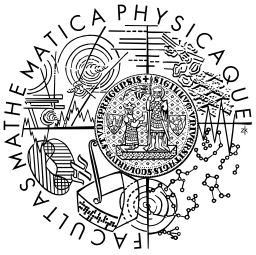


MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

BAKALÁŘSKÁ PRÁCE

Lucie Janečková

Klasifikace založená na směsových modelech

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: doc. RNDr. Arnošt Komárek, Ph.D.

Studijní program: Obecná matematika

Praha 2023

Prohlašuji, že jsem tuto bakalářskou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů. Tato práce nebyla využita k získání jiného nebo stejného titulu.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ....

Podpis autora

Zde bych ráda poděkovala mému vedoucímu práce, doc. RNDr. Arnoštu Komárkovi, Ph.D., za cenné rady a čas, který mi při psaní práce věnoval.

Název práce: Klasifikace založená na směsových modelech

Autor: Lucie Janečková

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: doc. RNDr. Arnošt Komárek, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Tato práce se zabývá klasifikací založenou na směsových modelech, a to převážně na modelech konečných normálních. Nejprve jsou zavedeny základní definice a vlastnosti konečných směsí. Následně je zde popsána metoda maximální věrohodnosti a její úskalí v kontextu konečných směsí, kterou používáme pro odhadování neznámých parametrů. Poté je popsán EM algoritmus, který je používán pro získání maximálně věrohodných odhadů a explicitně spočteny vzorce pro jednu iteraci EM algoritmu. V poslední části je ukázáno, jak lze konečné normální směsi využít ke klasifikaci.

Klíčová slova: konečná směs, normální směs, klasifikace, EM algoritmus, věrohodnost

Title: Classification based on mixture models

Author: Lucie Janečková

Department: Department of Probability and Mathematical Statistics

Supervisor: doc. RNDr. Arnošt Komárek, Ph.D., Department of Probability and Mathematical Statistics

Abstract: This thesis deals with classification based on mixture models, mainly on models finite normal. At first, there are introduced basic definitions and characteristics of finite mixtures. Afterwards there is described the maximum likelihood method and her obstacles in context of finite mixtures, which we are using for unknown parameters estimation. Then there is described EM algorithm, that is used to obtain the maximum likelihood estimator and there are calculated the formulae for one iteration of EM algorithm. In the last part there is shown, how can finite normal mixtures be used for classification.

Keywords: finite mixture, normal mixture, classification, EM algorithm, likelihood

Obsah

Úvod

Uvažujme situaci, kdy máme populaci rozdělenou do předem známého počtu skupin, ale nevíme, který jedinec patří do které skupiny. V každé skupině pozorujeme jiné vzorce chování. Naším cílem tedy bude každého jedince z populace zařadit do příslušné skupiny na základě jeho pozorovaného vzorce chování.

Tuto situaci modelujeme právě pomocí konečné směsi, kterou si v první kapitole zavedeme. Dále se také zmíníme o její možné interpretaci a matematicky popíšeme výše zmíněný cíl práce.

Pro správné zařazení jedinců do skupin budeme potřebovat odhadnout neznámé parametry. K tomu nám poslouží metoda maximální věrohodnosti, kterou popíšeme v druhé kapitole, včetně problémů, které nám použití této metody v kontextu konečných směsí může činit.

Ke konečnému nalezení odhadů použijeme EM algoritmus, který ve třetí kapitole popíšeme a vypočteme konkrétní vzorce pro odhady neznámých parametrů v normální směsi. Teoretická část práce vychází z knihy McLachlan a Peel (2000).

V poslední kapitole pak využijeme nabytých znalostí pro analýzu archeologických dat. Konkrétně z největších délek mozkovny a týlních úhlů budeme chtít roztřídit lebky podle místa nálezu. K tomu nám poslouží právě EM algoritmus.

1. Konečná směs

Mějme populaci sestávající z $g \in \mathbb{N}$ skupin, tu se budeme snažit popsat pomocí konečné, resp. normální směsi, což jsou pojmy, které si v této kapitole zavedeme. Dále si uvedeme jejich základní vlastnosti a popíšeme, jak lze směsi interpretovat a jak se dají použít ke klasifikaci.

1.1 Základní definice

Definice 1. *Nechť $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ je náhodný výběr p-rozměrných náhodných vektorů o rozsahu $n \in \mathbb{N}$, kde $p \in \mathbb{N}$. Nechť pro $k \in \{1, \ldots, g\}$ jsou f_k libovolné hustoty vzhledem k p-rozměrné Lebesgueově míře na $\mathbb{R}^p$, které odpovídají rozdělení se střední hodnotou $\boldsymbol{\mu_k} \in \mathbb{R}^p$ a varianční maticí $\boldsymbol{\Sigma_k} \in \mathbb{R}^{p\times p}$. Pokud má náhodný vektor $\boldsymbol{Y}_j$ rozdělení s hustotou danou*

$$f(\boldsymbol{y}) = \sum_{k=1}^{g} w_k f_k(\boldsymbol{y}), \quad \boldsymbol{y} \in \mathbb{R}^p, \tag{1.1}$$

vzledem k Lebesgueově míře na $\mathbb{R}^p$, kde

$$w_k \in (0,1), \quad \forall k = 1, \ldots, g$$

a

$$\sum_{k=1}^{g} w_k = 1,$$

potom říkáme, že $\boldsymbol{Y}_j$ má rozdělení, které nazýváme **konečná směs o g-složkách**. *Hustotu danou vztahem* (1.1) *nazýváme* **hustota směsi o g-složkách**. *Čísla w_k nazýváme* **váhy** *a hustoty f_k* **komponenty směsi**.

