayesov faktor pre H_1 proti H_0 je definovaný ako

$$B_{10} = \frac{P(H_1|\boldsymbol{X})}{P(H_0|\boldsymbol{X})} \Big/ \frac{P(H_1)}{P(H_0)} = \frac{P(\boldsymbol{X}|H_1)}{P(\boldsymbol{X}|H_0)}. \tag{2.61}$$

Je to vlastne pomer integrovaných vierohodností, teda

$$B_{10} = p(\mathbf{x}|H_1)/p(\mathbf{x}|H_0), \text{ kde } p(\mathbf{x}|H_i) = \int p(\boldsymbol{\Psi}_i|H_i)L(\boldsymbol{\Psi}_i|H_i)d\boldsymbol{\Psi}_i \ (i = 0,1). \tag{2.62}$$

$2\log_e(B_{10})$	B_{10}	Evidencia pre H_1
0 do 2	1 do 3	Slabá
2 do 5	3 do 12	Pozitívna
5 do 10	12 do 50	Silná
> 10	> 150	Presvedčivá

Tabuľka 2.1: Sprievodca pre použitie Bayesovho faktora pre výber modelu podľa Kass, Raftery (1995)

Prehľad histórie, vývoja a použitia Bayesových faktorov možno nájsť v Kass, Raftery (1995). V tabuľke 2.1 sa nachádza návrh pre použitie Bayesovho faktora pre výber modelu zmesi podľa týchto autorov. K vyčísleniu logaritmu integrovanej vierohodnosti možno použiť Laplaceovu metódu (2.35), ktorá je však značne komplikovaná. Raftery (1996) navrhol vylešenie pre jej použitie.

Toto kritérium sa niekedy označuje ako *Laplace-Metropolisove kritérium* a možno ho považovať aj za akúsi bayesovskú analógiu kritéria LEC.

2.7.2 Kritérium RW

Uvažujme nejakú hornú hranicu pre počet komponentov zmesi g_{max}, teda nech platí $1 \leq g \leq g_{max}$. V praxi často postačí $g_{max} = 10$. Ak by nás zaujímala hustota, ktorá by dobre popísala dáta, mohli by sme v prípade modelu zmesi použiť odhad $\hat{f}(\boldsymbol{y}) = \sum_g \hat{f}(\boldsymbol{y}|g)p(g|\boldsymbol{x})$, kde $\boldsymbol{x}$ reprezentuje vektor pozorovaní, $\hat{f}(\boldsymbol{y}|g)$ je hodnota hustoty odhadnutej zmesi s počtom komponentov g v bode $\boldsymbol{y}$ a $p(g|\boldsymbol{x})$ je aposteriórne rozdelenie počtu komponentov. Odporúča sa však najskôr odhadnúť počet komponentov a hustotu f odhadnúť ako $\hat{f}(\boldsymbol{y}) = \hat{f}(\boldsymbol{y}|\hat{g})$. Klasický bayesovský prístup na odhadnutie g spočíva v maximalizácii $p(g|\boldsymbol{y})$ cez $1 \leq g \leq g_{max}$.

Podľa Bayesovej vety platí

$$p(g|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|g)p(g)}{p(\boldsymbol{x})}, \tag{2.63}$$

kde $p(g)$ je apriórne rozdelenie počtu komponentov a

$$p(\boldsymbol{x}|g) = \int p(x|\boldsymbol{\Psi}_g)p_g(\boldsymbol{\Psi}_g), \tag{2.64}$$

$\boldsymbol{\Psi}_g$ je parameter $\boldsymbol{\Psi}$ modelu zmesi s počtom komponentov g a $p_g(\boldsymbol{\Psi}_g)$ je apriórne rozdelenie parametra $\boldsymbol{\Psi}$ v modeli s g komponentami.

Keď však za apriórne rozdelenie $p_g(\boldsymbol{\Psi}_g)$ zoberieme rozdelenie definované až na konštantu, teda $p_g(\boldsymbol{\Psi}_g)$ môžeme nahradiť $cp_g(\boldsymbol{\Psi}_g)$ pre ľubovoľné kladné c, zväčší sa c-násobne aj člen (2.64) a následne aj $p(g|\boldsymbol{x})$. Inak povedané, $p(g|\boldsymbol{x})$ obsahuje náhodnú konštantu.

Roeder, Wasserman (1997) navrhli za účelom odstránenia náhodnej konštanty z $p(g|\boldsymbol{x})$ nasledovný postup. Vychádzajú pri ňom z aproximácie použitej pri konštrukcii BIC

kritéria (viď (2.34) a (2.37)), na základe ktorej aproximujú člen $m_g = \int p(\boldsymbol{x}|\boldsymbol{\Psi}_g)p_g(\boldsymbol{\Psi}_g)$ členom

$$\hat{m}_g = n^{-d_g/2}p(\boldsymbol{x}|\hat{\boldsymbol{\Psi}}_g) = n^{-d_g/2}L(\hat{\boldsymbol{\Psi}}_g), \tag{2.65}$$

kde d_g je počet parametrov modelu s g komponentami. Za apriórne rozdelenie počtu komponentov volia rovnomerné rozdelenie, teda $p(g) = 1/g_{max}$ pre $g = 1, \ldots, g_{max}$ a odhad aposteriórneho rozdelenia zložitosti zmesi napokon volia ako

$$\hat{p}(g|\boldsymbol{x}) = \frac{\hat{m}_g}{\sum_h \hat{m}_h}. \tag{2.66}$$

Pre výpočet $\hat{m}_g$ potrebujeme odhad parametra $\hat{\boldsymbol{\psi}}_g$, pre ktorý môžeme použiť strednú hodnotu alebo modus aposteriórneho rozdelenia tohoto paramtera, ktoré získame aplikáciou MCMC metód.

Toto kritérium pre prehľadnosť označím symbolom RW[27]. Autori ukázali, že odhad aposteriórneho rozdelenia zložitosti zmesi je konzistentný v prípade, že skutočný počet komponentov zmesi nie je väčší ako uvažovaná horná hranica g_{max}. Konzistenciu napokon ukázali aj v prípade, že g_{max} rastie spolu s n, pričom však $g_{max} \equiv g_{max}(n)$ nesmie rásť príliš rýchlo (rád $o(n/\log n)$ zaručuje ešte konzistentnosť).

2.7.3 MCMC metódy

Phillips, Smith (1996) a Richardson, Green (1997a) navrhli použitie MCMC metód pre neznámy počet komponentov zmesi. Ide o pomerne komplikované metódy, vzhľadom k tomu, že MCMC sú aplikované na priestor s premenlivou dimenziou. V oboch prípadoch je za apriórne rozdelenie počtu komponentov g brané Poissonove rozdelenie useknuté v počiatku

$$p(g) = \frac{\lambda^g}{(e^\lambda - 1)g!}, \quad g = 1,2,\ldots.$$

V oboch prípadoch sa tiež podarilo zhotoviť spojité aposteriórne rozdelenia pre g a $\boldsymbol{\Psi}$. Algoritmus odvodený v prípade prvého uvedeného článku sa nazýva *jump-diffusion sampling* a v prípade druhého zase *reversible-jump Metropolis-Hastings*. Tento druhý algoritmus používa buď techniku *rodenia a zániku zhluku*, alebo techniku *rozdeľovania alebo zlučovania zhlukov*. Podrobne sa mu venuje Nemeček (2004).

U týchto MCMC metód je však problematické diagnostikovať konvergenciu a nevýhodou je aj ich komplikovanosť.

2.8 MLE - najnovšie výsledky

2.8.1 Konzistentný odhad počtu komponentov

Konzistencia odhadu MPLE bola dlho veľkým problémom. Veľkým prínosom v tejto oblasti bol už citovaný článok Leroux (1992). Tomu sa podarilo za určitých (nie príliš obmedzujúcich) predpokladov odvodiť, že odhad metódou penalizovanej vierohodnosti

[27] V odbornej literatúre sa nevyskytuje pod žiadnym špeciálnym názvom.

(MPLE) s penalizačnými členmi AIC a BIC asymptoticky nepodceňuje počet zhlukov (v zmysle skoro isto). S nadhodnotením počtu zhlukov ale už bol problém. Zádrhel bol v tom, že model je pri nadhodnotení neidentifikovateľný.

Dacunha-Castelle, Gassiat (1997, 1999) zaviedli špeciálnu reparametrizáciu zmesového modelu (*locally conic reparametrization*), pri ktorej je jeden parameter identifikovateľný úplne (aj v bode, v ktorom bol pri pôvodnej parametrizácii modelu neidentifikovateľný) a zvyšné parametre zahŕňajú celú neidentifikovateľnosť. V takomto prípade je možné použiť Taylorov rozvoj vierohodnosti v okolí identifikovateľného parametra. Tento prístup napokon umožnil odvodiť teoretické rozdelenie pre LRTS. Keribin (2000) urobil ďalší krok vpred, keď použil túto reparametrizáciu na odvodenie asymptotického správania odhadu MPLE.

Predpokladajme, že:

(Id) rodina rozdelení je identifikovateľná v zmysle **(Id)** a

(P) spĺňa niekoľko ďalších podmienok[28],

(C1) pre penalizačný člen $a_{n,g}$ (n-rozsah výberu, g počet komponentov) nech platí:

- $a_{n,g+1} \geq a_{n,g} > 0$ pre všetky $n \geq 1$,
- $\lim_{n\to\infty} a_{n,g} = +\infty$, $a_{n,g} = o(n)$ a
- $\lim_{n\to\infty} a_{n,g}/a_{n,g'} > 1$ pre $g' < g \leq G$, kde G je nejaká známa horná hranica pre počet komponentov.

Výsledky Keribina môžeme potom zhrnúť nasledovne:

(K1) Odhad počtu komponentov $\hat{g}$ konverguje v pravdepodobnosti ku skutočnému počtu komponentov g_0, teda $P(\lim_{n\to\infty} \hat{g}_n = g_0) = 1$.

(K2) V prípade trochu slabších predpokladov ako použil Leroux odhad počtu komponentov nepodhodnocuje v pravdepodobnosti skutočný počet komponentov, teda $P(\lim_{n\to\infty} \inf \hat{g}_n \geq g_0) = 1$.

(K3) Ak trochu rozšírime predpoklad (P) a pridáme ďalší predpoklad na penalizačný člen (C2) $\lim_{n\to\infty} \log\log n/a_{n,g} = 0$, tak potom platí, že

(a) $\hat{g}_n$ konverguje k g_0 *s.i.*,

(b) $(\hat{\boldsymbol{\pi}}_{\hat{g}}, \hat{\boldsymbol{\theta}}_{\hat{g}})$ konverguje k $(\boldsymbol{\pi}_0, \boldsymbol{\theta}_0)$ *s.i.*

Keribin vo svojom článku overil predpoklady potrebné ku konzistentnému odhadu počtu komponentov (výsledok (K3)) pre rodinu normálnych (pre prípad $\theta = \mu$ aj $\boldsymbol{\theta} = (\mu, \sigma^2)$) a Poissonových rozdelení. Ako vhodný penalizačný člen, ktorý vyhovuje Keribinovým

[28] Podmienky sú výrazne analytického charakteru a tiež komplikované na overenie.

podmienkam, sa ponúka penalizácia tvaru $a_{n,g} = \frac{g}{2}(\log n)^{\alpha}$, $\alpha = \frac{1}{2}$, 1, 2, 3,...[29]. Toto kritérium v tvare

$$-2\log L(\Psi) + \frac{d}{2}(\log n)^{\alpha}, \tag{2.67}$$

kde d je počet parametrov modelu, pre účely práce označím ako KER kritérium. V nasledujúcej kapitole ukážem dobré vlastnosti tohoto kritéria.

2.8.2 SMEM algoritmus

Ako už vieme EM algoritmus ponúka v prípade modelu so zmesou hustôt efektívne riešenie odhadu. Problém lokálnych maxím vierohodnosti, v ktorých EM algoritmus častokrát končí, je však stále tŕňom v oku tohoto algoritmu. Algoritmus preto zväčša opakujeme niekoľkokrát s inou inicializáciou, aby sme sa vyhli zlým lokálnym maximám, poprípade nejako testujeme, či získaný odhad možno považovať za MLE (teda, či je konzistentný a asymptoticky eficientný).

Naonori Ueda (2000) prišiel s novým prístupom. Všimol si, že lokálne maximá vierohodnosti v prípade (viacrozmerných) normálnych zmesí súvisia aj s tým, že v niektorej časti priestoru sa nahromadí príliš veľa normálnych zhlukov, pričom v inej časti priestoru je zhlukov nedostatok. Z tejto zlej konštalácie rozmiestnenia normálnych komponentov v zmesi nedokáže algoritmus v rozumnom čase vyskočiť, alebo sa mu nepodarí vyskočiť vôbec. Ueda za týmto účelom navrhol *SMEM algoritmus* (split and merge EM). Tento je založený na rozdeľovaní a spájaní určitých zhlukov vytvorených EM krokmi, pri ktorých sa na základe určitých kritérií preukáže, že by mohli byť rozdelené, resp. spojené. Pritom SMEM algoritmus nemení predom daný počet komponentov. Algoritmus v krokoch vyzerá nasledovne:

1. Aplikujeme klasický EM algoritmus. Dostaneme odhad Ψ^*, $Q^* = Q(\Psi^*)$ je stredná hodnota úplnej logaritmickej vierohodnosti podmienenej vektorom pozorovaní v bode Ψ^* (viď (1.32)).

2. Nájdeme kandidátov (kritéria pre ich nájdenie uvediem neskôr) na *zlúčenie a rozdelenie* v tvare trojíc $\{i,j,k\}_c$, $c = 1,\ldots,C_{max}$. i,j sú indexy zhlukov (komponentov), ktoré budú spojené do jedného zhluku a k je index zhluku ($k \neq i$ a $k \neq j$), ktorý bude rozdelený na dva nové. Počet zhlukov v zmesi sa tak procedúrou nezmení .

3. Pre všetky trojice z kroku 2 aplikujeme *parciálnu EM procedúru* (popíšem neskôr) a potom aplikujeme ešte *kompletnú EM procedúru* (normálny EM algoritmus). Dostaneme Ψ^{**} a Q^{**}.

4. Ak $Q^{**} > Q^*$, tak $\Psi^* \leftarrow \Psi^{**}$ a $Q^* \leftarrow Q^{**}$, vrátime sa späť na krok 2.

5. Koniec algoritmu. Odhad je Ψ^*.

[29] Keribin vo svojom článku uvádza, že v prípade $\alpha = 1$ táto penalizácia zodpovedá BIC kritériu, čo však nie je úplne pravda, keďže v BIC kritériu sa namiesto g používa počet neznámych parametrov modelu d. V Keribinovej teórii konzistencii to však nevadí, lebo počet neznámych parametrov je rastúcou funkciou počtu komponentov.

Parciálny EM algoritmus

Parciálna EM procedúra je proces, v ktorom z trojice zhlukov s indexami i,j,k vytvoríme nový zhluk i' spojením zhlukov i,j a nové zhluky j',k' rozdelením zhluku k, a odhadneme parametre týchto zhlukov tak, aby sme neovplyvnili zvyšné zhluky v zmesi. Počiatočné parametre pre zhluk i' skonštruujeme ako

$$\pi_{i'} = \pi_i^* + \pi_j^* \tag{2.68}$$

$$\theta_{i'} = \frac{\theta_i^*\pi_i^* + \theta_j^*\pi_j^*}{\pi_i^* + \pi_j^*}, \tag{2.69}$$

kde θ_i korešponduje s $\boldsymbol{\mu}_i$ a $\boldsymbol{\Sigma}_i$. Počiatočné parametre pre zhluky j' a k' zase skonštruujeme ako

$$\pi_{j'} = \pi_{k'} = \frac{1}{2}\pi_k^* \tag{2.70}$$

$$\boldsymbol{\Sigma}_{j'} = \boldsymbol{\Sigma}_{k'} = \det(\boldsymbol{\Sigma}_k^*)^{1/p}\mathbf{I}_p \tag{2.71}$$

a stredné hodnoty $\boldsymbol{\mu}_{j'}$ a $\boldsymbol{\mu}_{k'}$ získame aplikáciou algoritmu K-means na pozorovania s najväčšou aposteriórnou pravdepodobnosťou príslušnosti do zhluku k. Odhady parametrov pre nové zhluky i',j',k' dostaneme pomocou klasickej EM procedúry, v ktorej ako inicializáciu použijeme (2.68), (2.69), (2.70) a (2.71), pričom za aposteriórnu pravdepodobnosť príslušnosti k zhluku $m' = i',j',k'$ použijeme

$$\tau_{m'}(\mathbf{x},\boldsymbol{\Psi}^{(t)}) = \frac{\pi_{m'}^{(t)}\varphi(\mathbf{x};\boldsymbol{\mu}_{m'}^{(t)},\boldsymbol{\Sigma}_{m'}^{(t)})}{\sum_{l=i',j',k'}\pi_l^{(t)}\varphi(\mathbf{x};\boldsymbol{\mu}_l^{(t)},\boldsymbol{\Sigma}_l^{(t)})} \times \sum_{m=i,j,k}\tau_m(\mathbf{x};\boldsymbol{\Psi}^*). \tag{2.72}$$

Vzťah (2.72) zaručí rovnosť $\sum_{m'=i',j',k'}\tau_{m'}(\mathbf{x},\boldsymbol{\Psi}^{(t)}) = \sum_{m=i,j,k}\tau_m(\mathbf{x},\boldsymbol{\Psi}^*)$, ktorá zaručí, že odhady pre nové zhluky i',j' a k' získame pomocou tejto parciálnej EM procedúry bez toho, aby sme ovplyvnili zvyšné zhluky v uvažovanej zmesi.

Kritéria pre výber kandidátov na zlúčenie a rozdelenie

Maximum trojíc, na ktoré môžeme aplikovať rozdelenie a zlúčenie je $g(g-1)(g-2)/2$ (g je počet komponentov). Autor navrhuje prístup, podľa ktorého tieto trojice vhodne usporiada a pre procedúru rozdelenia a zlúčenia použije 5 najsilnejších kandidátov.

- **Kritérium pre zlúčenie** - dvojicu zhlukov s indexami i a j zlúčime, ak

$$J_{merge}(i,j,\boldsymbol{\Psi}^*) = \frac{T_i(\boldsymbol{\Psi}^*)^T T_j(\boldsymbol{\Psi}^*)}{\|T_i(\boldsymbol{\Psi}^*)\|\,\|T_j(\boldsymbol{\Psi}^*)\|}, \tag{2.73}$$

kde $T_i(\boldsymbol{\Psi}^*) = (\tau_i(\mathbf{x}_1;\boldsymbol{\Psi}^*),\ldots,\tau_i(\mathbf{x}_N;\boldsymbol{\Psi}^*))^T$, bude rozumne veľké (nadobúda hodnoty medzi 0 a 1). Toto kritérium je založené na myšlienke, že ak veľa dát má takmer rovnakú aposteriórnu pravdepodobnosť príslušnosti k dvom zhlukom, potom by tieto mali byť spojené.

- **Kritérium pre rozdelenie** - pre rozdelenie zhluku s indexom k použijeme *lokálnu Kullback Leiblerovu divergenciu*

$$J_{split}(k,\Psi^*) = \int p_k(\mathbf{x};\Psi^*) \log \frac{p_k(\mathbf{x};\Psi^*)}{\varphi(\mathbf{x};\boldsymbol{\mu}_k^*,\boldsymbol{\Sigma}_k^*)} d\mathbf{x}, \tag{2.74}$$

čo je vzdialenosť medzi lokálnou hustotou dát $p_k(\mathbf{x})$ okolo k-teho zhluku a k-tej normálnej hustoty danej aktuálnym odhadom Ψ^*. Lokálna hustota je definovaná ako

$$p_k(\mathbf{x},\Psi^*) = \frac{\sum_{j=1}^{N} \delta(\mathbf{x}-\mathbf{x}_j)\tau_k(\mathbf{x}_j;\Psi^*)}{\sum_{j=1}^{N} \tau_k(\mathbf{x}_j;\Psi^*)}, \tag{2.75}$$

čo je vlastne hustota modifikovaného empirického rozdelenia váženého aposteriórnymi pravdepodobnosťami tak, aby dáta okolo k-teho normálneho komponentu mali väčšiu váhu. Zhluk s najväčšou hodnotou J_{split} má teda najhorší odhad lokálnej hustoty, a preto sa ho pokúsime rozdeliť.