Nabízí se otázka, proč uvažujeme váhy v otevřeném intervalu $(0, 1)$ a ne v uzavřeném. Krajní hodnoty jsou totiž pro nás nezajímavé. Kdybychom uvažovali hodnotu 0 u některé z komponent, pak bychom pracovali s $g - 1$ složkovou směsí. Kdybychom uvažovali hodnotu 1, dostali bychom triviální jednosložkovou směs.

Tuto definici můžeme vyložit tak, že celková hustota populace $\boldsymbol{Y}_j$ je $f(\boldsymbol{y})$, hustoty jednotlivých skupin jsou $f_k(\boldsymbol{y})$ a poměrné zastoupení v jednotlivých skupinách jsou právě váhy w_k.

V následující větě určíme střední hodnotu a rozptyl náhodné veličiny s hustotou konečné směsi.

Věta 1. *Nechť $\boldsymbol{Y}$ je náhodný vektor s hustotou danou vzorcem* (1.1). *Označme $\boldsymbol{\Gamma_k}$ matici druhých momentů k-té složky. Pak platí*

$$E(\boldsymbol{Y}) = \sum_{k=1}^{g} w_k \boldsymbol{\mu_k},$$

$$var(\boldsymbol{Y}) = \sum_{k=1}^{g} w_k \boldsymbol{\Gamma_k} - (\sum_{k=1}^{g} w_k \boldsymbol{\mu_k})(\sum_{k=1}^{g} w_k \boldsymbol{\mu_k})^\top.$$

Důkaz. Pro $\boldsymbol{y} \in \mathbb{R}^p$ počítejme

$$\mathsf{E}(\boldsymbol{Y}) = \int_{\mathbb{R}^p} \boldsymbol{y} \sum_{k=1}^{g} w_k f_k(\boldsymbol{y}) \,\mathrm{d}\boldsymbol{y} = \sum_{k=1}^{g} w_k \int_{\mathbb{R}^p} \boldsymbol{y} f_k(\boldsymbol{y}) \,\mathrm{d}\boldsymbol{y} = \sum_{k=1}^{g} w_k \boldsymbol{\mu_k},$$

$$\mathsf{E}(\boldsymbol{Y}\boldsymbol{Y}^\top) = \int_{\mathbb{R}^p} \boldsymbol{y}\boldsymbol{y}^\top \sum_{k=1}^{g} w_k f_k(\boldsymbol{y}) \,\mathrm{d}\boldsymbol{y} = \sum_{k=1}^{g} w_k \int_{\mathbb{R}^p} \boldsymbol{y}\boldsymbol{y}^\top f_k(\boldsymbol{y}) \,\mathrm{d}\boldsymbol{y} = \sum_{k=1}^{g} w_k \boldsymbol{\Gamma_k},$$

kde používáme linearitu integrálu.

Tedy celkem pro varianční matici platí

$$\mathrm{var}(\boldsymbol{Y}) = \mathsf{E}(\boldsymbol{Y}\boldsymbol{Y}^\top) - (\mathsf{E}(\boldsymbol{Y}))(\mathsf{E}(\boldsymbol{Y}))^\top = \sum_{k=1}^{g} w_k \boldsymbol{\Gamma_k} - (\sum_{k=1}^{g} w_k \boldsymbol{\mu_k})(\sum_{k=1}^{g} w_k \boldsymbol{\mu_k})^\top.$$

□

Pokud v definici 1 vezmeme za f_k hustoty normálního rozdělení, můžeme definovat normální směs.

Definice 2. *Nechť $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ je náhodný výběr p-rozměrných náhodných vektorů s normálním rozdělením o rozsahu $n \in \mathbb{N}$, kde $p \in \mathbb{N}$. Nechť pro $k \in \{1, \ldots, g\}$ jsou $\phi_k(\boldsymbol{y}; \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})$ hustoty normálního rozdělení se střední hodnotou $\boldsymbol{\mu_k} \in \mathbb{R}^p$ a varianční maticí $\boldsymbol{\Sigma_k} \in \mathbb{R}^{p \times p}$. Označme $\boldsymbol{\psi} \equiv (w_1, \ldots, w_g, \boldsymbol{\mu_1}, \ldots, \boldsymbol{\mu_g}, \boldsymbol{\Sigma_1}, \ldots, \boldsymbol{\Sigma_g})$ sadu neznámých parametrů. Pokud má náhodný vektor $\boldsymbol{Y}_j$ rozdělení s hustotou danou*

$$\phi(\boldsymbol{y}; \boldsymbol{\psi}) = \sum_{k=1}^{g} w_k \phi_k(\boldsymbol{y}; \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}), \quad \boldsymbol{y} \in \mathbb{R}^p, \tag{1.2}$$

vzledem k Lebesgueově míře na $\mathbb{R}^p$, kde

$$w_k \in (0{,}1), \quad \forall k = 1, \ldots, g$$

a

$$\sum_{k=1}^{g} w_k = 1,$$

potom říkáme, že $\boldsymbol{Y}_j$ má rozdělení, které nazýváme **konečná normální směs o g-složkách**. *Hustotu danou* (1.2) *nazýváme* **hustota normální směsi o g-složkách**. *Čísla w_k nazýváme* **váhy** *a hustoty $\phi_k(\boldsymbol{y}; \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})$* **komponenty normální směsi**.

Dále se v práci budeme zabývat hlavně modelem homoskedastické normální směsi, tedy směsi, kde jsou všechny varianční matice $\boldsymbol{\Sigma_k}$ stejné. Směsi heteroskedastické (směsi s různými variančními maticemi $\boldsymbol{\Sigma_k}$) nám totiž mohou dělat problémy při odhadování parametrů pomocí metody maximální věrohodnosti, jak si ukážeme v kapitole 2.

1.2 Interpretace konečné směsi a klasifikace

Zde si uvedeme dvě možnosti, jak lze konečné směsi interpretovat. První z nich je pomocí náhodného vektoru, který nám identifikuje, z které komponenty směsi náš náhodný vektor pochází. Druhá nám popisuje, že chování našeho náhodného vektoru v celkové populaci lze popsat pomocí jeho chování v různých skupinách.

Jednou možností, jak konečnou směs interpretovat, je pomocí náhodného vektoru $\boldsymbol{Z}_j = (Z_{1j}, \ldots, Z_{gj})^\top$ s multinomickým rozdělením, kde

$$Z_{kj} = \begin{cases} 1, & \text{pokud náhodný vektor } \boldsymbol{Y}_j \text{ náleží do } k\text{-té komponenty směsi,} \\ 0, & \text{jinak.} \end{cases}$$

Jinými slovy

$$\boldsymbol{Z}_j \sim Mult_g(1, \boldsymbol{w}),$$

kde $\boldsymbol{w} = (w_1, \ldots, w_g)^\top$ je vektor vah. Náhodný vektor $\boldsymbol{Z}_j$ tedy určuje, z jaké komponenty směsi náhodný vektor $\boldsymbol{Y}_j$ pochází.

Místo toho lze k interpretaci použít náhodné veličiny $U_j, j = 1, \ldots, n$, které nabývají hodnot $1, \ldots, g$ s pravděpodobnostmi

$$P(U_j = k) = w_k, \quad k = 1, \ldots, g.$$

Pokud podmíněnou hustotu náhodného vektoru $\boldsymbol{Y}_j$ při jevu $U_j = k$ označíme f_k, pak pro nepodmíněnou hustotu $\boldsymbol{Y}_j$ z věty o úplné pravděpodobnosti platí

$$f(\boldsymbol{y}_j) = \sum_{k=1}^{g} f(\boldsymbol{y}_j | U_j = k) P(U_j = k) = \sum_{k=1}^{g} f_k(\boldsymbol{y}_j) w_k,$$

což odpovídá vztahu (1.1).

Nechť $\boldsymbol{y}_j$ je napozorovaná hodnota náhodného vektoru $\boldsymbol{Y}_j$. Označme $\tau_k(\boldsymbol{y}_j)$ pravděpodobnost, že $\boldsymbol{Y}_j$ náleží do k-té komponenty směsi. Pro $\tau_k(\boldsymbol{y}_j)$ pak platí

$$\tau_k(\boldsymbol{y}_j) := \frac{w_k f_k(\boldsymbol{y}_j)}{\sum_{h=1}^{g} w_h f_h(\boldsymbol{y}_j)} = P(Z_{kj} = 1 | \boldsymbol{Y}_j = \boldsymbol{y}_j) = \frac{P(U_j = k) f_k(\boldsymbol{y}_j)}{\sum_{h=1}^{g} P(U_j = h) f_h(\boldsymbol{y}_j)}, \tag{1.3}$$

což vychází z Bayesovy věty.

V samostatné klasifikaci právě tyto hodnoty $\tau_k(\boldsymbol{y}_j)$ hrají klíčovou roli. Ta probíhá následovně:

- Vezmeme pozorovanou hodnotu $\boldsymbol{y}_j$ a spočteme pro ni všechny pravděpodobnosti $\tau_k(\boldsymbol{y}_j)$, $k = 1, \ldots, g$.
- Spočteme $\arg\max_{k=1,\ldots,g} \tau_k(\boldsymbol{y}_j)$.
- Zařadíme pozorování $\boldsymbol{y}_j$ do této skupiny.

Jinými slovy určíme predikce hodnot náhodných veličin $Z_{kj}, \quad k = 1, \ldots, g,$ $j = 1, \ldots, n$, neboli $\hat{Z}_{kj} = \begin{cases} 1, & \text{pokud} \quad \arg\max_{k=1,\ldots,g} \tau_k(\boldsymbol{y}_j) = k, \\ 0, & \text{jinak.} \end{cases}$

Mohlo by se stát, že by se největší pravděpodobnosti $\tau_k(\boldsymbol{y}_j)$ mohli rovnat. Potom bychom vybrali libovolnou z nich.

2. Metoda maximální věrohodnosti

Pro klasifikaci napozorovaných subjektů z populace popsanou v první kapitole potřebujeme odhadnout neznámé parametry. Jeden z nějčastěji používaných způsobů hledání odhadů parametrů je metoda maximální věrohodnosti, jejíž použití pro model konečné normální směsi si v této kapitole popíšeme. Pak si ukážeme, kdy je věrohodnostní, resp. log-věrohodnostní funkce omezená a kdy nikoli.