Kandidátov na zlúčenie a rozdelenie usporiadame nasledovne. Najskôr zotriedime dvojice na zlúčenie $\{i,j\}_c$ podľa kritéria J_{merge} a ku každej zatriedenej dvojici $\{i,j\}$ pridáme zotriedených kandidátov na rozdelenie $\{k\}_c$ (s výnimkou zhlukou i,j) podľa kritéria J_{split}. Podľa autora je algoritmus SMEM zvlášť vhodný pre reálne viacrozmerné dáta s veľkou dimenziou.

V článku (Wang a kol., 2004) sa autori pokúsili použiť techniku rozdeľovania a zlučovania zhlukov v EM algoritme k odhadu počtu komponentov modelu. Túto modifikáciu EM algoritmu nazvali SSMEM (stepwise split and merge EM).

Kapitola 3

Model zmesi v praxi

V tejto kapitole sa pokúsim zhrnúť to, čo sme sa o metódach odhadu zmesi a najmä odhadu počtu komponentov dozvedeli na predchádzajúcich stránkach. Niektoré metódy som sa pokúsil implementovať v prostredí R, ktoré dnes čitateľovi zrejme nemusím predstavovať. Stručne predstavím zhotovené algoritmy. Okrem toho som sa vzhľadom na veľkú rozšírenosť a stále narastajúcu obľúbenosť tohoto jazyka rozhodol uviesť prehľad funkcií a metód pre prácu s modelmi zmesí, ktoré v R-ku môžeme nájsť. Na záver pridám ešte simulácie porovnávajúce vybrané metódy pre odhad zložitosti zmesi a pokúsim sa na reálnom príklade dvojrozmerných dát odhadnúť model zmesi.

3.1 Zopakujme si

Medzi najdôležitejšie prístupy pre odhad počtu komponentov zmesi patrí penalizovaný maximálne vierohodný odhad, odhad založený na minimálnej vzdialenosti (penalizácia alebo stanovanie hranice dostatočnej blízkosti), Bayesovské odhady a testovanie hypotézy o počte komponentov (predovšetkým test pomerom vierohodností). U metódy momentov sa nepodarilo preukázať konzistenciu a dnes sa teda používa zväčša len na určenie vhodnej inicializácie nejakého z iteratívnych aloritmov, grafickým metódam zase absentuje štatistický test.

Minulosť

V prípade **maximálne vierohodného odhadu** bol zlomom vo vývoji objav algoritmu EM, ktorý umožnil efektívne maximalizovať vierohodnosť zmesi pre známy počet komponentov. Vzhľadom k tomu, že s rastom počtu komponentov modelu jeho vierohodnosť rastie, bolo potrebné prírastok počtu komponentov v nejakom okamihu zastaviť. Vývoj sa potom uberal voľbou vhodného penalizačného členu, ktorý kvantifikoval nevhodnosť daného modelu, a spolu s vierohodnosťou modelu tak vytvoril kritérium pre voľbu počtu komponentov zmesi (informačné kritériá). Hlavným nedostatkom tohoto prístupu bol fakt, že sa nepodarilo preukázať konzistenciu odhadu. Podarilo sa však úspešne vyriešiť problém podhodnotenia počtu komponentov, a keďže existoval EM algoritmus, vhodná voľba penalizácie (ktorá by nenadhodocovala počet komponentov, alebo ho nadhodno-

covala čo najmenej) by poskytla efektívne riešenie. Toto bolo hnacím motorom vzniku rozličných informačných kritérií, z ktorých niektoré sa uberali príliš komplikovanou cestou, iné zase vykazovali príliš nejasné vlastnosti. Najlepšie výsledky v simuláciách dlhodobo dosahovalo kritérium BIC, ktoré sa zároveň vyznačuje aj jednoduchosťou. Toto kritérium napokon pretrvalo až do dnešných čias a v súčasnosti je to nepochybne najčastejšie používané kritérium pre odhad počtu komponentov v modeli zmesi.

Odhad založený na minimálnej vzdialenosti je presným opakom maximálne vierohodného odhadu. Konzistencia sa totiž v tomto prípade podarila dokázať vo väčšine prípadov, problémom však bola a stále je absencia efektívneho algoritmu na minimalizáciu vzdialenosti empirickej distribučnej funkcie a distribučnej funkcie hľadanej zmesi, resp. vzdialenosti odhadu hustoty zmesi od hustoty hľadanej zmesi. Tento problém totiž reprezentuje náročnú úlohu nelineárnej optimalizácie, ktorú riešime buď nejakou metódou stochastickej optimalizácie, alebo viacnásobnou optimalizáciou inicializovanou v rôznych počiatočných hodnotách. V prípade algoritmu HMIX je okrem numerickej maximalizácie dokonca potrebná aj numerická integrácia.
Okrem toho je problém s voľbou penalizácie v reálných úlohách. Pre použitie MPDE v praxi je potrebné si najskôr urobiť predstavu o tom, v akom rozsahu sa minimalizovaná vzdialenosť v našom konkrétnom prípade pohybuje, a podľa toho zvoliť penalizáciu alebo hranicu dostatočnej blízkosti. Tento zásah však značne zpochybňuje relevantnosť praktickej implemetácie týchto odhadov.

K pozitívnym vlastnostiam odhadov založených na minimalizácii vzdialenosti patrí ich robustnosť[30] (naznačená v Cutler, Cordero-Braña (1996)).

Testovanie hypotézy o počte komponentov založené na pomere vierohodností zase narazilo na problém asymptotického rozdelenia testovej štatistiky. To sa nepodarilo odvodiť pomocou klasickej asymptotickej teórie pre vierohodnosť, pretože model zmesi nespĺňa pre ňu potrebné predpoklady. Vývoj sa potom zameral hlavne na simulácie, ktoré sa snažili preskúmať rozdelenie LRTS. Vo výsledkoch jednotlivých autorov, ktorí chceli výsledky simulácií konfrontovať (či skôr porovnávať) s rozdelením χ^2, sa však vyskytli určité nezhody (viď napr. záver článku Schlattmann (2003)). Rozdelenia LRTS sa nakoniec podarilo dobre aproximovať použitím bootstrapu, ktorý je však v tomto prípade príliš výpočtovo náročný, a tak neponúka efektívne riešenie. Napokon sa v roku 1999 podarilo všeobecne odvodiť teoretické rozdelenie štatistiky LRTS (Dacunha-Castelle, Gassiat, 1999), ale toto nie je vhodné pre praktické použitie.

Okrem toho existuje ešte množstvo článkov venovaných **testovaniu homogenity**, ktoré sa zväčša venujú rôznym špeciálnym tvarom rodín alebo špeciálnym prípadom.

Bayesovské metódy vďaka existencii MCMC metód umožňujú dobre a aj celkom efektívne odhadnúť aposteriórne rozdelenia parametrov zmesi v prípade známeho počtu komponentov. Pre odhad počtu komponentov existujú tri hlavné metódy. Použitie Bayesovho faktora, ktoré je ale náročné na výpočet a nie je u neho diagnostikovaná

[30] Otázka robustnosti je dnes rozoberaná aj v prípade vierohodnotného prístupu (napr. použitie zmesi t-rozdelení namiesto normálych rozdelení - McLachlan, Peel (1998)).

konzistencia. Ďalej kritérium RW, ktoré k odvodeniu aposteriórneho rozdelenia zložitosti zmesi používa aposteriórne rozdelenie parametra Ψ_g pre modely s $g = 1, \ldots, g_{max}$ komponentami. Je teda náročné na implementáciu. Aposteriórne rozdelenie zložitosti modelu je ale konzistné, a to dokonca aj pre prípad, keď počet komponentov rastie s počtom pozorovaní s obmedzením na rýchlosť tohoto rastu. Poslednou metódou sú algoritmy Reversible Jump a jump-diffusion sampling, čo sú MCMC metódy pre reťazce s premenlivou dimenziou stavového priestoru. U týchto metód je však ťažké diagnostikovať konzistenciu a sú náročné na dobrú implementáciu.

SÚČASNOSŤ

Nové svetlo do oblasti odhadovania počtu komponentov zmesi vniesol Keribin (2000), ktorému sa pomocou špeciálnej reparametrizácie podarilo ukázať, že **odhad** počtu komponentov **metódou penalizovanej vierohodnosti** (pre určitý tvar penalizácie a za určitých predpokladov na rodinu rozdelení) **je konzistentný**. Podľa mojich znalostí sú podmienky overené len pre jednorozmerné normálne[31] a poissonove zmesi. Keribinovej penalizácii vyhovuje kritérium BIC. Použitie tvaru $(d/2)(\log n)^{\alpha}$, kde $\alpha > 0$ a d počet parametrov modelu, však umožňuje väčšiu flexibilitu. V časti venovanej simuláciám ukážem, že v prípade vzorky normálnych zmesí z článku Marron, Wand (1992) pre výbery s rozsahom medzi 100 až 1000 dosahuje Keribinova penalizáciu s $\alpha = 1/2$ veľmi dobré výsledky.

V súvislosti s odhadovaním zmesí si musíme ešte uvedomiť, že v prípade veľkého rozsahu dát a väčších dimenzií sme odkázaný len na metódu MLE resp. MPLE, ktorú jedinú vieme vďaka EM algoritmu a množstvu jeho modifikácií efektívne implementovať aj v týchto náročnejších situáciách. **Existencia EM algoritmu** je z tohoto pohľadu **veľmi dôležitá**. Význam algoritmu ešte v poslednej dobe narástol výsledkom o konzistencii odhadu zložitosti metódou MPLE a stále narastá so zavádzaním jeho nových modifikácií (algoritmy SMEM a SSMEM).

V závere druhej kapitoly som poukázal na zaujímavé vylepšenie EM algoritmu. Algoritmus SMEM zavádza spájanie a rozdeľovanie zhlukov za účelom odskakovania od zlých lokálnych extrémov a následne algoritmus SSMEM, ktorý spájanie a rozdeľovanie zhlukov v rámci EM algoritmu využíva aj ku odhadu počtu komponentov. Vývoj je tu podobný ako v prípade Bayesovských metód. Tam najskôr vznikli metódy MCMC a nasledoval vznik algoritmu Reversible Jump, ktorý je podobne ako algoritmy SMEM a SSMEM založený na spájaní a rozdeľovaní zhlukov.

PRETRVÁVAJÚCE NEDOSTATKY

V súvislosti s odhadom zmesí nesmieme zabudnúť na otázku **identifikovateľnosti** parametra modelu, ktorá patrí stále ešte k neuzavretým problémom. Pri používaní **EM algoritmu** musíme brať ohľad na jeho **základné nedostatky**, ktorými sú:

- *Pomalá konvergencia* - je známe, že EM algoritmus nedisponuje veľkou rýchlosťou.

[31] Rozptyl komponentov musí byť zdola obmedzený nejakou kladnou konštantou, čo je ale štandardná a nie nejak významne obmedzujúca podmienka.

Pri nevhodnej inicialiácii môže preto trvať veľmi dlho kým dostaneme rozumný odhad. Pri náhodnej inicializácii v záujme zvýšenia rýchlosti preto zväčša obmedzujeme maximálny počet iterácií a radšej povolíme väčší počet inicializácii.

- *Lokálne extrémy* - algoritmus častokrát skončí v lokálnom maxime, z ktorého sa mu už nepodarí vyskočiť, pričom niektoré lokálne maximá dávajú značne zlé odhady. Algoritmus preto opakujeme viac krát pre rôzne inicializácie a z dosiahnutých výsledkov potom vyberáme záverečný odhad.

- *Numerická nestabilita na okrajoch parametrického priestoru* - kvôli tejto nestabilite je problematické odhadovať komponenty s veľmi malým rozptylom, resp. veľmi malým zastúpením. Častokrát sa za týmto účelom stanovujú dolné medze na proporcie a rozptyly. Vo viacrozmernom (normálnom) prípade sa počas EM iterácií stáva, že nejaká variančná matica sa stane takmer singulárnou, čo má za následok zrútenie algoritmu[32]. Často sa preto používa spektrálny rozklad variančnej matice, ktorý umožňuje zníženie počtu parametrov modelu.

3.2 R - knižnice pre prácu s modelom zmesí

V tejto časti predstavím, aké funkcie a objekty ponúka pre prácu so zmesami prostredie R. V základnom balíku sa podľa mojich informácií nenachádza žiadna funkcia na analýzu zmesí[33], v archíve poskytovaných knižníc (CRAN) však môžeme nájsť tieto užitočné balíky:

- **norlmix** - umožňuje skonštruovať objekt jednorozmernej normálnej zmesi. Obsahuje funkciu pre hustotu, distribučnú funkciu, graf a generovanie náhodného výberu z tohoto objektu a okrem toho obsahuje objekty normálnych zmesí z článku Marron, Wand (1992).

- **mclust** - balík pre aplikáciu viacrozmerných normálnych zmesí z pohľadu zhlukovania (model based clustering). Sú v ňom implementované hierarchické zhlukovacie algoritmy (HC) pre normálne rozdelenie (napr. známe Wardovo kritérium, kritérium Scotta a Symonsa,viď Banfield, Raftery (1993)), a tiež EM algoritmus pre viacrozmerné normálne zmesi, ktorý vie pracovať aj s modelom rozšíreným o odľahlé pozorovania (1.50). Používa spektrálny rozklad variančných matíc (1.45) a na základe BIC kritéria vo viacrozmernom prípade vyberá najvhodnejší model z tabuľky 3.1. V prípade, že parametrizácia variančnej matice umožňuje aplikovať metódu HC, je EM algoritmus inicializovaný výsledkom tejto HC metódy. V jednorozmernom prípade balík rozlišuje tieto prípady: E - komponenty s rovnakým rozptylom a V - komponenty s rôznym rozptylom.

  Balík navyše umožňuje aj použitie diskriminančnej analýzy založenej na EM algoritme. Idea tejto varianty diskriminančnej analýzy spočíva v odhadnutí hustoty klasifikovaných trénovacích dát (M-krok), čo je model so zmesou hustôt, pomocou

[32] V článku Naonori Ueda (2000) sa v tejto súvislosti používa tzv. Bayesovská regularizácia.

[33] Funkcie pre zhlukovanie ako napr. metóda *k-means* sú v zákaldnom balíku prítomné.

Názov	Model	HC	EM	Rozdelenie	Veľkosť	Tvar	Orientácia
EII	$\lambda\mathbf{I}$	x	x	sférické	rovnaká	rovnaký	nie je
VII	$\lambda_k\mathbf{I}$	x	x	sférické	rôzna	rovnaký	nie je
EEI	$\lambda\mathbf{\Lambda}$		x	diagonálne	rovnaká	rovnaký	osi súr. sys.
VEI	$\lambda_k\mathbf{\Lambda}$		x	diagonálne	rôzna	rovnaký	osi súr. sys.
EVI	$\lambda\mathbf{\Lambda}_k$		x	diagonálne	rovnaká	rôzny	osi súr. sys.
VVI	$\lambda_k\mathbf{\Lambda}_k$		x	diagonálne	rôzna	rôzny	osi súr. sys.
EEE	$\lambda\mathbf{A}\mathbf{\Lambda}\mathbf{A}^T$	x	x	eliptické	rovnaká	rovnaký	rovnaká
VVV	$\lambda_k\mathbf{A}_k\mathbf{\Lambda}_k\mathbf{A}_k^T$	x	x	eliptické	rôzna	rôzny	rôzna
EEV	$\lambda\mathbf{A}_k\mathbf{\Lambda}\mathbf{A}_k^T$		x	eliptické	rovnaká	rovnaký	rôzna
VEV	$\lambda_k\mathbf{A}_k\mathbf{\Lambda}\mathbf{A}_k^T$		x	eliptické	rôzna	rovnaký	rôzna

Tabuľka 3.1: Parametrizácia variančnej matice dostupná v knižnici `mclust` pre hierarchické zhlukovanie HC a EM algoritmus pre viacrozmerné dáta ('x' znamená dostupnosť).

ktorej sa potom (E-krok) zkonštrujú podmienené pravdepodobnosti príslušnosti a následne aj samotné príslušnosti testovacích dát do jednotlivých zhlukov.

V prípade veľkého rozsahu dát môžeme aplikovať EM algoritmus pre nejaký výber pôvodných dát a použiť techniku diskriminačnej analýzy pre klasifikáciu zvyšných dát. Pri veľkej dimenzii pozorovaní v prípade modelov EEV, VEV a VVV, pri ktorých je počet parametrov maximálny - pre každý zhluk rádu $O(p^2)$, kde p je dimenzia dát - majú však výsledky malú kvalitu. Preto sa zväčša obmedzujeme na modely s menším počtom parametrov. Balík ponúka aj niekoľko grafických procedúr na zobrazovanie viacrozmerných dát.

- **bayesmix** - implementácia MCMC metód pre odhad jednorozmernej normálnej zmesi so známym počtom komponentov, ktorá pre realizáciu MCMC metód využíva program JAGS[34].

- **flexmix** - knižnica pre aplikáciu konečných zmesí regresných modelov

$$h(\mathbf{y}|\mathbf{x},\theta) = \sum_{i=1}^{g} \pi_i f(\mathbf{y}|\mathbf{x},\theta_k),$$

kde f je buď jednorozmerné normálne rozdelenie so strednou hodnotou $\boldsymbol{\beta}_k^T\mathbf{x}$ a rozpytlom σ_k^2 ($\boldsymbol{\theta}_k = (\boldsymbol{\beta}_k^T, \sigma_k^2)$) - vtedy hovoríme o zmesi (štandardných) lineárnych regresných modelov[35], alebo je f členom všeobecnej exponenciálnej rodiny - vtedy hovoríme o zmesi zobecnených lineárnych modelov. Regresori a regresant môžu byť aj viacrozmerné. Na odhad parametrov sa používa EM algoritmus.

- **fpc** - je knižnica od pána Henniga, ktorá sa venuje klasifikácii a zhlukovaniu v prípade viacrozmerných a rozsiahlych dát. Hennig nedávno vymyslel zaujímavú

[34] *Just Another Gibbs Sampler* - tento program sa dá priamo stiahnuť s knižnicou bayesmix a nevyžaduje žiadnu inštaláciu.

[35] V angličtine sa často používa aj termín *latent class regression* alebo *cluster-wise regression*

metódu na identifikáciu odľahlých pozorovaní v súvislosti so zhlukovaním (Hennig, 1998) a nazval ju *fixed point cluster analysis* (FPCA). Neskôr túto metódu použil k analýze zhlukovania v rozsiahlych dátach (Hennig, Christlieb, 2002). Knižnica je venovaná technike FPCA, obsahuje rôzne techniky redukcie dimenzie pre zhlukovanie a nájdeme v nej aj funkcie venované zmesiam lineárnych regresných modelov.

- **mmlcr**- v prvej kapitole som trochu predstavil model zmesi pre dáta, ktoré obsahujú spojité aj kategorické zložky, tzv. *mixed-mode data.* Knižnica `mmlcr` sa venuje zmesi regresných modelov práve pre tento typ dát[36].

- **mixreg** - knižnica pre zmesi regresných modelov s jedným regresorom. Zaujímavosťou je, že obsahuje funkciu pre test počtu komponentov metódou pomeru vierohodnotí, pričom štatistiku LRTS aproximuje buď parametrickým alebo semiparametrickým bootstrapom. Ako ale vieme, tento spôsob si vyžaduje veľkú dávku trpezlivosti.

- **depmix** - knižnica venovaná zmesiam pre závislé pozorovania (časové rady) - latent Markov models. Dáta môžu mať zmiešanú štruktúru (spojité aj kategoriálne zložky).

- **moc** - balík venovaný všeobecným nelineárnym zmesiam kriviek[37], pestrá paleta modelov zmesí pre viacrozmerné pozorovania s distribúciou a profilom definovanými samotným užívateľom.