Nechť $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ je náhodný výběr z rozdělení s hustotou (1.2). Označme $\boldsymbol{\psi} \equiv (w_1, \ldots, w_g, \boldsymbol{\mu_1}, \ldots, \boldsymbol{\mu_g}, \boldsymbol{\Sigma_1}, \ldots, \boldsymbol{\Sigma_g})$ sadu neznámých parametrů, kde w_g lze vynechat, neboť jej můžeme dopočítat z ostatních hodnot jako $1 - \sum_{k=1}^{g-1} w_k$. Sestavíme věrohodnostní funkci pro $\boldsymbol{\psi} \in \boldsymbol{\Psi}$:

$$L(\boldsymbol{\psi}) = \prod_{j=1}^{n} \phi(\boldsymbol{Y}_j; \boldsymbol{\psi}) = \prod_{j=1}^{n} \left(\sum_{k=1}^{g} w_k \phi_k(\boldsymbol{Y}_j; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right).$$

Obdobně sestavíme i log–věrohodnostní funkci pro $\boldsymbol{\psi}$:

$$\begin{aligned} \ell(\boldsymbol{\psi}) = \log(L(\boldsymbol{\psi})) &= \log \left(\prod_{j=1}^{n} \left(\sum_{k=1}^{g} w_k \phi_k(\boldsymbol{Y}_j; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right) \right) \\ &= \sum_{j=1}^{n} \left(\log \left(\sum_{k=1}^{g} w_k \phi_k(\boldsymbol{Y}_j; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right) \right). \end{aligned}$$

Odhad $\boldsymbol{\psi}$ metodou maximální věrohodnosti označíme jako $\hat{\boldsymbol{\psi}}$, pokud existuje. Nalezneme ho jako

$$\hat{\boldsymbol{\psi}} = \arg\max_{\boldsymbol{\psi} \in \Psi} L(\boldsymbol{\psi}) = \arg\max_{\boldsymbol{\psi} \in \Psi} \ell(\boldsymbol{\psi}).$$

2.1 Omezenost věrohodnostní funkce

Pro existenci $\hat{\boldsymbol{\psi}}$ potřebujeme, aby věrohodnostní, resp. log–věrohodnostní funkce byla omezená. To ale, jak ukazuje následující tvrzení, není vždy splněno.

Tvrzení 2. *Nechť $Y_1, \ldots, Y_n$ je náhodný výběr z rozdělení s hustotou jednorozměrné normální směsi o dvou složkách, tedy*

$$\phi(y; \boldsymbol{\psi}) = w\, \phi_1(y; \mu_1, \sigma_1^2) + (1 - w)\, \phi_2(y; \mu_2, \sigma_2^2),$$

kde $\boldsymbol{\psi} = (w, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)^\top$, $y \in \mathbb{R}$, $w \in (0, 1)$, $\mu_1, \mu_2 \in \mathbb{R}$, $\sigma_1^2, \sigma_2^2 \in (0, \infty)$. Označme $y_1, \ldots, y_n$ realizace $Y_1, \ldots, Y_n$. Pokud $\mu_1 = y_1$ a $\sigma_1^2 \to 0$, tak je log–věrohodnostní funkce neomezená.

Důkaz. Sestavíme log–věrohodnostní funkci pro $\boldsymbol{\psi}$:

$$\ell(\boldsymbol{\psi}) = \sum_{j=1}^{n} \log \left(w \phi_1(y_j; \mu_1, \sigma_1^2) + (1 - w) \phi_2(y_j; \mu_2, \sigma_2^2) \right), \tag{2.1}$$

kde

$$\phi_k(y_j;\mu_k,\sigma_k^2) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(y_j-\mu_k)^2}{2\sigma_k^2}\right), \quad k = 1, 2.$$

Dosadíme do (2.1) $\mu_1 = y_1$ a dostáváme

$$\ell(\boldsymbol{\psi}) = \sum_{j=1}^{n} \log\left(w\frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{(y_j-y_1)^2}{2\sigma_1^2}\right) + (1-w)\frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{(y_j-\mu_2)^2}{2\sigma_2^2}\right)\right).$$

Chceme $\ell(\boldsymbol{\psi})$ zespoda omezit tak, aby nám vyšel výraz tvaru

$$-\log(\sigma_1^2) + Z,$$

kde Z je funkce nezávisející na σ_1^2. Pak totiž

$$\lim_{\sigma_1^2 \to 0} \left(-\log(\sigma_1^2) + Z\right) = \infty,$$

což chceme dokázat.

V argumentu logaritmu vezmeme pro $j = 1$ pouze první člen a pro $j = 2, \ldots, n$ druhý člen, který na σ_1^2 nezávisí:

$$\ell(\boldsymbol{\psi}) > \log\left(w\frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{(y_1-y_1)^2}{2\sigma_1^2}\right)\right) + \sum_{i=2}^{n} \log\left((1-w)\frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{(y_i-\mu_2)^2}{2\sigma_2^2}\right)\right).$$

Označíme

$$Z := \sum_{i=2}^{n} \log\left((1-w)\frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{(y_i-\mu_2)^2}{2\sigma_2^2}\right)\right),$$

pak máme

$$\begin{aligned}\ell(\boldsymbol{\psi}) &> \log\left(w\frac{1}{\sqrt{2\pi\sigma_1^2}} \exp(0)\right) + Z \\ &= \left(\log\left(\frac{\sqrt{2\pi\sigma_1^2}}{w}\right)\right)^{-1} + Z \\ &= -\log\left(\frac{\sqrt{2\pi\sigma_1^2}}{w}\right) + Z,\end{aligned}$$

což je požadovaný tvar.

□

Naopak další věta ukazuje, že věrohodnostní funkce homoskedatické normální směsi, tedy směsi, kde všechny její složky mají stejný rozptyl, omezená je.

Věta 3. *Nechť $Y_1, \ldots, Y_n$ je náhodný výběr z rozdělení s hustotou g-složkové homoskedatické normální směsi, kde $g < n$, tedy s hustotou*

$$\phi(y; \boldsymbol{\psi}) = \sum_{k=1}^{g} w_k \phi_k(y; \mu_k, \sigma^2),$$

kde $\boldsymbol{\psi} = (w_1, \ldots, w_g, \mu_1, \ldots, \mu_g, \sigma^2)^\top$, $y \in \mathbb{R}$, $w_k \in (0,1)$, $\sum_{k=1}^{g} w_k = 1$, $\mu_k \in \mathbb{R}$, $k = 1, \ldots, g$, $\sigma^2 \in (0, \infty)$. Označme $y_1, \ldots, y_n$ realizace $Y_1, \ldots, Y_n$. Pak je věrohodnostní funkce shora omezená.

Důkaz. Sestavíme věrohodnostní funkci pro $\boldsymbol{\psi}$:

$$L(\boldsymbol{\psi}) = \prod_{j=1}^{n} \left(\sum_{k=1}^{g} w_k \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_j - \mu_k)^2}{2\sigma^2} \right) \right).$$

Platí:

$$L(\boldsymbol{\psi}) < \prod_{j=1}^{n} \left(\sum_{k=1}^{g} \frac{1}{\sigma} \exp\left(-\frac{(y_j - \mu_k)^2}{2\sigma^2} \right) \right).$$

Z předpokladů věty máme více pozorování než je složek směsi, a tedy:

$$\exists \delta > 0 \quad \exists n_0 \in 1, \ldots, n \,:\, \forall k \in 1, \ldots, g \,:\, |y_{n_0} - \mu_k| > \delta.$$

Upravíme zvlášť první člen součinu a zbylých $n-1$ členů.
Bez újmy na obecnosti nechť $n_0 = 1$ (jinak bychom upravovali jiný člen zvlášť).

$$\sum_{k=1}^{g} \frac{1}{\sigma} \exp\left(-\frac{(y_1 - \mu_k)^2}{2\sigma^2} \right) < \sum_{k=1}^{g} \frac{1}{\sigma} \exp\left(-\frac{\delta^2}{2\sigma^2} \right) = \frac{g}{\sigma} \exp\left(-\frac{\delta^2}{2\sigma^2} \right),$$

$$\prod_{j=2}^{n} \left(\sum_{k=1}^{g} \frac{1}{\sigma} \exp\left(-\frac{(y_1 - \mu_k)^2}{2\sigma^2} \right) \right) < \prod_{j=2}^{n} \left(\sum_{k=1}^{g} \frac{1}{\sigma} \exp\left(-\frac{0^2}{2\sigma^2} \right) \right) = \left(\frac{g}{\sigma} \right)^{n-1}.$$

Celkem tedy

$$L(\boldsymbol{\psi}) < \left(\frac{g}{\sigma} \right)^{n} \exp\left(-\frac{\delta^2}{2\sigma^2} \right).$$

A platí

$$\lim_{\sigma^2 \to \infty} \left(\frac{g}{\sigma} \right)^{n} \exp\left(-\frac{\delta^2}{2\sigma^2} \right) = 0,$$

$$\lim_{\sigma^2 \to 0} \left(\frac{g}{\sigma} \right)^{n} \exp\left(-\frac{\delta^2}{2\sigma^2} \right) = 0.$$

Z toho plyne, že věrohodnostní funkce je shora omezená.

□

2.2 Další vlastnosti

Odhad metodou maximální věrohodnosti můžeme také dostat jako řešení věrohodnostních rovnic, tedy

$$\frac{\partial L(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}} = \mathbf{0},$$

či ekvivalentně

$$\frac{\partial \ell(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}} = \mathbf{0},$$

a tak také budeme v kapitole 3 postupovat.

Zbývá nám odpovědět na otázku, zda existuje maximum pouze jedno. Jednou z vlastností konečných směsí je to, že je invariantní na „přečíslování“ složek směsi. Tedy pokud máme g-složkovou konečnou směs, existuje $g!$ permutací indexů složek, neboli $g!$ sad neznámých parametrů $\boldsymbol{\psi}$, které reprezentují stejné rozdělení, a tedy máme nejméně $g!$ (stejně hodnotných) maxim (to budeme nazývat sada maxim).

I přesto ale maxim, respektive sad maxim, je většinou více. Samotný EM algoritmus, který budeme pro odhad maxim v práci používat, zavedený v kapitole 3, se používá k nalezení ne nutně globálního maxima. Přesto se s tímto výsledkem spokojíme, neboť lze ukázat, že vlastnosti těchto lokálních maxim jsou velmi podobné.

3. EM algoritmus

EM algoritmus je jednou z možností hledání maximálně věrohodného odhadu. Pochází z anglického Expectation-Maximization. Je to iterativní algoritmus, který postupně optimalizuje parametry modelu na základě pozorovaných dat. Dělí se na dva kroky: E-krok (výpočet střední hodnoty) a M-krok (maximalizace). V E-kroku se vypočítávají pravděpodobnosti příslušnosti každého pozorování k jednotlivým komponentám. V M-kroku se pak aktualizují parametry komponent na základě těchto pravděpodobností. Tento algoritmus byl poprvé popsán v práci Dempstera, Lairdové a Rubina (Dempster, Laird a Rubin, 1977).

V této kapitole si nejprve představíme problém nekompletních dat, poté si řekneme pár potřebných vlastností věrohodnostní funkce kompletních dat a spočteme odhady neznámých parametrů homoskedatické normální směsi. Nakonec si shrneme, jak EM algoritmus funguje.