- **varmixt** - rozsiahla knižnica zaoberajúca sa modelom zmesí v súvislosti s analýzou dát genovej expresie.

- **vabayelmix** - knižnica, ktorá implementuje variačnú procedúru Bayesovského odhadu pre viacrozmerné normálne zmesi (súčasná verzia vyžaduje, aby variančné matice zmesi boli diagonálne). Implementovaný algoritmus (*iterative ensemble leraning algorithm*) umožňuje aj odhad počtu komponentov.

- **mixdist**[38] - knižnica, ktorá implementuje modely normálnych, log-normálnych, gama, Weibulových, binomických, negetívne binomických a poissonových zmesí pre dáta so skupinovou štruktúrou (dáta sú zoskupené do intervalov, nie sú teda zaznamenané individuálne pozorovania, ale len interval, do ktorého pozorovanie padlo). Pre odhad parametrov modelu sa používa kombinácia EM algoritmu a metódy Newtonovho typu. Je možné tiež nastaviť rozličné obmedzenia pre odhadované parametre, napr. aby stredné hodnoty susedných komponentov boli od

[36] Anglický termín pre takýto model je *mixed-mode latent class regression*, tiež známy ako *mixed-mode mixture regression model* alebo *mixed-mode mixture model regression.*

[37] Tomuto modelu som sa podrobne nevenoval. Knižnicu uvádzam len pre úplnosť.

[38] Pôvodne sa táto knižnica nazývala `Rmix` a ako jediná z uvedených knižníc sa nenachádza v archíve CRAN. Spolu s dokumentáciou je prístupná na `http://www.math.mcmaster.ca/peter/mix/mix.html`. Táto knižnica je vlastne akousi variantou komerčného programu `MIX` od Petra MacDonalda pre prostredie `R`.

seba rovnako vzdialené ($(\mu_2 - \mu_1) = (\mu_3 - \mu_2) = \ldots = (\mu_g - \mu_{g-1})$). Neobsahuje žiadnu funkciu, či metódu pre odhad počtu komponentov.

3.3 Poznámky k vlastným algoritmom

V tejto časti stručne predstavím vlastné algoritmy implemetované v prostredí `R` priložené v zadnej časti práce na CD médiu. Jedná sa o šesť skriptov: `EM.R`, `LRTS.R`, `mindistNM.R`, `mindistGA.R`, `multKer.R` a `Roeder.R`. Skripty sú pomerne podrobne komentované, takže veľkú časť je možné vyčítať z ich samotného kódu. Pre ich bezproblémovú funkčnosť je nutné v pracovnom adresári prostredia `R` vytvoriť adresár `mix` a do tohoto tieto skripty nahrať[39]. Vzhľadom k tomu, že občas sa budem zaoberať aj rýchlosťou výpočtu, je na mieste uviesť, že aplikácie boli prevedené na počítači s procesorom AMD Turion 64 Mobile, 1.79 GHz a operačným systémom MS Windows XP (2002). Na priloženom CD médiu sa okrem programových kódov nachádzajú aj všetky testované a analyzované dáta, ktoré boli pri analýzach použité.

3.3.1 EM

Skript `EM.R` obsahuje implementáciu MPLE pre jednorozmerný prípad zmesi normálnych, exponencionálnych alebo Poissonových rozdelení. Vierohodnosť je maximalizovaná pomocou EM algoritmu. Ako penalizáciu je možné zvoliť jedno z nasledujúcich kritérií: AIC, BIC, CLC, NEC, ICL-BIC, MIR[40], KER pre rôzne α a EIC (je možné zvoliť aj všetky kritériá súčasne pre účely porovnania s výnimkou penalizácie EIC, ktorá môže byť použitá len samostatne vzhľadom k svojej časovej náročnosti).

Hlavnou funkciou skriptu je funkcia `PEME.ic`, ktorej špecifikácia sa nachádza v dodatku B.1. Okrem tejto funkcie je v tomto skripte implementovaný aj prístup cross-validácie (kritérium CVIC), funkcia `PEME.cvic`. Táto metóda je však časovo veľmi náročná, podobne ako kritérium `EIC` a aproximácia štatistiky LRTS technikou bootstrap. Posledná uvedená metóda je implementovaná v rovnomennom skripte `LRTS.R`.

3.3.2 multKer

V tomto skripte som implementoval odhad zmesi pomocou vierohodnosti penalizovanej podľa Keribina aj pre prípad viacrozmerných normálnych rozdelení. Ide vlastne len o aplikáciu funkcie `Mclust` z predstavenej knižnice `mclust`. Skript obsahuje len jednu funkciu `MclustKer` so vstupnými parametrami `data, minG, maxG, alpha, plt`, ktorých význam je zrejmý z ich názvu. Dáta musia byť formátu `data.frame` (pozorovania sú v riadkoch, príslušné dimenzie jednotlivých pozorovaní zase v stĺpcoch), `alpha` je vektor rôznych α pre kritérium KER (defaultne $\alpha = 1/2,1,2,3$) a `plt` je parameter pre zobrazenie grafického výstupu. Pre každú hodnotu α sa potom vo výstupe nachádza zvolený

[39] Štandardne teda skripty uložíme do adresára `"C:\Program Files\R\R-2.2.1\mix\"`.

[40] Varianta založená na konvergencii EM algoritmu (McLachlan, Peel, 2000, str. 206)

počet komponentov a najvhodnejší model z `EII, VII, EEI, VVI, EEE, VVV`. Okrem týchto vo výstupe nájdeme aj odhad parametrov pre model vybratý na základe BIC kritéria. V prípade, že ide o dvojrozmerné dáta a vstupný parameter `plt` je nastavený na hodnotu `TRUE`, je vykreslená aj hustota a vrstevnicový graf pre tento model.

3.3.3 mindistNM

Skripty `mindistNM.R` a `mindistGA.R` sú venované odhadom založeným na minimálnej vzdialenosti. Pokúsil som sa implementovať Hennov algoritmus (Henna, 1985) založený na minimalizácii Cramér-von-Misesovej (CvM) vzdialenosti a algoritmus Chena (Chen, Kalbfleisch, 1996) pre suprémovú (KS), Kantorowichovu (označenie KANT, integrál absolútnej hodnoty rozdielu distribučných funkcií) a Cramér-von-Misesovu vzdialenosť.

Hlavné funkcie sú `MPDE` a `MPDE.ini` a umožňujú voľbu modelu metódou minimálnej vzdialenosti podľa Hennu aj Chena pre (jednorozmerné) normálne zmesi, pričom na minimalizáciu je použitý Nelder-Meadov simplexov algoritmus (NMSA). Špecifikácia týchto funkcií je umiestnená v dodatku B.2.

APLIKÁCIA

Pre ilustráciu som funkcie `MPDE` a `MPDE.ini` aplikoval na zmes troch dobre oddelených normálnych rozdelení $N(-3,1)$, $N(5,1)$, $N(20,16)$, pričom 50 pozorovaní bolo vygenerovaných z prvého zhluku, 30 z druhého a 50 z tretieho (dáta nazveme `data1`). Výstup funkcie `MPDE.ini` pre rôzne typy vzdialenosti (`$p, $mu, $sd` sú výsledné odhady pre proporcie, stredné hodnoty a rozptyly zmesi):

```
>MPDE.ini(data1,k=3,algorithm="Chen",dist="Kant",ini="sample")

User, System and Total elapsed time(s): 1206.07 0.83 1923.86
$p   0.3845334 0.2236523 0.3918143
$mu -3.109666  5.173643 18.947557
$sd  1.088789  1.052461  4.236215

> MPDE.ini(data1,k=3,algorithm="Chen",dist="KS",ini="sample")

User, System and Total elapsed time(s): 1879.33 1.58 2816.27
$p   0.2773973 0.3519950 0.3706078
$mu -2.277328  7.245524 15.434113
$sd  1.668226 12.019747 11.281520

> MPDE.ini(data1,k=3,algorithm="Chen",dist="CvM",delta=0.0001,ini="sample")

User, System and Total elapsed time(s): 478.05 2.96 516.93
$p  0.2254601 0.3973631 0.3771768
$mu 5.046499 18.553950 -3.175832
$sd 1.128971  4.200309  1.049458

> MPDE.ini(data1,k=3,algorithm="Henna",dist="CvM",ini="sample")

User, System and Total elapsed time(s): 524.12 0.32 532.76
$p   0.3993386 0.2202354 0.3804260
```

```
$mu 18.523681  5.057628 -3.161928
$sd  4.225705  1.093522  1.063823
```

Z výstupu vidíme, že o poznanie lepšie a rýchlejšie výsledky v porovnaní so suprémovou vzdialenosťou dosahujeme pri použití KANT a CvM vzdialenosti, pričom najrýchlejšie je to v prípade CvM vzdialenosti (tu bolo potrebné zmenšiť parameter delta). Použitie suprémovej vzdialenosti sa teda z praktického hľadiska ukazuje ako neefektívne.

Výstup aplikácie funkcie `MPDE` aplikovanej na datový súbor `data1` (pripustil som len modely s dvoma, troma a štyrmi komponentami):

```
>MPDE(data1,k.min=2,k.max=4,dist="CvM",delta=0.0001,ini="emklas")

[1] 2
User, System and Total elapsed time: 160.02 0.22 165.97
$p   0.4437681 0.5562319
$mu -2.828887 15.538720
$sd  1.402869  7.524375

[1] 3
User, System and Total elapsed time: 299.75 0.04 307.28
$p   0.4016213 0.3825158 0.2158629
$mu 18.488686 -3.152926  5.057283
$sd  4.255078  1.073128  1.063894

[1] 4
User, System and Total elapsed time: 3358.37 3.38 3679.41
$p  0.24028276 0.37770530 0.31247851 0.06953343
$mu 5.167429  -3.173591  19.597961  16.454012
$sd 1.2304869  1.0517808  4.2737967  0.8995325

[1] "Best theta:"
$p   0.4016213 0.3825158 0.2158629
$mu 18.488686 -3.152926  5.057283
$sd  4.255078  1.073128  1.063894
```

Všimnime si, že použitie inicializácie `emklas` znížilo čas potrebný na nájdenie minima (v prípade troch komponentov). V prípade štyroch komponentov sme si ale na výsledok museli počkať takmer hodinu. Nájdené riešenie je napokon správne, ale čas potrebný na jeho nájdenie je príliš dlhý, vzhľadom k tomu, že ide len o veľmi jednoduchú zmes odhadovanú na malom počte pozorovaní[41].

Pokúsim sa teraz aplikovať funkcie na prípad náročnejšej zmesi. Uvažujme zmes, ktorú Marron a Wand vo svojom článku z roku 1992 označili názvom *Trimodal*. Zmes označíme ako MW9. Jej parametre sú (použijeme knižnicu `norlmix`):

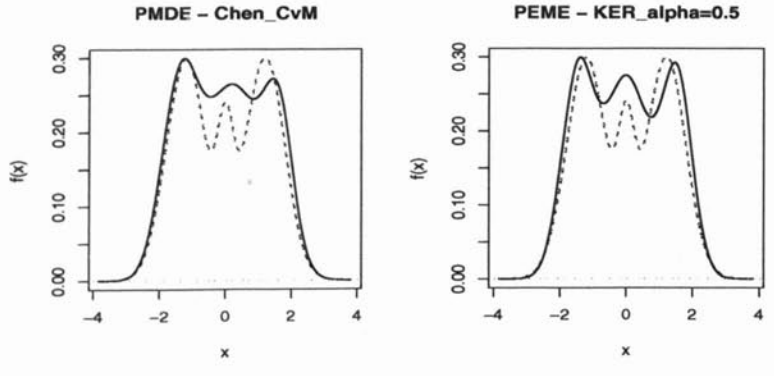
```
> MW.nm9
'Normal Mixture' object          ''#9 Trimodal''
      mu   sig2    w
[1,] -1.2 0.3600 0.45
[2,]  1.2 0.3600 0.45
[3,]  0.0 0.0625 0.10
```

[41] Aplikácia metódy MPLE (kritérium BIC) v tomto prípade nevyžaduje žiadne čakanie

Aplikujeme metódu Chena pre CvM vzdialenosť na výber z tejto zmesi o rozsahu 200 (opäť som pripustil len 2,3 a 4 komponentov), data označíme ako `data2`:

```
>data2=rnorMix(200,MW.nm9)

>MPDE(data2,k.min=2,k.max=4,dist="CvM",delta=0.0001,ini="emklas")

[1] 2
User, System and Total elapsed time: 253.5 0.14 259.79
$p   0.5590267 0.4409733
$mu -1.015504  1.139938
$sd  0.8523159 0.7354323

[1] 3
User, System and Total elapsed time: 4237.95 2.72 4344.66
$p   0.2959034  0.5211662  0.1829304
$mu -1.4895660  0.1704048  1.5934087
$sd  0.5539964  0.9089898  0.4527268

[1] 4
User, System and Total elapsed time: 4228.33 2.06 4287.95
$p  0.1613978    0.3916389    0.1227522    0.3242111
$mu 1.185935525 -0.006294422  1.926305444 -1.495883936
$sd 0.2409526    0.6303262    0.2224622    0.5197757

[1] "Best theta:"
$p   0.2959034  0.5211662  0.1829304
$mu -1.4895660  0.1704048  1.5934087
$sd  0.5539964  0.9089898  0.4527268
```

Chenovou metódou sa nám podarilo odhadnúť správny počet komponentov, ale čas potrebný pre tento odhad sa v tomto prípade blíži až trom hodinám. Na obrázku 3.1 vľavo môžeme ešte vidieť porovnanie hustoty skutočnej zmesi a hustoty odhadnutej zmesi. Pre porovnanie si ešte uveďme výsledok metódy MPLE pre rôzne informačné



Obrázok 3.1: Porovnanie hustoty zmesi MW9 (červená prerušovaná čiara) s hustotou odhadu zmesi získaného metódou MPDE podľa Chena (ľavý obrázok) a s hustotou odhadu zmesi metódou PEME pomocou Keribinovho kritéria s parametrom $\alpha = 0.5$ (pravý obrázok).

kritéria (z výstupu uvádzam to najpodstatnejšie):

```
> PEME.ic(data2,criterion="all")

User, System and Total elapsed time: 93.22 0 96.28

IC for different number of components: 2 3 4 5 6 7 8 9 10
AIC:   654  651  651  654  660  666  672  678  684
KER1:  655  653  654  659  665  672  679  686  692
BIC:   670  677  687  701  716  732  748  763  779
KER3:  784  860  938 1019 1103 1187 1271 1356 1440
KER4: 1387 1825 2265 2709 3155 3600 4047 4493 4939
CLC:   716  772  759  775  905  889  926  982  998
ICL:   743  814  817  849  995  995 1048 1120 1151
NEC:  2.85 3.83 3.16 5.39 6.63 7.13 7.72 7.93 8.72
MIR:  1.09 0.42 0.42 0.34 0.53 0.64 0.46 0.61 0.63

Number of components for best models:
AIC :    4 KER1:    3
BIC :    2 KER3:    2
KER4:    2 CLC :    2
NEC :    2 ICL :    2
MIR :    5

ker1
$p   0.4276790   0.2796211   0.2926999
$mu -0.04316884  1.49945935 -1.53754127
$sd  0.7039327   0.4718796   0.4897725
```

Vidíme, že správny počet komponentov sa v tomto prípade podarilo odhaliť len v prípade Keribinovho kritéria s parametrom $\alpha = 0.5$ (pričom voľba medzi dvoma, troma a štyrmi komponentami je veľmi tesná). Na obrázku 3.1 vpravo môžeme opäť vidieť porovnanie skutočnej hustoty zmesi MW9 s hustotou danou týmto odhadom. Napokon si ešte všimnime, že k výsledku sme sa v tomto prípade dopracovali približne za jeden a pol minúty, čo je neporovnateľný rozdiel s tromi hodinami potrebnými v prípade metódy MPDE.

3.3.4 mindistGA

Za účelom zlepšenia implementácie metódy MPDE som vyvinul *genetický algoritmus* (GA), ktorým som nahradil algoritmus NMSA. Celý prístup je pre prehľadnosť implementovaný v samostatnom skripte `mindistGA.R`. Popis hlavnej funkcie `MPDE.ga` aj stručné popísanie genetického algoritmu sa nachádzajú v dodatku B.3.

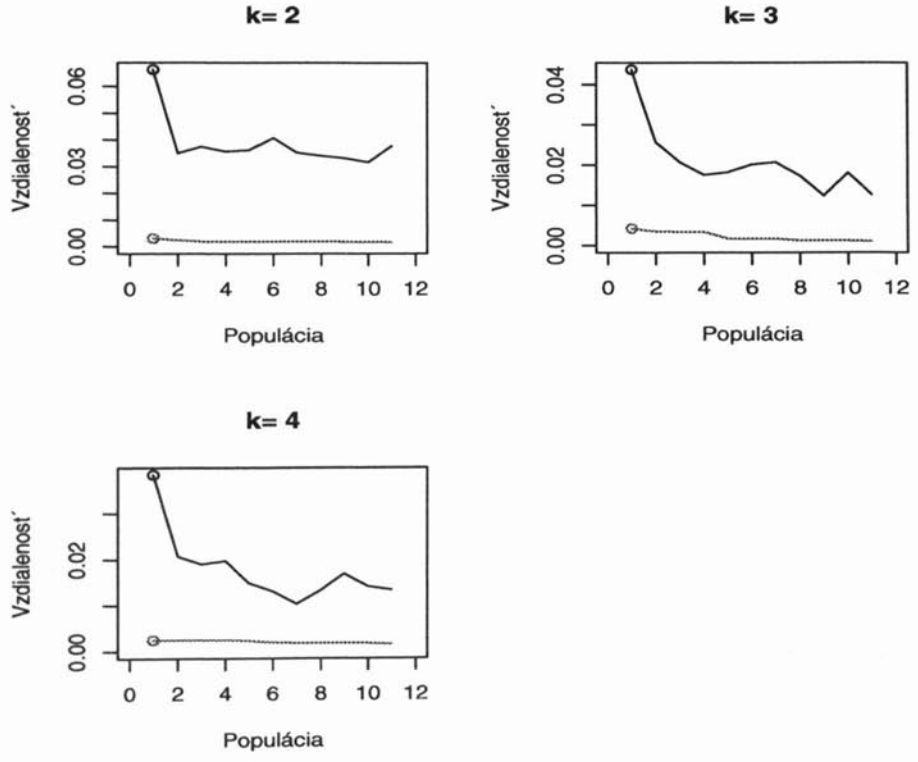
APLIKÁCIA

Výhodou implementovaného GA je, že ho môžeme pomerne dobre regulovať, sledovať priebežné výsledky a ovplyvňovať aj čas výpočtu. Nevýhodou je samozrejme istá jeho vágnosť, keďže ide o akési na náhode založené prehľadávanie priestoru. Vzhľadom k tomu, že používame elitárstvo (prenášanie najlepších jedincov populácie do novej populácie), máme zaručené, že najlepší fitness (minimálna vzdialenosť) aktuálnej populácie bude vždy aspoň taký dobrý ako najlepší fitness populácie predchádzajúcej. Dôležité však ešte je, aby klesal aj priemerný fitness populácie, alebo prinajmenšom výrazne

nerástol, poprípade nerobil veľké skoky nahor. Algoritmus som aplikoval na datový súbor `data1` (počet iterácií 10):

```
> MPDE.ga(data1,K.min=2,K.max=4,dist="CvM",delta=0.0001,init="rand")

Penalized distance(k =2:4): 0.001685570 0.001066024 0.001518626
$p  0.2414473 0.3598802 0.3986724
$mu 4.786963 19.247450 -3.252283
$sd 1.709765  2.893708  0.880381

User, System and Total elapsed time(s): 319.22 2.34 325.4
```

Počet komponentov sa mi podarilo odhadnúť správne, odhad parametrov nie je síce excelentný, možno ho však považovať za dostačujúci. V porovnaní s NMSA sa však výrazne skrátil čas potrebný na výpočet (predtým hodina, teraz približne 5 minút). Na obrázku 3.2 je vývoj fitnessu populácie, ktorý vyzerá celkom rozumne. Na získanie



Obrázok 3.2: Genetický algoritmus na hľadanie odhadu zmesi metódou MPDE -vývoj fitnessu populácie pre odhad zmesi (počet komponentov - k) v prípade datového súboru `data1` (čierna čiara - priemerný fitness populácie, červená čiara - najlepší fitness populácie).

vierohodnejšieho výsledku je potrebné použiť viac iterácií. Pri použití 30 iterácií som dostal tento výsledok:

```
> MPDE.ga(data1,K.min=2,K.max=5,dist="CvM",delta=0.0001,init="rand")

Penalized distance(k =2:5): 0.001122622 0.0006150428 0.001058659 0.001248547
$p   0.4057340 0.1884841 0.4057819
$mu 17.787096  5.546846 -3.079713
$sd  5.525674  1.087241  0.960204

User, System and Total elapsed time(s): 1210.84 5.35 1443.34
```

Čas potrebný na výpočet síce narástol (pripustil som aj zmes s piatimi komponentami), ale stále je to výrazne lepšie ako v prípade NMSA. Napokon sa ešte pozrime, ako si GA poradil s náročnejšou zmesou reprezentovanou výberom `data2` (30 iterácií):

```
> MPDE.ga(data,K.min=2,K.max=4,dist="CvM",delta=0.0001,init="rand")

Penalized distance(k =2:4): 0.0001945033 0.0002940073 0.0004719044
$p  0.5065705  0.4934295
$mu 1.004719  -1.108383
$sd 0.8779308  0.7821236

User, System and Total elapsed time(s): 1492.86 2.48 1540.95
```

Skutočný počet komponentov sa po 30-tich iteráciách nepodarilo odhaliť (bolo by zrejmé potrebné iterovať ďalej alebo sa pokúsiť zmeniť nastavenie parametrov pre GA), odhad pre dva komponenty je však slušný.

Výhodou použitia GA je, že umožňuje vstupovať do prehľadávacieho procesu, poprípade sa zamerať viac na lokálne prehľadávanie, ktoré môže byť vhodné ak chceme vylepšiť nejaké iné dostupné riešenie, alebo využiť nejakú apriórnu znalosť. Taktiež môžeme ohraničiť čas prehľadávania, čo je tiež zaujímavý aspekt. V praxi sa totiž často stáva, že do určitého času potrebujeme nájsť aspoň nejaké riešenie. Algoritmus ponúka väčšiu voľnosť a dáva nejaké riešenie aj v náročnejších úlohách, kde nemáme k dispozícii iné riešenie. To bol aj dôvod prečo som sa pri implementácii metódy MPDE vybral týmto smerom.

3.3.5 Roeder

Na záver som ešte implementoval grafickú techniku MDP (Roeder, 1994) pre testovanie počtu komponentov v prípade jednorozmernej normálnej zmesi predstavenú v závere predchádzajúcej kapitoly. Skript som nazval `Roeder.R` a na jeho zhotovenie mi dobre poslúžil spomenutý článok, ktorý sa venuje aj algoritmizácii zavedenej techniky.

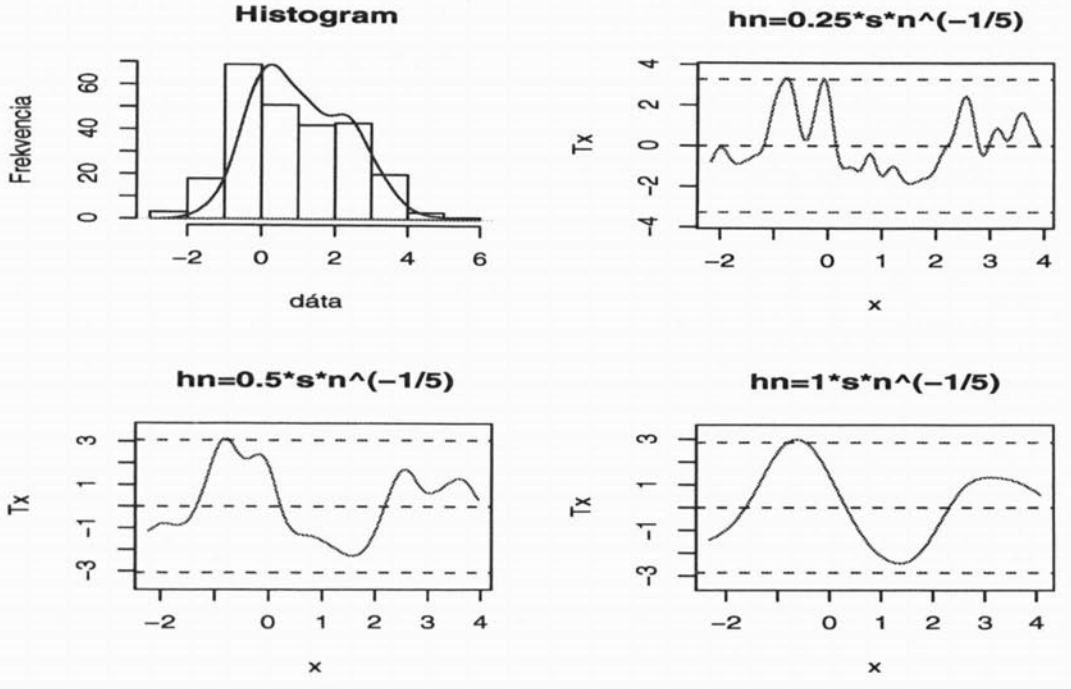
Hlavná funkcia `MDP` je venovaná testovaniu homogenity. Hlavné parametere sú `data` a vektor koeficientov `a.plt`, ktorý určuje šírku vyhladzovacieho parametra $h_n = (\texttt{a.plt}) * s * n^{-1/5}$ (s je odmocnina z rozptylu dát a n je rozsah dát). Vo výstupe sa nachádza detekčný graf MDP pre rôzne vyhladzovacie parametre dané vektorom `a.plt` (defaultne vektor troch hodnôt: 0.25, 0.5, 1).

Funkcia `MDPk` je zase venovaná testovaniu hypotézy $H_0 : g = k$ proti $H_1 : g = k + 1$ pre $k > 1$. Hlavné parametre sú `data, k, a.plt`. V tomto prípade ide však len o aproximáciu originálnej techniky, ktorá využíva odhad zmesi metódou momentov. Namiesto metódy momentov som použil maximálne vierohodný odhad.

APLIKÁCIA

Pre ilustráciu som použil funkciu `MDP` na zmes $0.67N(0,1)+0.33N(2.5,1)$ reprezentovanú výberom o rozsahu 250 (tieto dáta označíme ako `data3`). Ide o prípad unimodálnej

zmesi[42]. Detekcia pomocou funkcie MDP (obr. 3.3) nám odhalí, že dáta skutočne pochádzajú zo zmesi. Po vyhladení lokálnych perturbácií sa totiž prejaví bimodálny signál - detekčný graf má dva modusy a výrazné dno, a hoci oscilácie mierne prekračujú konfidečnú hranicu, výsledok možno ešte považovať za signifikantný.



Obrázok 3.3: Výstup funkcie `MDP(data3)`. Prvý graf zobrazuje histogram preložený hustotou a ostatné grafy znázorňujú MDP pre tri rôzne vyhladzovacie parametre. Horná a dolná prerušovaná čiara indikujú 90% konfidenčný interval pre maximálnu osciláciu.

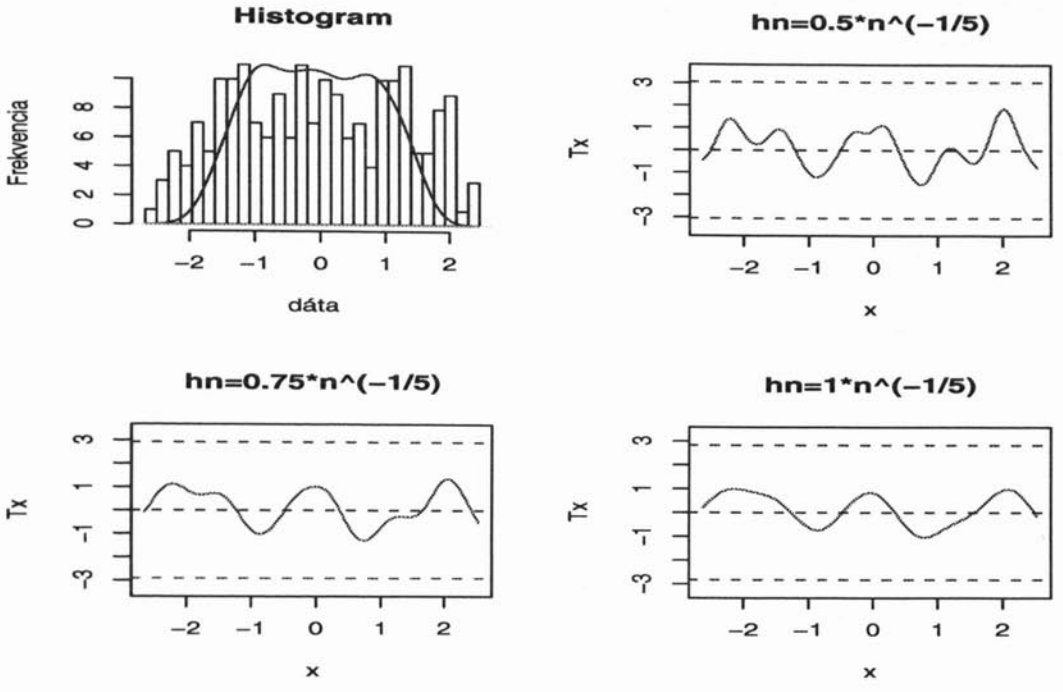
Funkciu `MDPk` som aplikoval na datový súbor `data2`. Výstup testovania hypotézy $H_0 : g = 2$ proti $H_1 : g = 3$ je na obrázku 3.4. Po vyhladení lokálnych oscilácií vykazuje detekčný graf tri modusy s dvomi dnami, pričom mení znamienko 6-krát v poradí $(-, +, -, +, -, +, -)$, čo znamená, že v tejto normálnej zmesi sú podľa tejto detekcie prítomné tri komponenty.

3.4 Simulácie - porovnanie informačných kritérií

V tejto časti predstavím výsledky simulácií pre odhad počtu komponentov v prípade rôznych informačných kritérií, ktoré som získal aplikáciou funkcie `PEME.ic`.

Testovanie som uskutočnil na 12-tich typoch normálnej zmesi z článku Marron, Wand (1992), ktorých definície sa nachádzajú v prehľadnej tabuľke v dodatku A a grafy hustôt boli predstavené v prvej kapitole. Zmesi budem označovať písmenami MW a číslom, ktoré odpovedá číslovaniu v dodatku A. Vzhľadom k tomu, že zmesi MW11 a MW13

[42]Táto zmes pripomína zošikmené rozdelenie. Odlíšiť zošikmené rozdelenie od normálnej zmesi je však náročná úloha, o ktorej som sa už zmienil v prvej kapitole.



Obrázok 3.4: Výstup funkcie `MDPk(data2,k=2,verbose=F)`. Prvý graf zobrazuje histogram preložený hustotou, ostatné grafy znázorňujú MDP pre tri rôzne vyhladzovacie parametre. Horná a dolná prerušovaná čiara indikujú 90% konfidenčný interval pre maximálnu osciláciu.

sú si podobné, rovnako ako dvojica zmesí MW14, MW15, z testovania som vylúčil zmes MW13 a zmes MW15. Počet komponentov v uvažovanej vzorke zmesí sa pohybuje medzi 2 až 9. Treba podotknúť, že v niektorých prípadoch ide o veľmi komplikované zmesi, a to predovšetkým zmes MW11, MW3 a MW14 (obsahujú komponenty s veľmi malým rozptylom poprípade aj malým zastúpením). Z každej zmesi som 50-krát realizoval náhodný výber o rozsahu $n = 100$, 200, 500 a 1000, v poslednom prípade som náhodny výber opakoval len 20-krát, a na každom výbere som aplikáciou funkcie `PEME.ic` odhadol počet komponentov v prípade viacerých informačných kritérií.

Pri odhade počtu komponentov som uvažoval hornú hranicu pre počet komponentov 10 a použil som tieto kritéria: AIC, Keribinovu penalizáciu pre 4 rôzne $\alpha = 1/2$, 1, 2, 3 (označenie KER1, BIC, KER3, KER4), CLC, ICL a MIR[43]. Výsledky simulácií sú uvedené v tabuľkách 3.2, 3.3, 3.4 a 3.5, ktoré odpovedajú rôznym rozsahom náhodných výberov. Pre kažé kritérium a každú zmes je v tabuľke zaznamenané, koľkokrát bol zvolený konkrétny počet komponentov. Namiesto nuly je pre prehľadnosť použitá bodka. Správny počet komponentov je v prípade každej zmesi označený hviezdičkou. EM algoritmus pre určenie maximálnej vierohodnosti pre daný počet komponentov bol vždy inicializovaný výsledkom klasifikačného EM algorimtu (pozri str. 24) pre daný počet komponentov (tento algoritmus bol zasa inicializovaný náhodne, a to 10 krát). Pre

[43]Pôvodne som testoval aj NEC kritérium, ale neskôr som zistil, že som ho implementoval chybne. Chybu som síce opravil, ale simuláciu sa mi pre toto kritérium už z časového dôvodu nepodarilo zopakovať.

ukončenie EM algoritmu bolo použité Aitkenovo ukončovacie kritérium s toleranciou 1.10^{-5} alebo prekročenie maximálneho počtu iterácií stanoveného na hodnotu 500.

3.4.1 Poznámky k výsledkom simulácií

V prípade náhodného výberu o rozsahu $n = 100$ si môžeme všimnúť nestabilitu výsledkov u jednotlivých kritérií, ktorá sa najviac prejavuje v prípade kritéria CLC. Toto kritérium aj v jasných prípadoch, ako sú zmesi MW6 a MW7, veľmi preceňuje počet zhlukov. Ďalšiu nestabilitu možno pozorovať u kritérií AIC a KER1, čo je spôsobené zrejme tým, že reagujú na väčší počet komponentov v zmesi (MW2, MW3, MW9 až MW14). Tieto kritéria však nadhodnocujú počet komponentov aj v jednoduchých prípadoch (KER1 trochu menej ako AIC). Kritérium BIC, ale najmä kritériá KER3 a KER4 naopak ostávajú voči väčšiemu počtu komponentov chladné.

Pre rozsah $n = 200$ sú výsledky zretelnejšie a jasnejšie. Nestabilita CLC kritéria sa výrazne zredukovala. Jasné zlepšenie vidíme aj u kritérií AIC a KER1, ktoré už tak bezhlavo nenadhodnocujú a v prípade náročnejších zmesí s väčším počtom komponentov sa odhad hromadí okolo v blízkosti skutočného počtu komponentov (s výnimkou zmesi MW11). Taktiež si môžeme všimnúť, že na väčší počet komponentov začína reagovať aj kritérium BIC.

Pri ďalšom zväčšení rozsahu na $n = 500$ sú výsledky opäť o poznanie lepšie a stabilnejšie. Môžeme si všimnúť znovu zlepšenie u AIC a KER1, a tiež to, že BIC už výraznejšie reaguje na väčší počet komponentov (výnimkou je opäť zmes MW11). Možno už pozorovať, že kritérium KER1 dáva lepšie výsledky ako kritérium AIC (menej preceňuje u zmesí s malým počtom komponentov a dáva stabilnejšie a lepšie výsledky pre zmesi s väčším počtom komponentov). MIR, ICL a CLC zaostávajú.

Napokon v prípade rozsahu $n = 1000$ sú výsledky najstabilnejšie. KER1 dosahuje excelentný výsledok pre zmes MW9 a v priemere možno povedať, že spolu s BIC kritériom dosahuje najlepší výsledok. Úplne nerozlúsknutým orieškom zostáva zmes MW11, čo však nie je nič prekvapujúce vzhľadom k tomu, že jej šesť komponentov s miniatúrnym rozptylom (tzv. „hrebienkov") má také malé zastúpenie, že v náhodnom výbere o rozsahu $n = 1000$ sa v priemere nachádzajú len tri pozorovania z každého tohoto komponentu. Tu by sme rozhodne potrebovali väčší rozsah náhodného výberu.

Na záver teda možno konštatovať, že simulácie odhalili pozitívny výsledok kritéria KER1 pri výbere modelu normálnej zmesi. Ukázalo sa, že toto kritérium vypĺňa priestor medzi AIC kritériom, ktoré príliš nadhodnocuje počet komponentov a BIC kritériom, ktoré pri malom rozsahu zase príliš penalizuje a tak neumožňuje odhaliť väčší počet komponentov v zmesi v náročnejších prípadoch. Dá sa povedať, že s rastúcim n sme mohli u kritérií KER1 a BIC pozorovať aj konvergenciu odhadu ku skutočnému počtu komponentov. Penalizácia KER3 a KER4 sa ukázala pri našom rozsahu (do 1000) ako príliš silná. Kritérium AIC očakávane príliš preceňovalo počet komponentov a kritéria MIR, ICL a CLC sa nejak významne neosvedčili.