3.1 Nekompletní data

Uvažujme náhodný výběr $\boldsymbol{Y}_c = \left(\begin{pmatrix} \boldsymbol{Y}_1 \\ \mathbf{Z}_1 \end{pmatrix}, \ldots, \begin{pmatrix} \boldsymbol{Y}_n \\ \mathbf{Z}_n \end{pmatrix} \right)^\top$, kde $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)^\top$ je náhodný výběr z rozdělení s hustotou homoskedastické normální směsi a $\boldsymbol{Z} = (\boldsymbol{Z}_1^\top, \ldots, \boldsymbol{Z}_n^\top)^\top$ je tedy náhodný vektor určující, z jaké komponenty směsi náhodné vektory $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ pochází popsaný v kapitole 1.2. Tento náhodný výběr nazýváme kompletní data. Náhodný výběr $(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)^\top$ nazveme nekompletní data.

Mějme $\boldsymbol{y}_c = \left(\begin{pmatrix} \boldsymbol{y}_1 \\ \mathbf{z}_1 \end{pmatrix}, \ldots, \begin{pmatrix} \boldsymbol{y}_n \\ \mathbf{z}_n \end{pmatrix} \right)^\top$ potenciálně napozorovaný kompletní náhodný výběr, kde $\boldsymbol{z}_j = (z_{1j}, \ldots, z_{gj})^\top$. Dále označme $\boldsymbol{\xi} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_g, \boldsymbol{\Sigma})$ sadu vybraných neznámých parametrů a $\boldsymbol{\psi} = (w_1, \ldots, w_g, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_g, \boldsymbol{\Sigma})$ sadu všech neznámých parametrů.

Sestavíme log-věrohodnostní funkci pro $\boldsymbol{\psi} = (w_1, \ldots, w_g, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_g, \boldsymbol{\Sigma})$ a náhodný výběr $Y_1, \ldots, Y_n$

$$\ell(\boldsymbol{\psi}) = \sum_{j=1}^{n} \log \left(\sum_{k=1}^{g} w_k \phi_k(\boldsymbol{Y}_j; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}) \right).$$

Maximalizace $\ell(\boldsymbol{\psi})$ je obtížná (neboť s logaritmem součtu se velmi špatně pracuje). Proto v této situaci používáme právě EM algoritmus. Dále také proto, že je schopen z maximalizace nekompletních dat přejít na maximalizaci dat kompletních.

Našim cílem je dostat $\hat{\boldsymbol{\psi}} = (\hat{w}_1, \ldots, \hat{w}_g, \hat{\boldsymbol{\mu}}_1, \ldots, \hat{\boldsymbol{\mu}}_g, \hat{\boldsymbol{\Sigma}})$ (maximálně věrohodný odhad parametru $\boldsymbol{\psi}$) jako řešení soustavy věrohodnostních rovnic

$$\frac{\partial \ell(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}} = \mathbf{0}.$$

3.2 Věrohodnostní funkce kompletních dat a její vlastnosti

Zásadní úlohu v EM algoritmu hraje log-věrohodnostní funkce pro $\boldsymbol{\psi}$ z kompletních dat. Pro tu si nejprve vyjádříme hustotu náhodného vektoru $(\boldsymbol{Y}_j^\top, \boldsymbol{Z}_j^\top)^\top$, pro $j = 1, \ldots, n$:

$$\begin{aligned} f_c(\boldsymbol{Y}_j, \boldsymbol{Z}_j; \boldsymbol{\psi}) &= f(\boldsymbol{Y}_j|\boldsymbol{Z}_j; \boldsymbol{\psi}) f(\boldsymbol{Z}_j; \boldsymbol{\psi}) \\ &= \sum_{k=1}^{g} Z_{kj} \phi_k(\boldsymbol{Y}_j; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}) \sum_{k=1}^{g} Z_{kj} w_k = \prod_{k=1}^{g} (\phi_k(\boldsymbol{Y}_j; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}) w_k)^{Z_{kj}}. \end{aligned}$$

Z toho pro log-věrohodnostní funkci pro $\boldsymbol{\psi}$ z kompletních dat platí

$$\ell_c(\boldsymbol{\psi}) = \sum_{k=1}^{g} \sum_{j=1}^{n} Z_{kj} \left(\log w_k + \log \left(\phi_k(\boldsymbol{Y}_j; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}) \right) \right).$$

V následujícím lemmatu si shrneme základní vlastnosti $\ell_c(\boldsymbol{\psi})$, které budeme dále potřebovat.

Lemma 4. *Nechť* $\tau_k(\boldsymbol{y}_j; \boldsymbol{\psi}) = \frac{w_k \phi_k(\boldsymbol{y}_j; \boldsymbol{\psi})}{\sum_{h=1}^{g} w_h \phi_h(\boldsymbol{y}_j; \boldsymbol{\psi})}$, *pak platí*

1.

$$\mathsf{E}_{\boldsymbol{\psi}}[Z_{kj}|\boldsymbol{Y}_j = \boldsymbol{y}_j] = \tau_k(\boldsymbol{y}_j; \boldsymbol{\psi}),$$

2.

$$\begin{aligned} &\mathsf{E}_{\boldsymbol{\psi}}[\ell_c(\boldsymbol{\psi})|\boldsymbol{Y} = \boldsymbol{y}] = \\ &\sum_{k=1}^{g} \sum_{j=1}^{n} \tau_k(\boldsymbol{y}_j; \boldsymbol{\psi}) \left(\log w_k - \frac{1}{2} \log (2\pi)^p - \frac{1}{2} \log |\boldsymbol{\Sigma}| - \frac{1}{2} (\boldsymbol{y}_j - \boldsymbol{\mu}_k)^\top (\boldsymbol{\Sigma})^{-1} (\boldsymbol{y}_j - \boldsymbol{\mu}_k) \right), \end{aligned}$$

kde $|\boldsymbol{\Sigma}|$ *značí determinant matice* $\boldsymbol{\Sigma}$*,*

3.