Na obrázkoch 3.5 a 3.6 môžeme ešte vidieť porovnanie hustoty maximálne vierohodného odhadu zmesi pri použití penalizácie KER1 a BIC v prípade náhodného výberu o rozsahu 500 u šiestich vybraných normálnych zmesí.

n = 100	**MW2**									**MW3**										**MW4**									
	2	*3*4*		*5*	*6*	*7*	*8*	*9*	*10*	*2*	*3*	*4*	*5*	*6*	*7*	*8*9*		*10*		*2*3*		*4*	*5*	*6*	*7*	*8*	*9*	*10*	
AIC	30	4	5	3	2	1	1	1	3	.	4	9	13	6	2	5	5	6		8	11	8	7	2	2	4	5	3	
KER1	33	4	3	3	2	1	1	.	3	.	10	9	13	7	1	5	3	2		11	15	10	4	2	2	1	2	3	
BIC	50	.	.	.	.	.	.	.	.	8	38	4	.	.	.	.	.	.		15	32	2	1	.	.	.	.	.	
KER3	50	.	.	.	.	.	.	.	.	50	.	.	.	.	.	.	.	.		48	2	.	.	.	.	.	.	.	
KER4	50	.	.	.	.	.	.	.	.	50	.	.	.	.	.	.	.	.		50	.	.	.	.	.	.	.	.	
CLC	9	1	1	.	.	3	4	3	29	.	.	.	.	4	4	11	10	21		7	.	.	1	3	2	7	9	21	
ICL	35	9	3	.	2	1	.	.	.	36	4	4	5	1	.	.	.	.		44	3	1	2	.	.	.	.	.	
MIR	17	7	4	4	1	3	4	5	5	30	4	5	2	1	2	5	.	1		24	7	5	2	4	1	.	2	5	
$n = 100$	**MW5**									**MW6**										**MW7**									
	*2*3*		*4*	*5*	*6*	*7*	*8*	*9*	*10*	*2*3*		*4*	*5*	*6*	*7*	*8*	*9*	*10*		*2*3*		*4*	*5*	*6*	*7*	*8*	*9*	*10*	
AIC	15	13	7	.	3	6	1	4	1	21	8	7	6	4	2	.	1	1		22	9	9	3	1	.	2	2	2	
KER1	19	13	5	.	3	5	1	3	1	23	10	6	4	4	2	.	1	.		27	9	8	3	1	1	.	.	1	
BIC	47	2	1	.	.	.	.	.	.	48	2	.	.	.	.	.	.	.		48	2	.	.	.	.	.	.	.	
KER3	49	.	1	.	.	.	.	.	.	50	.	.	.	.	.	.	.	.		50	.	.	.	.	.	.	.	.	
KER4	49	.	1	.	.	.	.	.	.	50	.	.	.	.	.	.	.	.		50	.	.	.	.	.	.	.	.	
CLC	14	12	3	1	2	2	3	8	5	.	1	1	3	6	3	6	11	19		16	4	3	2	6	.	2	7	10	
ICL	40	9	1	.	.	.	.	.	.	42	7	.	1	.	.	.	.	.		49	1	.	.	.	.	.	.	.	
MIR	36	5	2	1	1	1	.	2	2	24	12	5	1	2	2	2	1	1		46	2	1	.	.	1	.	.	.	
$n = 100$	**MW8**									**MW9**										**MW10**									
	*2*3*		*4*	*5*	*6*	*7*	*8*	*9*	*10*	*2*	*3*4*		*5*	*6*	*7*	*8*	*9*	*10*		*2*	*3*	*4*	*5*	*6*7*		*8*	*9*	*10*	
AIC	15	10	4	4	2	4	6	2	3	18	6	6	6	5	1	2	6	.		3	6	9	7	3	3	4	10	5	
KER1	20	8	4	4	1	4	4	2	3	25	4	4	4	5	2	1	5	.		5	7	10	7	3	3	4	6	5	
BIC	47	2	.	.	.	.	1	.	.	47	3	.	.	.	.	.	.	.		39	8	1	.	.	1	.	1	.	
KER3	49	.	.	.	.	.	1	.	.	50	.	.	.	.	.	.	.	.		48	.	.	.	.	1	.	1	.	
KER4	49	.	.	.	.	.	1	.	.	50	.	.	.	.	.	.	.	.		48	.	.	.	.	1	.	1	.	
CLC	2	2	1	.	1	3	7	10	24	2	.	.	1	1	5	6	17	18		1	.	.	.	1	2	5	16	25	
ICL	33	13	3	1	.	.	.	.	.	46	3	1	.	.	.	.	.	.		24	12	8	6	.	.	.	.	.	
MIR	13	18	2	3	4	1	1	3	5	22	7	4	3	1	5	5	1	2		6	8	9	5	5	3	3	6	5	
$n = 100$	**MW11**									**MW12**										**MW14**									
	2	*3*	*4*	*5*	*6*	*7*	*8*	*9*10*		*2*	*3*	*4*	*5*	*6*7*		*8*	*9*	*10*		*2*	*3*	*4*	*5*	*6*7*		*8*	*9*	*10*	
AIC	16	7	9	4	2	7	2	1	2	8	7	10	4	3	5	1	7	5		.	1	3	5	12	9	10	4	6	
KER1	18	6	11	4	2	5	2	.	2	9	8	10	4	3	7	1	6	2		.	3	5	8	11	8	8	2	5	
BIC	47	1	1	.	.	.	1	.	.	39	7	2	2	.	.	.	.	.		2	26	13	6	2	.	.	.	1	
KER3	49	.	.	.	.	.	1	.	.	49	1	.	.	.	.	.	.	.		49	.	.	.	.	.	.	.	1	
KER4	49	.	.	.	.	.	1	.	.	49	1	.	.	.	.	.	.	.		49	.	.	.	.	.	.	.	1	
CLC	1	1	3	3	2	4	5	5	26	.	.	3	2	5	2	7	10	21		.	1	.	3	7	11	9	8	11	
ICL	39	9	2	.	.	.	.	.	.	21	15	5	5	1	1	1	1	.		3	24	17	3	2	1	.	.	.	
MIR	15	12	9	2	6	.	4	1	1	15	7	6	4	3	4	2	5	4		28	15	2	2	1	.	1	.	1	

Tabuľka 3.2: Voľba počtu komponentov podľa rôznych informačných kritérií pre výbery z 12-tich typov normálnych MW zmesí o rozsahu $n = 100$ (počet náhodných výberov pre jeden typ zmesi je 50).

n = 200	**MW2**									**MW3**										**MW4**									
	2	*3*4*		*5*	*6*	*7*	*8*	*9*	*10*	*2*	*3*	*4*	*5*	*6*	*7*	*8*9*		*10*		*2*3*		*4*	*5*	*6*	*7*	*8*	*9*	*10*	
AIC	32	9	2	2	3	2	.	.	.	.	3	12	9	9	6	4	1	6		14	12	10	4	2	3	1	4	.	
KER1	36	7	3	1	1	2	.	.	.	.	4	19	12	6	3	3	.	3		14	20	7	2	2	1	2	2	.	
BIC	50	.	.	.	.	.	.	.	.	2	43	5	.	.	.	.	.	.		20	29	1	.	.	.	.	.	.	
KER3	50	.	.	.	.	.	.	.	.	50	.	.	.	.	.	.	.	.		41	9	.	.	.	.	.	.	.	
KER4	50	.	.	.	.	.	.	.	.	50	.	.	.	.	.	.	.	.		50	.	.	.	.	.	.	.	.	
CLC	24	2	3	.	1	.	4	5	11	5	.	1	.	1	11	5	18	9		31	1	1	3	2	3	4	2	3	
ICL	44	2	3	1	.	.	.	.	.	40	2	3	3	1	1	.	.	.		48	1	1	.	.	.	.	.	.	
MIR	16	6	5	2	4	2	4	6	5	37	6	1	1	.	.	2	1	2		29	1	1	4	4	3	4	4	.	
$n = 200$	**MW5**									**MW6**										**MW7**									
	*2*3*		*4*	*5*	*6*	*7*	*8*	*9*	*10*	*2*3*		*4*	*5*	*6*	*7*	*8*	*9*	*10*		*2*3*		*4*	*5*	*6*	*7*	*8*	*9*	*10*	
AIC	33	6	3	2	2	1	.	3	.	30	10	6	.	1	.	1	2	.		30	7	4	4	1	1	3	.	.	
KER1	34	9	4	1	.	.	.	2	.	35	7	6	.	1	.	.	1	.		38	6	4	2	.	.	.	.	.	
BIC	48	1	1	.	.	.	.	.	.	49	1	.	.	.	.	.	.	.		50	.	.	.	.	.	.	.	.	
KER3	48	1	1	.	.	.	.	.	.	49	1	.	.	.	.	.	.	.		50	.	.	.	.	.	.	.	.	
KER4	48	1	1	.	.	.	.	.	.	49	1	.	.	.	.	.	.	.		50	.	.	.	.	.	.	.	.	
CLC	33	15	1	1	.	.	.	.	.	23	4	2	3	3	.	3	6	6		45	4	.	.	1	.	.	.	.	
ICL	42	7	.	1	.	.	.	.	.	50	.	.	.	.	.	.	.	.		49	1	.	.	.	.	.	.	.	
MIR	39	4	2	.	2	.	.	2	1	31	6	5	3	.	.	2	2	1		43	2	.	.	1	1	1	1	1	
$n = 200$	**MW8**									**MW9**										**MW10**									
	*2*3*		*4*	*5*	*6*	*7*	*8*	*9*	*10*	*2*	*3*4*		*5*	*6*	*7*	*8*	*9*	*10*		*2*	*3*	*4*	*5*	*6*7*		*8*	*9*	*10*	
AIC	29	8	5	2	2	2	.	1	1	21	13	5	5	1	3	1	.	1		1	4	2	7	12	11	4	5	4	
KER1	33	7	4	2	2	2	.	.	.	27	13	3	4	1	2	.	.	.		2	5	5	8	11	10	4	4	1	
BIC	49	1	.	.	.	.	.	.	.	46	4	.	.	.	.	.	.	.		32	10	8	.	.	.	.	.	.	
KER3	49	1	.	.	.	.	.	.	.	48	2	.	.	.	.	.	.	.		49	1	.	.	.	.	.	.	.	
KER4	49	1	.	.	.	.	.	.	.	48	2	.	.	.	.	.	.	.		49	1	.	.	.	.	.	.	.	
CLC	14	6	2	1	4	2	7	3	11	20	4	4	3	2	3	3	5	6		7	11	5	2	2	3	2	8	10	
ICL	40	7	2	1	.	.	.	.	.	45	5	.	.	.	.	.	.	.		17	18	11	3	1	.	.	.	.	
MIR	20	7	5	5	3	2	.	4	4	22	9	2	3	2	.	4	2	6		5	8	11	9	3	2	1	3	8	
$n = 200$	**MW11**									**MW12**										**MW14**									
	2	*3*	*4*	*5*	*6*	*7*	*8*	*9*10*		*2*	*3*	*4*	*5*	*6*7*		*8*	*9*	*10*		*2*	*3*	*4*	*5*	*6*7*		*8*	*9*	*10*	
AIC	28	10	7	2	1	1	.	.	1	2	6	7	9	9	3	5	3	6		.	.	6	8	7	8	11	8	2	
KER1	34	10	5	.	1	.	.	.	.	2	8	9	13	8	3	3	2	2		.	.	10	8	6	7	11	7	1	
BIC	49	1	.	.	.	.	.	.	.	40	9	1	.	.	.	.	.	.		.	11	25	11	2	1	.	.	.	
KER3	50	.	.	.	.	.	.	.	.	50	.	.	.	.	.	.	.	.		50	.	.	.	.	.	.	.	.	
KER4	50	.	.	.	.	.	.	.	.	50	.	.	.	.	.	.	.	.		50	.	.	.	.	.	.	.	.	
CLC	25	8	5	1	3	1	2	2	3	9	2	2	4	5	8	5	6	9		.	4	15	13	11	2	3	1	1	
ICL	46	4	.	.	.	.	.	.	.	22	9	5	9	3	1	1	.	.		.	12	25	11	2	.	.	.	.	
MIR	30	4	3	2	2	2	2	2	3	24	4	2	5	5	4	3	.	3		19	17	5	7	.	.	1	1	.	

Tabuľka 3.3: Voľba počtu komponentov podľa rôznych informačných kritérií pre výbery z 12-tich typov normálnych MW zmesí o rozsahu $n = 200$ (počet náhodných výberov pre jeden typ zmesi je 50).

n = 500	MW2									MW3									MW4								
	2	*3*4*		*5*	*6*	*7*	*8*	*9*	*10*	*2*	*3*	*4*	*5*	*6*	*7*	*8*9*		*10*	*2*3*		*4*	*5*	*6*	*7*	*8*	*9*	*10*
AIC	37	5	3	1	1	1	.	2	.	.	.	5	18	7	7	7	1	5	26	12	5	3	1	1	.	.	2
KER1	43	6	.	.	.	1	.	.	.	.	.	10	25	3	6	4	.	2	31	16	2	1	.	.	.	.	.
BIC	50	.	.	.	.	.	.	.	.	.	11	37	2	.	.	.	.	.	32	17	1	.	.	.	.	.	.
KER3	50	.	.	.	.	.	.	.	.	48	2	.	.	.	.	.	.	.	32	17	1	.	.	.	.	.	.
KER4	50	.	.	.	.	.	.	.	.	50	.	.	.	.	.	.	.	.	50	.	.	.	.	.	.	.	.
CLC	42	1	1	2	1	.	.	3	.	34	.	.	.	2	.	1	6	7	48	.	.	.	1	.	1	.	
ICL	46	2	1	1	.	.	.	.	.	48	1	.	1	.	.	.	.	.	50	.	.	.	.	.	.	.	.
MIR	13	4	5	5	5	4	4	6	4	35	10	1	.	.	.	.	3	1	27	.	3	2	4	1	5	6	2

n = 500	MW5									MW6									MW7								
	*2*3*		*4*	*5*	*6*	*7*	*8*	*9*	*10*	*2*3*		*4*	*5*	*6*	*7*	*8*	*9*	*10*	*2*3*		*4*	*5*	*6*	*7*	*8*	*9*	*10*
AIC	34	6	4	2	.	2	1	1	.	38	2	3	5	.	2	.	.	.	30	8	1	2	3	4	1	.	1
KER1	39	8	2	.	.	.	.	1	.	45	2	1	.	.	2	.	.	.	39	5	3	1	1	1	.	.	.
BIC	48	2	.	.	.	.	.	.	.	50	.	.	.	.	.	.	.	.	49	1	.	.	.	.	.	.	.
KER3	48	2	.	.	.	.	.	.	.	50	.	.	.	.	.	.	.	.	49	1	.	.	.	.	.	.	.
KER4	48	2	.	.	.	.	.	.	.	50	.	.	.	.	.	.	.	.	49	1	.	.	.	.	.	.	.
CLC	42	8	.	.	.	.	.	.	.	49	1	.	.	.	.	.	.	.	49	1	.	.	.	.	.	.	.
ICL	46	4	.	.	.	.	.	.	.	49	1	.	.	.	.	.	.	.	50	.	.	.	.	.	.	.	.
MIR	41	4	1	.	1	.	1	1	1	45	2	1	.	.	1	1	.	.	40	.	.	.	2	2	2	1	3

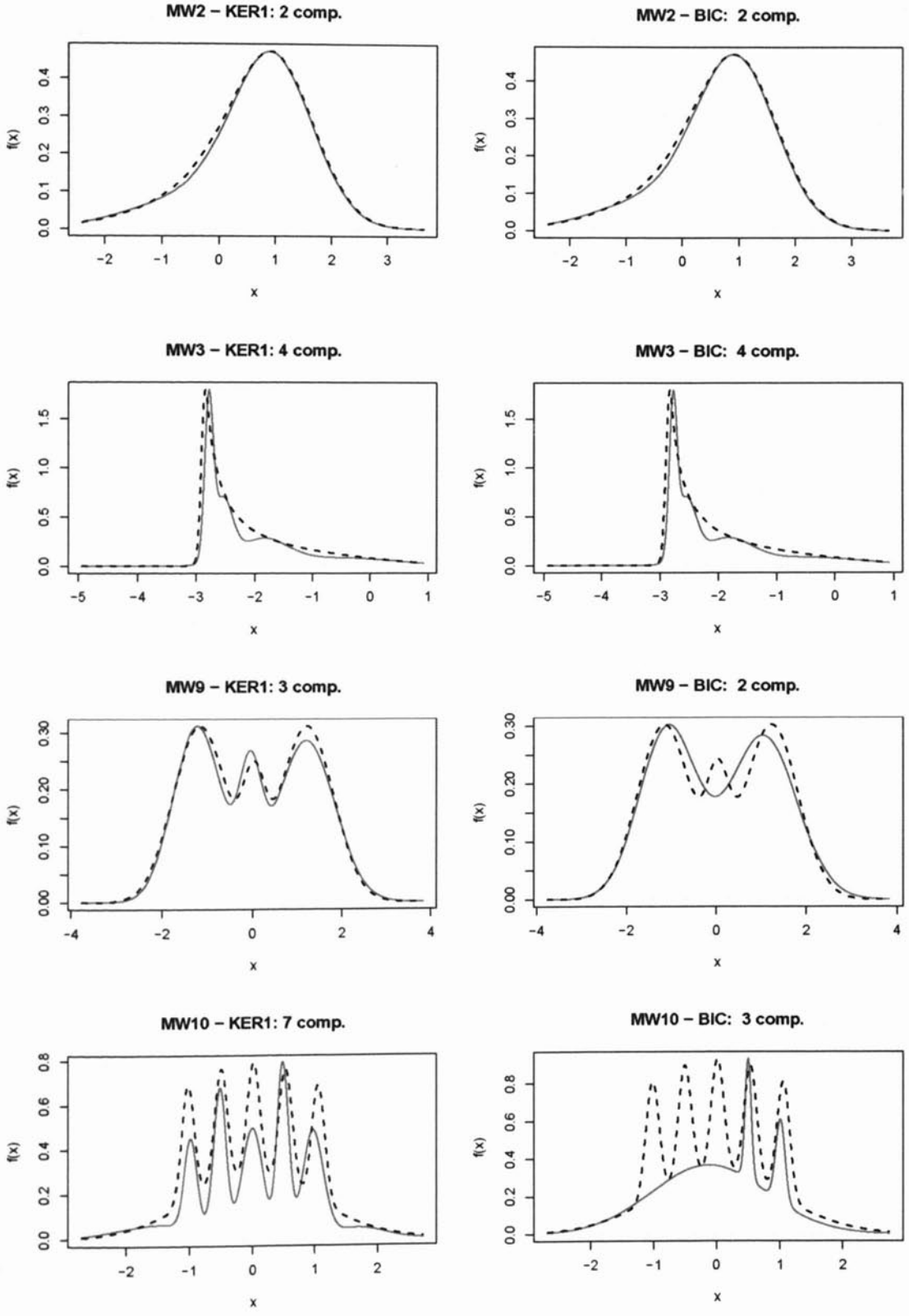
n = 500	MW8									MW9									MW10								
	*2*3*		*4*	*5*	*6*	*7*	*8*	*9*	*10*	*2*	*3*4*		*5*	*6*	*7*	*8*	*9*	*10*	*2*	*3*	*4*	*5*	*6*7*		*8*	*9*	*10*
AIC	35	6	3	.	2	2	1	1	.	4	22	7	7	5	.	1	3	1	.	.	.	.	3	21	11	8	7	
KER1	43	3	2	.	1	1	.	.	.	6	30	6	4	2	1	.	1	.	.	.	1	.	4	25	9	5	6	
BIC	49	1	.	.	.	.	.	.	.	39	11	.	.	.	.	.	.	.	19	8	6	4	2	10	1	.	.	
KER3	49	1	.	.	.	.	.	.	.	50	.	.	.	.	.	.	.	.	47	3	.	.	.	.	.	.	.	
KER4	49	1	.	.	.	.	.	.	.	50	.	.	.	.	.	.	.	.	47	3	.	.	.	.	.	.	.	
CLC	44	2	2	.	1	.	.	1	.	40	10	.	.	.	.	.	.	.	9	8	13	2	.	8	3	2	5	
ICL	47	2	.	.	1	.	.	.	.	46	4	.	.	.	.	.	.	.	13	12	13	3	.	5	3	.	1	
MIR	23	9	5	4	2	4	1	1	1	17	13	2	.	1	6	7	2	2	1	9	13	9	2	4	2	3	7	

n = 500	MW11									MW12									MW14								
	2	*3*	*4*	*5*	*6*	*7*	*8*	*9*10*		*2*	*3*	*4*	*5*	*6*7*		*8*	*9*	*10*	*2*	*3*	*4*	*5*	*6*7*		*8*	*9*	*10*
AIC	27	9	1	5	4	1	1	1	1	.	.	1	12	7	8	7	11	4	.	.	.	5	6	10	12	8	9
KER1	34	7	2	1	4	1	.	.	1	.	.	3	12	10	7	6	10	2	.	.	1	6	6	13	10	6	8	
BIC	49	1	.	.	.	.	.	.	.	17	8	14	8	2	.	1	.	.	.	.	27	14	5	3	.	1	.	
KER3	49	1	.	.	.	.	.	.	.	50	.	.	.	.	.	.	.	.	22	28	.	.	.	.	.	.	.	
KER4	49	1	.	.	.	.	.	.	.	50	.	.	.	.	.	.	.	.	50	.	.	.	.	.	.	.	.	
CLC	47	3	.	.	.	.	.	.	.	12	3	5	10	7	7	4	1	1	.	4	29	13	3	1	.	.	.	
ICL	50	.	.	.	.	.	.	.	.	17	9	4	10	4	4	1	1	.	.	4	33	12	1	.	.	.	.	
MIR	34	1	.	2	.	2	2	2	7	18	4	16	5	1	2	.	3	1	13	12	12	3	.	1	1	2	6	