$$\begin{aligned} &\arg\max_{\boldsymbol{\psi}} \mathsf{E}_{\boldsymbol{\psi}}[\ell_c(\boldsymbol{\psi})|\boldsymbol{Y} = \boldsymbol{y}] = \\ &\Bigg(\frac{1}{n} \sum_{j=1}^{n} \tau_1(\boldsymbol{y}_j; \boldsymbol{\psi}), \frac{1}{n} \sum_{j=1}^{n} \tau_2(\boldsymbol{y}_j; \boldsymbol{\psi}), \ldots, \frac{1}{n} \sum_{j=1}^{n} \tau_g(\boldsymbol{y}_j; \boldsymbol{\psi}), \\ &\frac{\sum_{j=1}^{n} \tau_1(\boldsymbol{y}_j; \boldsymbol{\psi}) \boldsymbol{y}_j}{\sum_{j=1}^{n} \tau_1(\boldsymbol{y}_j; \boldsymbol{\psi})}, \frac{\sum_{j=1}^{n} \tau_2(\boldsymbol{y}_j; \boldsymbol{\psi}) \boldsymbol{y}_j}{\sum_{j=1}^{n} \tau_2(\boldsymbol{y}_j; \boldsymbol{\psi})}, \ldots, \frac{\sum_{j=1}^{n} \tau_g(\boldsymbol{y}_j; \boldsymbol{\psi}) \boldsymbol{y}_j}{\sum_{j=1}^{n} \tau_g(\boldsymbol{y}_j; \boldsymbol{\psi})}, \\ &\frac{1}{n} \sum_{k=1}^{g} \sum_{j=1}^{n} \tau_k(\boldsymbol{y}_j; \boldsymbol{\psi}) (\boldsymbol{y}_j - \boldsymbol{\mu}_k)(\boldsymbol{y}_j - \boldsymbol{\mu}_k)^\top \Bigg)^\top. \end{aligned}$$

Důkaz.

1.

$$\mathsf{E}_{\boldsymbol{\psi}}[Z_{kj}|\boldsymbol{Y}_j = \boldsymbol{y}_j] = P_{\boldsymbol{\psi}}(Z_{kj} = 1|\boldsymbol{Y}_j = \boldsymbol{y}_j) = \tau_k(\boldsymbol{y}_j; \boldsymbol{\psi}),$$

kde druhá rovnost plyne ze vztahu (1.3).

2. Vypočteme hodnotu této funkce a upravíme do tvaru vhodného k derivování.

$$\begin{aligned}
\mathsf{E}_{\boldsymbol{\psi}}[l_c(\boldsymbol{\psi})|\boldsymbol{Y}=\boldsymbol{y}] &= \sum_{k=1}^{g}\sum_{j=1}^{n}\mathsf{E}_{\boldsymbol{\psi}}[Z_{kj}|\boldsymbol{Y}_j=\boldsymbol{y}_j]\left(\log w_k + \log\left(\phi_k(\boldsymbol{y}_j;\boldsymbol{\mu}_k,\boldsymbol{\Sigma})\right)\right)\\
&= \sum_{k=1}^{g}\sum_{j=1}^{n}\tau_k(\boldsymbol{y}_j;\boldsymbol{\psi})\left(\log w_k + \log\left(\phi_k(\boldsymbol{y}_j;\boldsymbol{\mu}_k,\boldsymbol{\Sigma})\right)\right)\\
&= \sum_{k=1}^{g}\sum_{j=1}^{n}\tau_k(\boldsymbol{y}_j;\boldsymbol{\psi})\left(\log w_k + \log\left(\frac{1}{\sqrt{(2\pi)^p|\boldsymbol{\Sigma}|}}\exp\left(\frac{1}{2}(\boldsymbol{y}_j-\boldsymbol{\mu}_k)^\top(\boldsymbol{\Sigma})^{-1}(\boldsymbol{y}_j-\boldsymbol{\mu}_k)\right)\right)\right)\\
&= \sum_{k=1}^{g}\sum_{j=1}^{n}\tau_k(\boldsymbol{y}_j;\boldsymbol{\psi})\left(\log w_k - \frac{1}{2}\log(2\pi)^p - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}(\boldsymbol{y}_j-\boldsymbol{\mu}_k)^\top(\boldsymbol{\Sigma})^{-1}(\boldsymbol{y}_j-\boldsymbol{\mu}_k)\right).
\end{aligned}$$

3. Chceme najít řešení rovnice $\hat{\boldsymbol{\psi}} = \arg\max_{\boldsymbol{\psi}}\mathsf{E}_{\boldsymbol{\psi}}[\ell_c(\boldsymbol{\psi})|\boldsymbol{Y}=\boldsymbol{y}]$. Parametr $\boldsymbol{\psi}$ lze rozdělit na $\boldsymbol{w} = (w_1,\ldots,w_g)^\top$ a $\boldsymbol{\xi} = (\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_g,\boldsymbol{\Sigma})$. Stejně tak lze $\mathsf{E}_{\boldsymbol{\psi}}[\ell_c(\boldsymbol{\psi})|\boldsymbol{Y}=\boldsymbol{y}]$ rozdělit na část závisející na $\boldsymbol{w}$ a část závisející na $\boldsymbol{\xi}$ jako