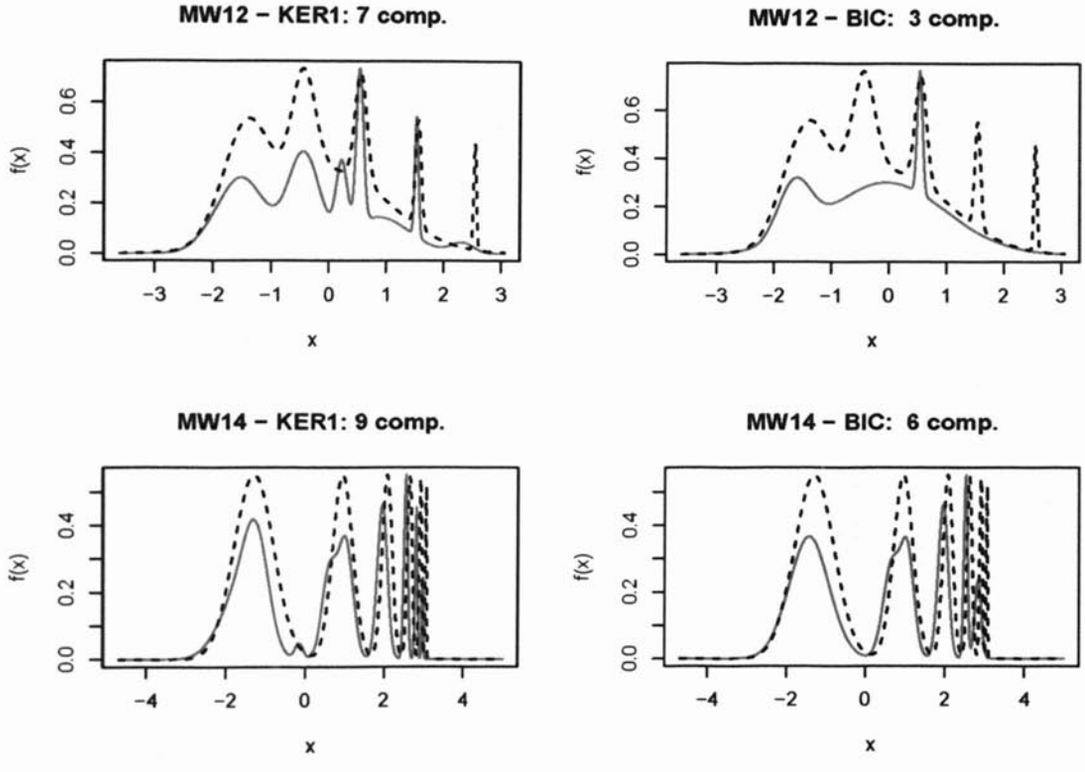
Tabuľka 3.4: Voľba počtu komponentov podľa rôznych informačných kritérií pre výbery z 12-tich typov normálnych MW zmesí o rozsahu $n = 500$ (počet náhodných výberov pre jeden typ zmesi je 50).

n = 1000	**MW2**									**MW3**										**MW4**									
	2	*3*4*		*5*	*6*	*7*	*8*	*9*	*10*	*2*	*3*	*4*	*5*	*6*	*7*	*8*9*		*10*		*2*3*		*4*	*5*	*6*	*7*	*8*	*9*	*10*	
AIC	17	1	1	.	1	.	.	.	.	.	.	.	10	2	3	2	3	.		10	8	.	1	1	.	.	.	.		
KER1	19	.	.	.	1	.	.	.	.	.	.	3	13	4	.	.	.	.		11	8	.	.	1	.	.	.	.		
BIC	20	.	.	.	.	.	.	.	.	.	.	19	1	.	.	.	.	.		13	7	.	.	.	.	.	.	.		
KER3	20	.	.	.	.	.	.	.	.	5	15	.	.	.	.	.	.	.		13	7	.	.	.	.	.	.	.		
KER4	20	.	.	.	.	.	.	.	.	20	.	.	.	.	.	.	.	.		20	.	.	.	.	.	.	.	.		
CLC	19	.	.	.	1	.	.	.	.	20	.	.	.	.	.	.	.	.		20	.	.	.	.	.	.	.	.		
ICL	19	.	.	.	1	.	.	.	.	20	.	.	.	.	.	.	.	.		20	.	.	.	.	.	.	.	.		
MIR	13	2	2	.	.	.	1	2	.	14	3	.	.	.	.	1	2	.		16	.	.	.	.	1	.	1	2		
$n = 1000$	**MW5**									**MW6**										**MW7**									
	*2*3*		*4*	*5*	*6*	*7*	*8*	*9*	*10*	*2*3*		*4*	*5*	*6*	*7*	*8*	*9*	*10*		*2*3*		*4*	*5*	*6*	*7*	*8*	*9*	*10*	
AIC	15	4	.	.	.	1	.	.	.	18	2	.	.	.	.	.	.	.		14	4	.	.	.	1	.	1	.		
KER1	17	3	.	.	.	.	.	.	.	18	2	.	.	.	.	.	.	.		17	3	.	.	.	.	.	.	.		
BIC	18	2	.	.	.	.	.	.	.	19	1	.	.	.	.	.	.	.		20	.	.	.	.	.	.	.	.		
KER3	18	2	.	.	.	.	.	.	.	20	.	.	.	.	.	.	.	.		20	.	.	.	.	.	.	.	.		
KER4	18	2	.	.	.	.	.	.	.	20	.	.	.	.	.	.	.	.		20	.	.	.	.	.	.	.	.		
CLC	15	5	.	.	.	.	.	.	.	19	1	.	.	.	.	.	.	.		20	.	.	.	.	.	.	.	.		
ICL	17	3	.	.	.	.	.	.	.	20	.	.	.	.	.	.	.	.		20	.	.	.	.	.	.	.	.		
MIR	13	1	.	.	.	1	1	1	3	19	.	.	.	.	.	1	.	.		9	.	.	.	1	3	3	2	2		
$n = 1000$	**MW8**									**MW9**										**MW10**									
	*2*3*		*4*	*5*	*6*	*7*	*8*	*9*	*10*	*2*	*3*4*		*5*	*6*	*7*	*8*	*9*	*10*		*2*	*3*	*4*	*5*	*6*7*		*8*	*9*	*10*	
AIC	18	2	.	.	.	.	.	.	.	.	15	1	1	2	1	.	.	.		.	.	.	.	2	5	6	5	2		
KER1	18	2	.	.	.	.	.	.	.	.	20	.	.	.	.	.	.	.		.	.	.	.	2	6	8	3	1		
BIC	20	.	.	.	.	.	.	.	.	9	11	.	.	.	.	.	.	.		.	.	1	1	2	9	7	.	.		
KER3	20	.	.	.	.	.	.	.	.	20	.	.	.	.	.	.	.	.		20	.	.	.	.	.	.	.	.		
KER4	20	.	.	.	.	.	.	.	.	20	.	.	.	.	.	.	.	.		20	.	.	.	.	.	.	.	.		
CLC	20	.	.	.	.	.	.	.	.	20	.	.	.	.	.	.	.	.		2	3	11	4	.	.	.	.	.		
ICL	20	.	.	.	.	.	.	.	.	20	.	.	.	.	.	.	.	.		2	3	11	4	.	.	.	.	.		
MIR	14	.	.	.	.	1	1	1	3	8	3	.	.	2	3	.	1	3		.	2	6	1	.	.	3	5	3		
$n = 1000$	**MW11**									**MW12**										**MW14**									
	2	*3*	*4*	*5*	*6*	*7*	*8*	*9*10*		*2*	*3*	*4*	*5*	*6*7*		*8*	*9*	*10*		*2*	*3*	*4*	*5*	*6*7*		*8*	*9*	*10*	
AIC	18	1	1	.	.	.	.	.	.	.	.	1	3	3	3	2	2	6		.	.	.	2	2	1	4	5	6		
KER1	19	.	1	.	.	.	.	.	.	.	.	1	4	4	3	2	2	4		.	.	.	2	4	3	4	4	3		
BIC	20	.	.	.	.	.	.	.	.	1	2	4	3	7	2	.	1	.		.	.	1	4	5	4	3	3	.		
KER3	20	.	.	.	.	.	.	.	.	20	.	.	.	.	.	.	.	.		.	19	1	.	.	.	.	.	.		
KER4	20	.	.	.	.	.	.	.	.	20	.	.	.	.	.	.	.	.		20	.	.	.	.	.	.	.	.		
CLC	20	.	.	.	.	.	.	.	.	5	1	.	6	6	.	2	.	.		.	4	9	5	1	1	.	.	.		
ICL	20	.	.	.	.	.	.	.	.	6	2	.	6	5	.	1	.	.		.	4	10	5	.	1	.	.	.		
MIR	13	.	.	.	.	.	1	2	4	2	12	1	1	.	.	1	1	2		5	6	2	3	.	1	1	.	2		

Tabuľka 3.5: Voľba počtu komponentov podľa rôznych informačných kritérií pre výbery z 12-tich typov normálnych MW zmesí o rozsahu $n = 1000$ (počet náhodných výberov pre jeden typ zmesi je 20).



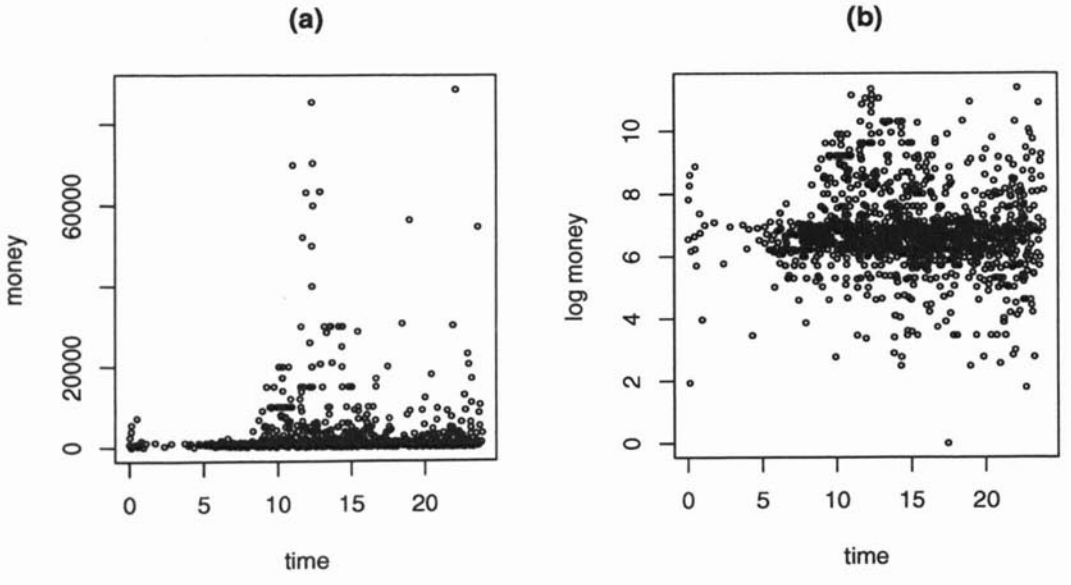
Obrázok 3.5: Hustota odhadu zmesi v prípade kritéria KER1 a BIC z náhodných výberov vybraných normálnych zmesí o rozsahu 500. Prerušovaná čiara znázorňuje hustotu skutočnej zmesi. Číslo vedľa kritéria v názve každého grafu vyjadruje počet komponentov odhadu zmesi (skutočný počet komponentov je postupne 3, 8, 3, 6).



Obrázok 3.6: Hustota odhadu zmesi v prípade kritéria KER1 a BIC z náhodných výberov vybraných normálnych zmesí o rozsahu 500. Prerušovaná čiara znázorňuje hustotu skutočnej zmesi. Číslo vedľa kritéria v názve každého grafu vyjadruje počet komponentov odhadu zmesi (skutočný počet komponentov v oboch prípadoch je 6).

3.5 Bilancia platieb - odhad reálnej zmesi

V tejto časti sa pokúsim odhadnúť model zmesi na reálnom príklade dvojrozmerných dát. Analyzované dáta predstavujú bilanciu platieb kartami na šiestich termináloch. Prvý rozmer dát reprezentuje čas výberu (spojitá premenná s hodnotami medzi 0 a 24) a druhý zase vyberanú sumu. Data sú súhrnom 280-tich denných bilancií, pričom pre každý deň máme maximálne 30 záznamov. Celkový rozsah dát je 2746, z toho asi 100 reprezentujú výber menší ako sto korún, približne 1700 výber do tisíc korún, 900 je potom výber medzi tisíc a desať tisíc korún a niečo cez sto pozorovaní je výber väčší ako desať tisíc korún, pričom maximálny výber má hodnotu 88600 korún. Vzhľadom k tomuto nerovnomernému rozloženiu som sa rozhodol veľkosť výberu pred analýzou logaritmicky transformovať. Dáta sa nachádzajú na obrázku 3.7, z ktorého môžeme pekne vidieť ako sa transformáciou stratili odľahlé pozorovania (výbery nad 10 resp. viac tisíc). Výbery s hodnotou menšou ako sto by sa mohli z analýzy vylúčiť (ide zrejme o chybné dáta), ale vzhľadom k tomu, že ich nie je príliš veľa a vzhľadom k zavedenej transformácii si myslím, že výsledok významne neovplyvnia. Dáta som



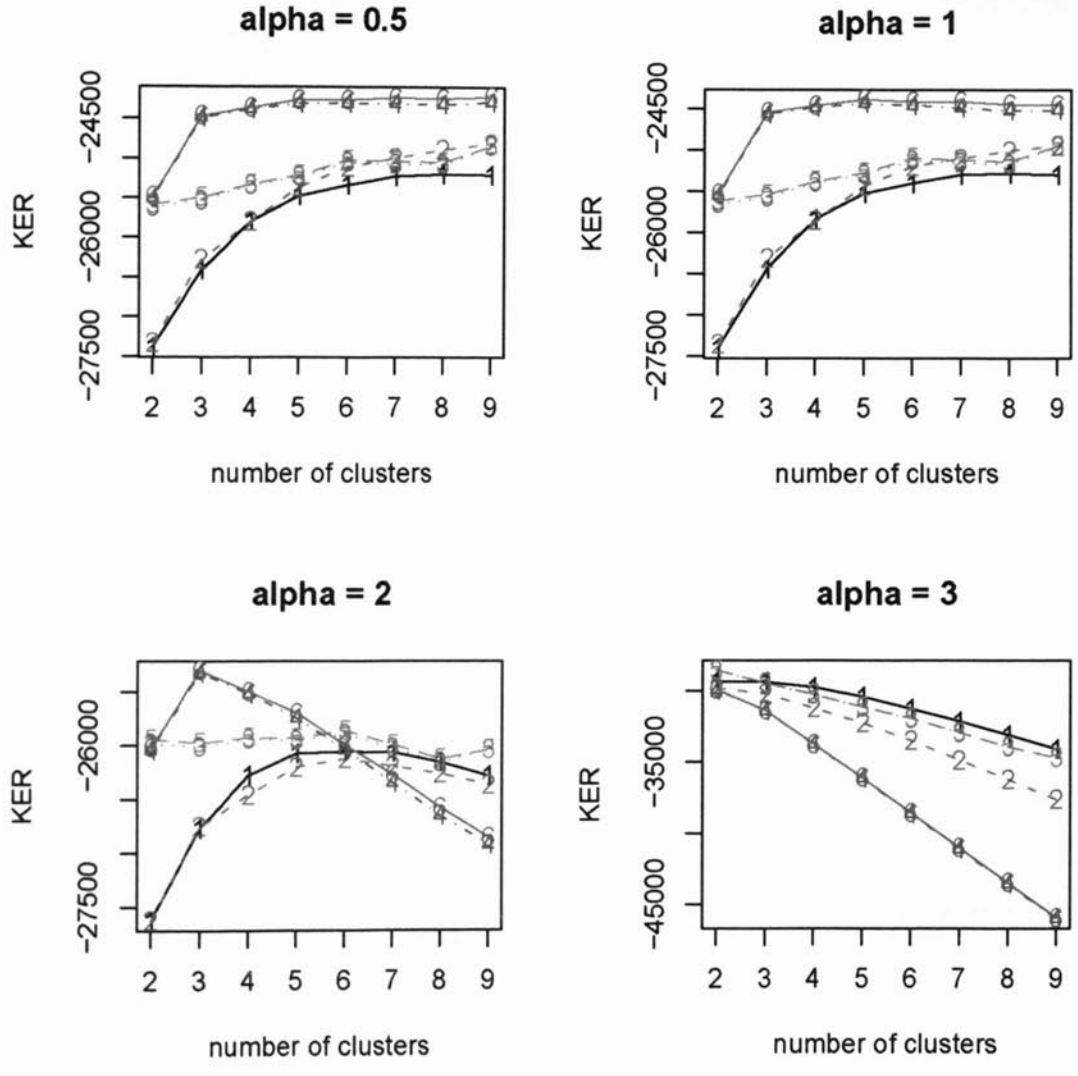
Obrázok 3.7: Bilancia platieb kartami - (a) originálne dáta, (b) transformované dáta

modeloval pomocou dvojrozmernoej Gaussovskej zmesi. Na nájdenie odhadu som použil metódu MPLE s Keribinovou penalizáciou pre $\alpha = 1/2$, 1, 2 a 3 aplikovanú použitím funkcie `Mclust` a `MclustKer`[44]. Pri použití BIC kritéria bol zvolený model zmesi s piatimi zhlukmi s rôznou veľkosťou, tvarom aj orientáciou (model VVV). Výsledný odhad algoritmu je:

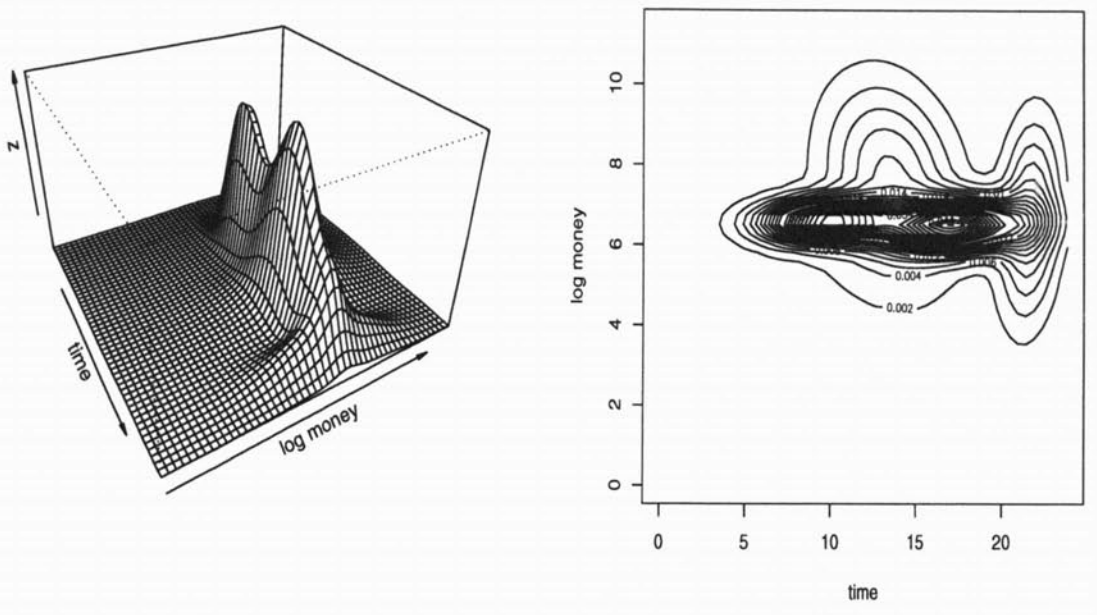
[44]Implementovaný algoritmus sa pri hľadaní odhadu niekoľkokrát zrútil, zrejme kvôli nejakej singularite, do ktorej sa po ceste dostal. Bolo teda potrebné opakovať ho viac krát

```
pro 0.26293961 0.22530260 0.30755330 0.13614278 0.06806171

mu
             1        2         3         4        5
[1,] 13.540022 9.832861 17.339135 21.689692 9.850493
[2,]  7.744889 6.608830  6.492791  6.664267 5.352578

sigma

, , 1                    , , 2                          , , 3
          [,1]      [,2]           [,1]       [,2]           [,1]      [,2]
[1,]  6.804819 -0.895997 [1,] 5.66239587 0.06664964 [1,] 6.3981986 0.1072761
[2,] -0.895997  2.470491 [2,] 0.06664964 0.11989126 [2,] 0.1072761 0.1764101

, , 4                    , , 5
          [,1]      [,2]           [,1]       [,2]
[1,] 1.3599163 0.3141198 [1,] 34.452055 -5.107934
[2,] 0.3141198 2.6271714 [2,] -5.107934  2.620507
```

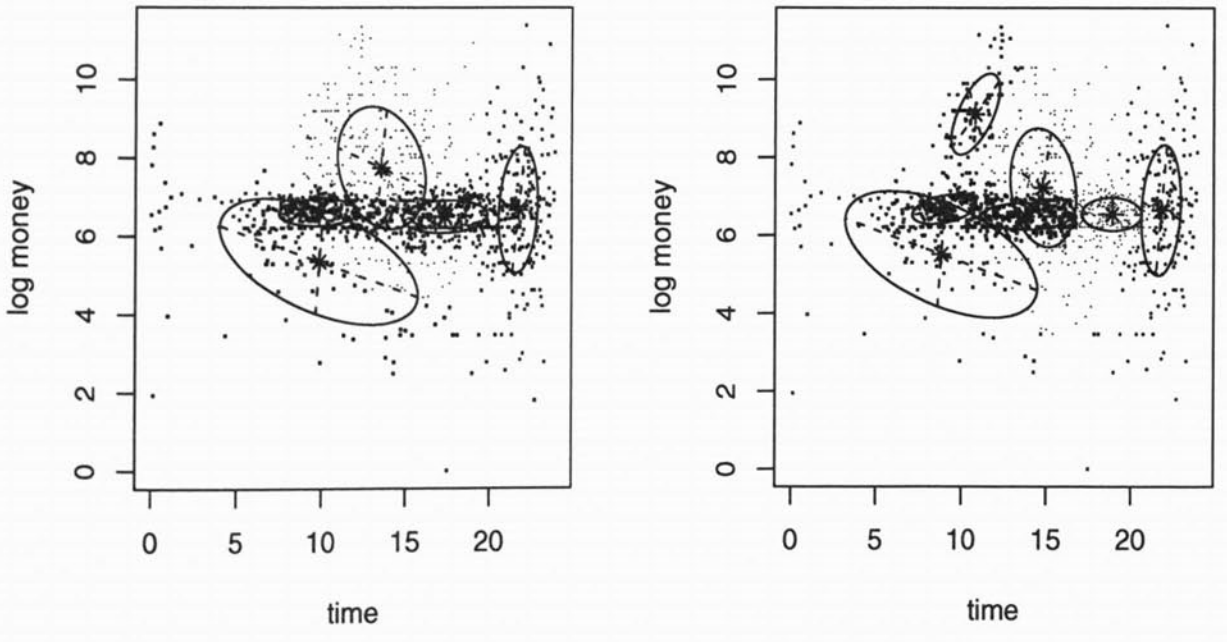
Rozhodovanie medzi modelom s piatimi až deviatimi zhlukmi je však veľmi tesné, čo môžeme vidieť na obrázku 3.8, ktorý znázorňuje hodnotu Keribinovho kritéria pre rôzne modely (EII, VII, EEI, VVI, EEE, VVV) a pre rôzne hodnoty parametra α. V prípade BIC kritéria ($\alpha = 1$) sa hodnota penalizovanej vierohodonosti líši v rámci modelu zmesi s tromi až deviatimi komponentami len minimálne. S rastúcou hodnotou parametra α Keribinovej penalizácie sa mení výsledný počet komponentov postupne z 9 (VVV) na spomínaných 5, ďalej na 3 (VVV) až napokon na 2 (EEE). Na obrázku 3.9 je znázornený graf hustoty aj vrstevnicový graf pre odhad založený na BIC kritériu. Pri opakovaní EM algoritmu som sa viac-krát dostal aj k modelu so siedmimi zhlukmi, čo nie je prekvapivé vzhľadom k tesnému výsledku medzi modelom s 5-timi až modelom s 9-timi zhlukmi. Na obrázku 3.10 je znázornená klasifikácia dát (zaradenie do zhlukov) pre model s piatimi a siedmimi zhlukmi, na ktorom môžeme vidieť, kde sa v bohatšom modele sformovali dva nové zhluky (obidva s proporciou približne 0.05). Výber modelu (medzi modelmi s 5-timi až 9-timi komponentami) záleží už na subjektívnom rozhodnutí. Ja osobne by som volil strednú cestu, a to model so siedmimi komponentami. Podstatné však je, že modelom sa nám podarilo preukázať prítomnosť zhlukov v dátach.



Obrázok 3.8: Keribinovo kritérium pre voľbu počtu komponentov v prípade dát o platobnej bilancii pre šesť uvažovaných modelov (1-EII, 2-VII, 3-EEI, 4-VVI, 5-EEE, 6-VVV) a pre štyri rôzne hodnoty α.



Obrázok 3.9: Graf hustoty a vrstevnicový graf pre model zmesi pri použití BIC kritéria v prípade dát o platobnej bilancii.



Obrázok 3.10: Klasifikácia v prípade modelu s piatimi zhlukmi a modelu so siedmimi zhlukmi.

Záver

Pokúsim sa teraz zhrnúť obsah predchádzajúcich stránok, to čo sa mi v práci podarilo a čo naopak nie.

V prvej časti práce som sa venoval základnej chrakteristike modelu so zmesou hustôt. Predstavil som model konečnej zmesi s hustotami pochádzajúcimi z rovnakej parametricky formulovanej rodiny rozdelení, ktorý sa stal ťažiskom môjho záujmu. Vysvetlil som problém identifikovateľnosti parametra modelu a predstavil základné metódy pre odhad parametrov modelu zmesi s pevným (známym) počtom komponentov. Medzi týmito dominuje EM algoritmus, ktorý poskytuje približné riešenie v prípade maximálne vierohodného odhadu. Algoritmus bol pôvodne navrhnutý pre dáta s chýbajúcimi pozorovaniami, ale uplatnenie našiel aj v prípade zmesí. V štatistike je tento algoritmus skutočne pojem a v súvislosti s modelmi zmesí mu bolo venované veľké množstvo pozornosti. V časti kapitoly som sa zaoberal aj týmto algoritmom a jeho základnými vlastnosťami a modifikáciami. Vysvetlil som, ako sa model zmesi použije k odhaľovaniu a klasifikácii zhlukov a ako pomocou neho možno šikovne modelovať pravdepodobnostné rozdelenia. Venoval som trochu aj problému mnohorozmerných dát a konkrétnym aplikáciam modelu zmesí v špeciálnych prípadoch.

Druhá a hlavná časť práce bola zameraná na odhad zložitosti modelu zmesi, teda prípadu, keď počet komponentov modelu bolo potrebné odhadnúť z dát. Predstavil som štyri hlavné prístupy, a to metódu založenú na maximálnej vierohodnosti, minimálnej vzdialenosti, testovanie počtu komponentov pomerom vierohodností a bayesovský prístup. Urobil som prehľad základných existujúcich metód pre každý prípad. Popísal som problém, na ktorý narazíme pri testovaní pomerom vierohodností a akým spôsobom bol tento problém riešený. Poukázal som na ťažkosti s diagnostikou konzistencie v prípade maximalizovania penalizovanej vierohodnosti, ktorá viedla ku vzniku množstva informačných kritérií. Ukázal som, že v prípade metód založených na minimálnej vzdialenosti sa dá vzdialenosť penalizovať tak, aby bol odhad zložitosti zmesi konzistentný. Zároveň som poukázal na to, že s praktickou voľbou týchto penalizácií je problém, a že absentuje efektívny algoritmus na minimalizáciu vzdialenosti. Predstavil som hlavné bayesovské prístupy na odhad zložitosti modelu zmesi, ktoré sú založené z väčšej časti na MCMC metódach. MCMC metódy sú ale pomerne komplikované a ich aplikácia si vyžaduje dobrú znalosť problematiky - je potrebné vedieť vhodne zvoliť model, apriórne pravdepodobnosti a vysporiadať sa s permutáciou indexov komponentov zmesi. V záverečnej časti som poukázal na existenciu konzistentného odhadu metódou maximálnej penalizovanej vierohodnosti. Pekné na tomto výsledku je, že vzišiel z náročného analytického prístupu, ktorý má korene v testovaní hypotézy o počte komponentov pomerom viero-

hodností. Poukázal som aj na moderný vývoj v oblasti EM algoritmu, ktorý sa zaoberá procesom zlučovania a rozdeľovania zhlukov v priebehu iterovania. Tento pochádza z dielne teoretických informatikov, na čom je vidieť široké uplatnenie modelov zmesí, čo by mohlo byť hnacou silou ďalšieho vývoja v tejto oblasti.

V tretej a poslednej časti som sa pokúsil prezentovať môj výskum v oblasti modelu zmesi. Prezentoval som algoritmy, ktoré sa mi podarilo implementovať v prostredí `R` a pokúsil som sa s nimi experimentovať. V prípade metód založených na minimálnej vzdialenosti som pre účely minimalizácie vzdialenosti vyvinul jednoduchý genetický algoritmus, pomocou ktorého sa mi podarilo v jednoduchých prípadoch dospieť k riešeniu rýchlejšie ako pri použití klasického Nelder-Meadovho simplexovho algoritmu. Predovšetkým som si ale overil náročnosť aplikácie týchto metód. Predviedol som aj použitie jednej grafickej techniky pre odhad zložitosti normálnej zmesi. Pomocou simulácií sa mi podarilo ukázať výhodu Keribinovej modifikácie kritéria BIC (KER s $\alpha = 1/2$) pre normálne zmesi s rozsahom výberu do 1000. Tu si táto penalizácia počínala veľmi dobre a dala lepšie odhady pre počet komponentov skúmaných zmesí a aj lepší odhad hustoty ako kritérium BIC. Napokon sa mi ešte podarilo úspešne aplikovať model zmesi k diagnostike zhlukov na reálnom príklade dvojrozmerných dát.

Na záver by som chcel spomenúť to, čo sa mi do práce nepodarilo zahrnúť a čo by z môjho pohľadu bolo ešte zaujímavé preskúmať. Je to implementácia bayesovského kritéria RW, variačná procedúra bayesovského prístupu implementovaná v knižnici `vabayelmix` s odhadom počtu komponentov a algoritmus SSMEM uvedený v závere druhej kapitoly. Zaujímavá by mohla byť aj simulácia za účelom porovnania kritérií pre výber modelu exponenciálnych zmesí. Priestor je aj v oblasti vyšetrovania identifikovateľnosti a overovaní podmienok, potrebných na odvodenie konzistencie odhadu zložitosti zmesi pomocou kritéria KER.

Zoznam použitých skratiek

AIC	Akaikeho informačné kritérium
AKM	Algoritmus z rodiny metód MPDE
BIC	Bayesovské informačné kritérium
CLC	Klasifikačné informačné kritérium
CVIC	Informačné kritérium založené na cross-validácii
CvM	Cramér-von-Misesova vzdialenosť distribučných funkcií
EDF	Empirická distribučná funkcia
EIC	Efronovo informačné kritérium
GA	Genetický algoritmus
GLM	Zobecnený lineárny model
HC	Hierarchický zhlukovací algoritmus
ICL	Kritérium založené na integrovanej klasifikačnej vierohodnosti
ICL-BIC	Modifikovaná verzia kritéria ICL
ICOMP	Kritérium informačnej zložitosti
KANT	Kantorowichova vzdialenosť distribučných funkcií
KL	Kullback-Leiblerova divergencia, resp. vzdialenosť
KS	Kolmogorov-Smirnovova (suprémová) vzdialenosť distribučných funkcí
LRT	Test pomerom vierohodností
LRTS	Štatistika pre test pomerom vierohodností
MCMC	Monte Carlo Markov Chain - metóda bayesovskej analýzy
MDE	Odhad metódou minimálnej vzdialenosti
MDP	Detekčný graf pre identifikáciu zmesi zavedený v Roeder (1994)
MIR	Kritérium založené na informačnej matici pomeru
MLE	Maximálne vierohodný odhad
MLA	Metóda zhluk. analýzy založená na maximálnej vierohodnosti
MPDE	Odhad metódou minimálnej penalizovanej vzdialenosti
MPLE	Odhad založený na maximalizácii penalizovanej vierohodnosti
MW	Normálna zmes z článku Marron, Wand (1992)
NEC	Kritérium založené na normalizovanej entropii
NMSA	Nelder-Meadov simplexov algoritmus
RW	Bayesovské kritérium (Roeder, Wasserman, 1997)
SMEM	EM algoritmus s rozdeľovaním a zlučovaním zhlukov (Ueda, 2000)
SSMEM	EM algoritmus s rozdeľovaním a zlučovaním zhlukov (Wang a kol., 2004)

Dodatok A

Príklady hustôt normálnej zmesi

Názov	$f(x)$
1. Gaussian	$N(0,1)$
2. Skewed unimodal	$\frac{1}{5}N(0,1) + \frac{1}{5}N(\frac{1}{2},(\frac{2}{3})^2) + \frac{3}{5}N(\frac{13}{15},(\frac{5}{9})^2)$
3. Strongly skewed	$\sum_{i=0}^{7} \frac{1}{8}N(3\{(\frac{2}{3})^i - 1\},(\frac{2}{3})^{2i})$
4. Kurtotic unimodal	$\frac{2}{3}N(0,1) + \frac{1}{3}N(0,(\frac{1}{10})^2)$
5. Outlier	$\frac{1}{10}N(0,1) + \frac{9}{10}N(0,(\frac{1}{10})^2)$
6. Bimodal	$\frac{1}{2}N(-1,(\frac{2}{3})^2) + \frac{1}{2}N(1,(\frac{2}{3})^2)$
7. Separated bimodal	$\frac{1}{2}N(-\frac{3}{2},(\frac{1}{2})^2) + \frac{1}{2}N(\frac{3}{2},(\frac{1}{2})^2)$
8. Skewed bimodal	$\frac{3}{4}N(0,1) + \frac{1}{4}N(\frac{3}{2},(\frac{1}{3})^2)$
9. Trimodal	$\frac{9}{20}N(-\frac{6}{5},(\frac{3}{5})^2) + \frac{9}{20}N(\frac{6}{5},(\frac{3}{5})^2) + \frac{1}{10}N(0,(\frac{1}{4})^2)$
10. Claw	$\frac{1}{2}N(0,1) + \sum_{i=0}^{4} \frac{1}{10}N(\frac{i}{2} - 1,(\frac{1}{10})^2)$
11. Double claw	$\frac{49}{100}N(-1,(\frac{2}{3})^2) + \frac{49}{100}N(1,(\frac{2}{3})^2) + \sum_{i=0}^{6} \frac{1}{350}N(\frac{i-3}{2},(\frac{1}{100})^2)$
12. Assymetric claw	$\frac{1}{2}N(0,1) + \sum_{i=-2}^{2} \frac{2^{1-i}}{31}N(i + \frac{1}{2},(\frac{2^{-i}}{10})^2)$
13. Assymetric double claw	$\sum_{i=0}^{1} \frac{46}{100}N(2i - 1,(\frac{2}{3})^2) + \sum_{i=1}^{3} \frac{1}{300}N(\frac{-i}{2},(\frac{1}{100})^2) +$ $+ \sum_{i=1}^{3} \frac{7}{300}N(\frac{i}{2},(\frac{7}{100})^2)$
14. Smooth comb	$\sum_{i=0}^{5} \frac{2^{5-i}}{63}N((65 - 96(\frac{1}{2})^i)/21,(\frac{32}{63})^2/2^{2i})$
15. Discrete comb	$\sum_{i=0}^{2} \frac{2}{7}N(\frac{12i-15}{7},(\frac{2}{7})^2) + \sum_{i=8}^{10} \frac{1}{21}N(\frac{2i}{7},(\frac{1}{21})^2)$

Tabuľka A.1: Parametre hustôt 15-tich príkladov normálnej zmesi z článku Marron, Wand (1992)

Dodatok B

Špecifikácia funkcií implementovaných v jazyku R

B.1 Funkcia PEME.ic

Maximálne vierohodný odhad pre modely jednorozmerných normálnych, exponenciálnych a Poissonových zmesí, ktorý umožňuje voliť model na základe viacerých informačných kritérií.

Vstupné parametre:

- `data` - vstupné dáta, ktoré by mali byť triedy `numeric` alebo triedy `data frame`
- `family` - rodina rozdelení, možnosti: `normal`, `exp`, `poiss`, defaultne funkcia pracuje s normálnou zmesou
- `criterion` - informačné kritérium pre výber počtu komponentov zmesi, možnosti: `AIC, BIC, KER, CLC, NEC, ICL-BIC, MIR, all` a EIC[45], v prípade `all` sú vyčíslené všetky kritéria kritériá s výnimkou EIC, defaultne sa používa `BIC` kritérium
- `alpha` - vektor štyroch (rôznych) hodnôt α pre Keribinovu penalizáciu, defaultné nastavenie `alpha=c(0.5,1,2,3)`
- `k.min, k.max` - minimálna a maximálna hodnota pre počet komponentov zmesi, defaultne `k.min=2, k.max=9`
- `em.options` - list hodnôt pre nastavenie EM algoritmu
  - `klas` - booleanovská premenná, ak je nastavená na `TRUE`, tak je použitá klasifikačná alternatíva EM algoritmu

[45] I napriek použitej (zrýchlujúcej) aproximácii od Konishi a Kitagawu je výpočet v tomto prípade veľmi dlhý.

- `ini.type` - typ náhodnej inicializácie (implementované sú dva typy 1 a 2), okrem toho je možnosť zadať namiesto typu náhodnej inicializácie do tohoto parametra list konkrétnych inicializácií, ktorých počet musí odpovedať počtu opakovaní EM algoritmu,
- `s.values` - počet opakovaní EM algoritmu (pre viacero náhodných inicializácií)
- `tol` - tolerancia pre ukončenie algoritmu, pre zastavenie iterácií je použité Aitkenovo ukončovacie kritérium (McLachlan, Peel, 2000, str. 52-53))
- `maxiter` - maximálny počet iterácií

defaultne
`list(klas=FALSE, ini.type=1, s.values=10, tol=1e-5, maxiter=500)`

- `eic.options` - nastavenie hodnôt pre `EIC` kritérium:
  - `B` - počet bootstrapových výberov
  - `s.values.eic` - počet náhodných inicializácií pre EM algoritmus aplikovaný na jednotlivé bootstrapové výbery
- `prnt` - booleanovská premenná, ak je nastavená na `TRUE`, tak výsledky sa v prehľadnej podobe zobrazia v pracovnej konzole
- `plt, path, name` - `plt` je tiež booleanovská premenná, ak je nastavená na `TRUE`, tak je zostrojený graf porovnávajúci hustotu odhadnutej zmesi s jadrovým odhadom hustoty (len v prípade normálnej a exponenciálnej zmesi). Tento graf sa uloží vo formáte pdf pod menom „`name`“ + „názov kritéria“ do adresára s cestou `path`. Defaultne sa pre tieto účely používa aktuálny pracovný adresár, do ktorého sa grafický výstup uloží do poddresára `"\mix\em\"` (pokiaľ takýto neexistuje, bude vytvorený)

Výstup:

V prípade, že parameter `prnt` je nastavený na `TRUE` je výstupom z funkcie prehľadný výpis najdôležitejších výsledkov do konzoly. Okrem toho v prípade, že vstupný parameter `criterion` je nastavený na hodnotu `all`, je výstupom list hodnôt informačných kritérií pre rôzny počet komponentov (`$criterion`), potom list odhadov parametra modelu pre jednotlivé kritériá (`$theta`) a napokon ešte list odhadnutého počtu komponentov v prípade jednotlivých kritérií (`$G`). V prípade individuálneho informačného kritéria je výstupom odhad parametrov modelu (`$theta`) a hodnota tohoto informačného kritéria pre rôzny počet komponentov (`$ic`).

B.2 Funkcia MPDE

Odhad jednorozmernej normálnej zmesi metódou minimálnej vzdialenosti. Voľba modelu je možná buď stanovením hranice dostatočnej blízkosti podľa Henna (1985) alebo

penalizáciou navrhnutou v Chen, Kalbfleisch (1996). Na minimalizáciu vzdialensoti je použitý algoritmus NMSA[46].

Vstupné parametre

- `data` - vstupné dáta
- `k.min, k.max` - minimálny a maximálny počet komponentov, defaultné nastavenie je 2 a 9
- `ini` - typ inicializácie pre Nelder Meadov algoritmus, možnosti: `em` - za inicializáciu sa zoberie riešenie EM algoritmu, `emklas` - za incializáciu sa berie riešenie klasifikačnej verzie EM algoritmu, `sample` - náhodná inicializácia založená na výbere z dát, `rnorm` - náhodná inicializácia založená na výbere z normálneho rozdelenia s parametrami odvodenými z dát, defaultne sa používa nastavenie `emklas`
- `algorithm` - typ algoritmu, možnosti:
  - `Henna` - v tomto prípade je minimalizovaná CvM vzdialenosť medzi EDF a distribučnou funkciou zmesi (MCDF), za počet komponentov je zvolené najmenšie k, pre ktoré je minimalizovaná vzdialenosť menšia ako `h.delta` $* \frac{(\log \mathtt{n})^2}{\mathtt{n}}$, kde `n` je počet pozorovaní, ak pre žiadne k minimalizovaná vzdialenosť nepodlezie uvedenú hranica, za riešenie sa zoberie k odpovedajúce prípadu, v ktorom bola dosiahnutá najmenšia vzdialenosť
  - `Chen` - v tomto prípade je minimalizovaná vzdialenosť `dist` medzi EDF a MCDF penalizovaná o člen `delta` $* c_n * \sum_{i=1}^{k} \log \mathtt{p}_i$, postupnosť konštánt c_n určuje funkcia `c(n)` implementovaná v úvode skriptu, ktorá je pôvodne stanovená na hodnotu `c(n)` $= \mathtt{n}^{-1/2} \log \mathtt{n}$ (`n` rozsah dát)

  defaultne sa používa algoritmus `Chen`
- `dist` - typ vzdialenosti medzi EDF a MCDF pre prípad Chenovho algortimu, možnosti: `KS` - suprémová vzdialenosť, `Kant` - Kantorowichova vzdialensoť a `CvM` - Cramér-von-Misesova vzdialenosť, defaultne sa používa `CvM` vzdialenosť
- `delta` - konštanta pre penalizáciu v prípade algoritmu `Chen`, defaultne `delta=0.01`
- `h.delta` - konštanta pre hranicu dostatočnej blízkosti v prípade algoritmu `Henna` defaultne `h.delta=0.0001`
- `s.values` - počet opakovaní algoritmu NMSA pre rôzne inicializácie (prvá inicializácia je typu `emklas` a zvyšné sú typu `sample`), defaultne je kvôli časovej náročnosti je hodnota `s.values` nastavená len na hodnotu 1
- `maxit` - hranicu maximálneho počtu iterácií pre NMSA algoritmus, defaultná hodnota je 500

[46] Tento algoritmus je implementovaný použitím funkcie `constrOptim` zo základnej knižnice programu R.

- `prnt` - booleanovská premenná - indikátor pre vypísanie stručných výsledkov do konzoly

Výstup

Výstupom je odhad parametra a v prípade pozitívnej hodnoty `prnt` aj prehľadný výpis výsledkov (v prípade algortimu `Henna` to je aj poznámka o tom, či bola dosiahnutá hranica minimálnej blízkosti).

Funkcia MPDE.ini

V prípade známeho k možno namiesto funkcie `MPDE` používať funkciu `MPDE.ini`, ktorá obsahuje rovnaké vstupné parametre ako funkcia `MPDE` s výnimkou `k.min` a `k.max`, namiesto ktorých obsahuje len parameter `k`, ktorý vyjadruje konkrétny počet komponentov v zmesi. Táto funkcia je vlastne základom pre funkciu `MPDE`.

B.3 Funkcia MPDE.ga

Odhad jednorozmernej normálnej zmesi metódou minimálnej vzdialenosti. Ide o analógiu funkcie `MPDE`, ktorá namiesto algoritmu NMSA používa na minimalizáciu vzdialenosti genetický algoritmus.

Genetický algoritmus v krokoch:

1. Je zvolená počiatočná populácia jedincov (jedinec = vektor parametrov zmesi). Defaultný rozsah populácie je 100 jedincov.
2. Spočíta sa „fitness" každého jedinca populácie, teda hodnota vzdialenosti EDF a MCDF prislúchajúca tomuto jedincovi (do aktuálneho grafu v konzole sa nanesie hodnota priemerného a najlepšieho fitnessu celej populácie).
3. Vytvorí sa nová rovnako veľká populácia jedincov. Tá bude obsahovať niekoľko jedincov z najlepším fitnessom zo starej populácie (*elitárstvo*). Časť jedincov do novej populácie bude náhodne vygenerovaných. Ostatní jedinci novej populácie budú vytvorený buď *mutáciou* alebo *krížením* určitých jedincov z predchádzajúcej populácie, pričom na kríženie a mutovanie sú jedinci vyberaný podľa nejakého kľúča na základe fitnessu jedincov (*selekcia*).
4. Opakujú sa kroky 2. a 3. až kým algoritmus neukončíme.

Mutácia parametra $\boldsymbol{\Psi} = (\pi_1, \ldots, \pi_g, \mu_1, \ldots, \mu_g, \sigma_1^2, \ldots, \sigma_g^2)$ - zvolím vektor parametrov $\boldsymbol{r} = (r_0, r_1, \ldots, r_g, r_{g+1}, \ldots, r_{2g})$, ktorého súčet je jedna, pri mutácii potom s pravdepodobnosťou r_0 zmením proporcie, s pravdepodobnosťou r_i zmením strednú hodnotu i-teho komponentu a s pravdepodobnosťou r_{g+i} zmením rozptyl tohoto komponentu ($i = 1, \ldots, g$). Zmenu u proporcií robím tak, že jednu (náhodne zvolenú) z proporcií náhodne zmutujem a celý vektor proporcií potom znormujem, tak aby jeho súčet

bol jedna. Pri mutácii strednej hodnoty nejakého komponentu postupujem tak, že z dát náhodne vyberiem nejaké pozorovanie, ku ktorému pričítam realizáciu z rozdelenia $N(0,s^2)$, kde s^2 je predom volený parameter. V prípade mutácie rozptylu pôvodný rozptyl nahradím realizáciou z rovnomerného rozdelenia na množine (0,var(data)).

Kríženie parametrov Ψ_1, Ψ_2 - z parametrov Ψ_1, Ψ_2 vytvorím novú dvojicu Ψ_1^*, Ψ_2^* nasledovne: náhodne vyberiem dvojicu indexov i,j (i,$j \in \{1,\ldots,g\}$) a potom vytvorím nový parameter Ψ_1^* tak, že v Ψ_1 nahradím i-ty komponent j-tym komponentom z Ψ_2 (teda π_i,μ_i,σ_i nahradím s π_j^*,μ_j^*,σ_j^{2*}) a parameter Ψ_2^* vytvorím tak, že v parameteri Ψ_2 nahradím j-ty komponent i-tym komponentom z Ψ_1. Na dokončenie kríženia je ešte potrebné znormovať vektory proporcií v nových parametroch Ψ_1^*, Ψ_2^*.

Selekcia - jedincov v populácii usporiadom od jedinca s najlepším fitnessom po jedinca s najhorším, časť najhorších jedincov zahodím a selekciu jedincov realizujem použitím gama rozdelenia, pomocou ktorého pri výbere uprednostňujem jedincov s lepším fitnessom.

Vstupné parametre

Vstupné parametre sú podobné ako v prípade funkcie `MPDE`, a to `data, K.max, K.min, algorithm, dist, delta, h.delta, init, path, fname` a navyše je tu parameter `ga.options`, ktorý slúži na podrobné nastavenie genetického algoritmu. Pomocou neho napríklad nastavujeme počet iterácií (počet opakovní kroku 4.), ktorý je defaultne nastavený na 10, ďalej rozsah populácie a parametre pre kríženie, mutovanie a selekciu.

Výstup

Počas behu algoritmu je priebežne vykreslovaný priemerný a najlepší fitness jednotlivých populácií. Aktuálna populácia a zároveň aj zoznam najlepších jedincov jednotlivých populácií pre každý uvažovaný počet komponentov (medzi `K.min` a `K.max`) sa ukladajú do súborov `"\mix\ga\MPDE-last.popul_k.txt"` a `"\mix\ga\MPDE-best.solution_k.txt"`. Ak by sme chceli po prvých 10-tich iteráciách algoritmu pokračovať v ďalšom iterovaní, stačí zavolať funkciu `MPDE.ga` s parametrom `init="read"`. V tomto prípade sa načíta posledná uložená populácia a iterovanie bude pokračovať od tejto poplácie.

Nastavením parametrov pre kríženie a mutovanie (poprípade modifikáciou kríženia a mutovania, resp. ich úplnou zmenou) a nastavením ďlaších parametrov GA algoritmu možno potom vstupovať do prehľadávacieho procesu a prispôsobovať algoritmus konkrétnemu prípadu.

Literatúra

Jeffrey D. Banfield, Adrian E. Raftery. Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49**, 803–821, 1993.

E.H. Blackstone, D.C. Naftel, M.E. Turner. The decomposition of timevarying hazard into phases, each encorporating a separate stream of concomitant information. *J. Amer. Statist. Assoc.* **81**, 615–624, 1986.

N.A. Campbell. Mixture models and atypical values. *Mathematical geology* **16**, 465–477, 1984.

J. Chen. Optimal rate of convergence for finite mixture models. *Ann. Statist.* **23**, 221–234, 1995.

Jiahua Chen, J. D. Kalbfleisch. Penalized minimum-distance estimates in finite mixture models. *Canad. J. Statist.* **24**, 167–175, 1996.

D.A. Coleman, D.L. Woodruff. Cluster analysis for large dataset: Efficient alghorithms for maximizing the mixture likelihood. *Journal of Computational and Graphical Statistics* **9**, 672–688, 2000.

Harald Cramér. *Mathematical Methods of Statistics.* Princeton Mathematical Series, vol. 9. Princeton University Press, Princeton, N. J., 1946.

Adele Cutler, Olga I. Cordero-Braña. Minimum Hellinger distance estimation for finite mixture models. *J. Amer. Statist. Assoc.* **91**, 1716–1723, 1996.

D. Dacunha-Castelle, E. Gassiat. Testing the order of a model using locally conic parametrization: population mixtures and stationary ARMA processes. *Ann. Statist.* **27**, 1178–1209, 1999.

Didier Dacunha-Castelle, Elisabeth Gassiat. The estimation of the order of a mixture model. *Bernoulli* **3**, 279–299, 1997.

A.P. Dempster, N.M. Laird, D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *J. Roy. Statist. Soc. Ser. B* **39**, 1–38, 1977.

Jean Diebolt, Eddie H. S. Ip. Stochastic EM: method and application. In *Markov chain Monte Carlo in practice*, Interdiscip. Statist., 259–273. Chapman & Hall, London, 1996.

Isidore Eisenberger. Genesis of bimodal distributions. *Technometrics* **6**, 357–363, 1964.

B. S. Everitt. *An introduction to latent variable models.* Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1984.

Z.D. Feng, C.E. McCulloch. Using bootstrap likelihood ratios in finite mixture models. *J.R. Statist. Soc. B* **58**, 609–617, 1996.

N. I. Fisher, E. Mammen, J. S. Marron. Testing for multimodality. *Comput. Statist. Data Anal.* **18**, 499–512, 1994.

Li Gan, Jiming Jiang. A test for global maximum. *J. Amer. Statist. Assoc.* **94**, 847–854, 1999.

Jayanta Kumar Ghosh, Pranab Kumar Sen. On the asymptotic performance of the log likelihood ratio statistic for the mixture model and related results. In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer, Vol. II (Berkeley, Calif., 1983)*, Wadsworth Statist./Probab. Ser., 789–806, Belmont, CA, 1985. Wadsworth.

J. A. Hartigan. A failure of likelihood asymptotics for normal mixtures. In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer, Vol. II (Berkeley, Calif., 1983)*, Wadsworth Statist./Probab. Ser., 807–810, Belmont, CA, 1985. Wadsworth.

J. A. Hartigan, P. M. Hartigan. The dip test of unimodality. *Ann. Statist.* **13**, 70–84, 1985.

R. J. Hathaway. A constrained formulation of maximum-likelihood estimation for normal mixture distribution. *Ann. Statist.* **13**, 795–800, 1985.

James J. Heckman, Richard Robb, James R. Walker. Testing the mixture of exponentials hypothesis and estimating the mixing distribution by the method of moments. *J. Amer. Statist. Assoc.* **85**, 1990.

Jōgi Henna. On estimating of the number of constituents of a finite mixture of continuous distributions. *Ann. Inst. Statist. Math.* **37**, 235–240, 1985.

C. Hennig. Clustering and outlier identification: fixed point cluster analysis. In *Advances in data science and classification (Rome, 1998)*, Stud. Classification Data Anal. Knowledge Organ., 37–42. Springer, Berlin, 1998.

Christian Hennig, Norbert Christlieb. Validating visual clusters in large datasets: fixed point clusters of spectral features. *Comput. Statist. Data Anal.* **40**, 723–739, 2002.

Makio Ishiguro, Yosiyuki Sakamoto, Genshiro Kitagawa. Bootstrapping log likelihood and EIC, an extension of AIC. *Ann. Inst. Statist. Math.* **49**, 411–434, 1997.

Lancelot F. James, Carey E. Priebe, David J. Marchette. Consistent estimation of mixture complexity. *Ann. Statist.* **29**, 1281–1296, 2001.

Norman L. Johnson, Samuel Kotz. *Distributions in statistics: Discrete distributions.* Houghton Mifflin Co., Boston, Mass., 1969.

M. Jorgensen, L. Hunt. Mixture model clustering of data sets with categorical and continuous variables. In *Proceedings of the Conference ISIS '96*, 375–384, Australia, 1996.

Evdokia Karlis, Dimitris a Xekalaki. Choosing initial values for the EM algorithm for finite mixtures. *Comput. Statist. Data Anal.* **41**, 577–590, 2003. Recent developments in mixture models (Hamburg, 2001).

R.E. Kass, A.E. Raftery. Bayes factors. *J. Amer. Statist. Assoc.* **90**, 773–795, 1995.

C. Keribin. Consistent estimation of the order of mixture models. *Sankhyā Ser. A* **62**, 49–66, 2000.

J. Kiefer, J. Wolfowitz. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.* **27**, 887–906, 1956.

Sadanori Konishi, Genshiro Kitagawa. Generalised information criteria in model selection. *Biometrika* **83**, 875–890, 1996.

Nan Laird. Nonparametric maximum likelihood estimation of a mixed distribution. *J. Amer. Statist. Assoc.* **73**, 805–811, 1978.

C.J. Lawrence, W.J. Krzanowski. Mixture separation for mixed-mode data. *Statistics and Computing* **6**, 85–92, 1996.

Brian G. Leroux. Consistent estimation of a mixing distribution. *Ann. Statist.* **20**, 1350–1360, 1992.

Bruce G. Lindsay. The geometry of mixture likelihoods: a general theory. *Ann. Statist.* **11**, 86–94, 1983.

Bruce G. Lindsay. Moment matrices: Application in mixtures. *Ann. Statist.* **17**, 722–740, 1989.

Bruce G. Lindsay, Prasanta Basak. Multivariate normal mixtures: a fast consistent method of moments. *J. Amer. Statist. Assoc.* **88**, 468–476, 1993.

Bruce G. Lindsay, David W. Furman. Measuring the relative effectiveness of moment estimators as starting values in maximizing likelihoods. *Comput. Statist. Data Anal.* **17**, 493–507, 1994.

D.J. Marchette, C.E. Priebe, G.W. Rogers, J.L. Solka. Filtered kernel density estimation. *Comput. Statist.* **11**, 95–112, 1996.

M. Markatou. Mixture models, robustness and weighted likelihood methodology. *Technical Report No 1998-9. Stanford: Department of Statistics, Stanford University.*, 1998.

J. S. Marron, M. P. Wand. Exact mean integrated squared error. *Ann. Statist.* **20**, 712–736, 1992.

G. J. McLachlan. On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Applied Statistics* **36**, 318–324, 1987.

Geoffrey McLachlan, David Peel. *Finite mixture models.* Wiley Series in Probability and Statistics: Applied Probability and Statistics. Wiley-Interscience, New York, 2000.

Geoffrey J. McLachlan, David Peel. Robust cluster analysis via mixtures of multivariate t-distributions. In *Advances in pattern recognition (Sydney, 1998)* **1451** of *Lecture Notes in Comput. Sci.*, 658–666. Springer, Berlin, 1998.

X.L. Meng, D.V. Dyk. The em algorithm - an old folk-song sung to a fast new tune. *J. Roy. Statist. Soc. Ser. B* **59**, 511–567, 1997.

Ryohei Nakano Naonori Ueda. Split and merge EM algorithm for improving gaussian mixture density estimates. *The Journal of VLSI Signal Processing* **26**, 133–140, 2000.

Jan Nemeček. Klasifikace a rozpoznávaní. Diplomová práca, Univerzita Karlova v Praze, Matematicko-fyzikálna fakulta, 2004.

David Peel, William J. Whiten, Geoffrey J. McLachlan. Fitting mixtures of Kent distributions to aid in joint set identification. *J. Amer. Statist. Assoc.* **96**, 56–63, 2001.

B. Charles Peters, Jr., Homer F. Walker. An iterative procedure for obtaining maximum-likelihood estimates of the parameters for a mixture of normal distributions. *SIAM J. Appl. Math.* **35**, 362–378, 1978.

D.B. Phillips, A.F.M. Smith. *Practical Markov chain Monte Carlo*, kapitola Bayesian model comparison via jump difusions, 215–239. London, 1996.

A. Polynemis, D.M. Titterington. On the determination of the number of components in a mixture. *Statistics and Computing* **38**, 295–298, 1998.

Carey E. Priebe, David J. Marchette. Alternating kernel and mixture density estimates. *Comput. Statist. Data Anal.* **35**, 43–65, 2000.

Adrian E. Raftery. Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika* **83**, 251–266, 1996.

R. Redner. Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions. *Ann. Statist.* **9**, 225–228, 1981.

Richard A. Redner, Homer F. Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.* **26**, 195–239, 1984.

Sylvia Richardson, Peter J. Green. On Bayesian analysis of mixtures with an unknown number of components. *J. Roy. Statist. Soc. Ser. B* **59**, 731–792, 1997a.

Sylvia Richardson, Peter J. Green. On Bayesian analysis of mixtures with an unknown number of components. *J. Roy. Statist. Soc. Ser. B* **59**, 731–792, 1997b.

Kathryn Roeder. A graphical technique for determining the number of components in a mixture of normals. *J. Amer. Statist. Assoc.* **89**, 487–495, 1994.

Kathryn Roeder, Larry Wasserman. Practical Bayesian density estimation using mixtures of normals. *J. Amer. Statist. Assoc.* **92**, 894–902, 1997.

Peter Schlattmann. Estimating the number of components in a finite mixture model: the special case of homogeneity. *Comput. Statist. Data Anal.* **41**, 441–451, 2003.

N.J. Schork, A.B. Weder, A. Schork. On the asymmetry of biological frequency distributions. *Genetic Epidemiology* **7**, 427–446, 1990.

G. Schwarz. Estimating the dimension of a model. *Ann. Statist.* **6**, 461–464, 1978.

Léopold Simar. Maximum likelihood estimation of a compound Poisson process. *Ann. Statist.* **4**, 1200–1209, 1976.

P. Smyth. Model selection for probabilistic clustering using cross-validated likelihood. *Statistics and Computing* **10**, 63–72, 2000.

R. Sundberg. Maximum likelihood theory for incomplete data from an exponential family. *Scand. J. Statist.* **1**, 49–58, 1974.

Edward Susko. Weighted tests of homogeneity for testing the number of components in a mixture. *Comput. Statist. Data Anal.* **41**, 367–378, 2003. Recent developments in mixture models (Hamburg, 2001).

D. M. Titterington, A. F. M. Smith, U. E. Makov. *Statistical analysis of finite mixture distributions.* Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Ltd., Chichester, 1985.

Abraham Wald. Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statistics* **20**, 595–601, 1949.

H. X. Wang, B. Luo, Q. B. Zhang, S. Wei. Estimation for the number of components in a mixture model using stepwise split-and-merge em algorithm. *Pattern Recogn. Lett.* **25**, 1799–1809, 2004.

M.P. Windham, A. Cutler. Information ratios for validating mixture analyses. *J. Amer. Statist. Assoc.* **87**, 1188–1192, 1992.