$$\arg\max_{\boldsymbol{\psi}}\mathsf{E}_{\boldsymbol{\psi}}[\ell_c(\boldsymbol{\psi})|\boldsymbol{Y}=\boldsymbol{y}] = S(\boldsymbol{w},\boldsymbol{\psi}) + T(\boldsymbol{\xi},\boldsymbol{\psi}),$$

kde

$$S(\boldsymbol{w},\boldsymbol{\psi}) = \sum_{k=1}^{g}\sum_{j=1}^{n}\tau_k(\boldsymbol{y}_j;\boldsymbol{\psi})\log w_k$$

a

$$T(\boldsymbol{\xi},\boldsymbol{\psi}) = \sum_{k=1}^{g}\sum_{j=1}^{n}\tau_k(\boldsymbol{y}_j;\boldsymbol{\psi})\log\left(\phi_k(\boldsymbol{y}_j;\boldsymbol{\mu}_k,\boldsymbol{\Sigma})\right).$$

Tím můžeme maximalizaci $\mathsf{E}_{\boldsymbol{\psi}}[\ell_c(\boldsymbol{\psi})|\boldsymbol{Y}=\boldsymbol{y}]$ rozdělit na dvě části. Maximalizací $S(\boldsymbol{w},\boldsymbol{\psi})$ získáme nový odhad $\boldsymbol{w}$ a maximalizací $T(\boldsymbol{\xi},\boldsymbol{\psi})$ získáme nový odhad $\boldsymbol{\xi}$.

Pro hledání $\hat{w}_k$ použijeme metodu Lagrangeových multiplikátorů s vazebnou podmínkou

$$\sum_{k=1}^{g} w_k = 1.$$

Řešíme rovnici

$$\frac{\partial}{\partial w_k}\left(\sum_{k=1}^{g}\sum_{j=1}^{n}\tau_k(\boldsymbol{y}_j;\boldsymbol{\psi})\log w_k + \lambda(\sum_{k=1}^{g} w_k - 1)\right) = 0, \tag{3.1}$$

kde $\lambda\in\mathbb{R}$ je Lagrangeův multiplikátor. Z (3.1) plyne, že

$$\sum_{j=1}^{n}\frac{\tau_k(\boldsymbol{y}_j;\boldsymbol{\psi})}{\hat{w}_k} + \lambda = 0,$$

tedy

$$\sum_{j=1}^{n}\tau_k(\boldsymbol{y}_j;\boldsymbol{\psi}) + \lambda\hat{w}_k = 0$$

a neboť

$$\sum_{k=1}^{g} \hat{w_k} = 1$$

a

$$\sum_{k=1}^{g}\sum_{j=1}^{n} \tau_k(\boldsymbol{y}_j;\boldsymbol{\psi}) = n,$$

tak

$$n + \lambda * 1 = 0 \implies \lambda = -n.$$

Z čehož dostáváme

$$\hat{w_k} = \frac{1}{n}\sum_{j=1}^{n} \tau_k(\boldsymbol{y}_j;\boldsymbol{\psi}), \quad k = 1,\ldots,g.$$

Pro hledání $\hat{\boldsymbol{\mu}_k}$ zderivujeme $T(\boldsymbol{\xi},\boldsymbol{\psi})$ podle $\boldsymbol{\mu}_k$, položíme rovno nule a vyřešíme

$$\sum_{j=1}^{n} \tau_k(\boldsymbol{y}_j;\boldsymbol{\psi})\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}_k}) = 0, \quad k = 1,\ldots,g,$$

a protože je $\boldsymbol{\Sigma}$ regulární, upravíme rovnici do tvaru

$$\sum_{j=1}^{n} \tau_k(\boldsymbol{y}_j;\boldsymbol{\psi})\boldsymbol{y}_j = \sum_{j=1}^{n} \tau_k(\boldsymbol{y}_j;\boldsymbol{\psi})\hat{\boldsymbol{\mu}_k}.$$

Z čehož dostáváme

$$\hat{\boldsymbol{\mu}_k} = \frac{\sum_{j=1}^{n} \tau_k(\boldsymbol{y}_j;\boldsymbol{\psi})\boldsymbol{y}_j}{\sum_{j=1}^{n} \tau_k(\boldsymbol{y}_j;\boldsymbol{\psi})}, \quad k = 1,\ldots,g.$$

Obdobně pro hledání $\hat{\boldsymbol{\Sigma}}$ zderivujeme $T(\boldsymbol{\xi},\boldsymbol{\psi})$ podle $\boldsymbol{\Sigma}^{-1}$, položíme rovno nule a vyřešíme:

$$\sum_{k=1}^{g}\sum_{j=1}^{n} \tau_k(\boldsymbol{y}_j;\boldsymbol{\psi}) \left(\hat{\boldsymbol{\Sigma}} - (\boldsymbol{y}_j - \hat{\boldsymbol{\mu}_k})(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}_k})^\top\right) = 0,$$

z toho

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{k=1}^{g}\sum_{j=1}^{n} \tau_k(\boldsymbol{y}_j;\boldsymbol{\psi})(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}_k})(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}_k})^\top.$$

□

3.3 EM algoritmus

Nyní si představíme samotný EM algoritmus a jeho fungování na základě informací uvedených výše.

Začneme s počátečním odhadem parametrů $\boldsymbol{\psi}^{(0)}$, který volíme libovolně z parametrického prostoru $\boldsymbol{\Psi}$. Provedeme jeden iterační krok EM algoritmu (E-krok a M-krok) a dostaneme $\boldsymbol{\psi}^{(1)} \in \boldsymbol{\Psi}$ takové, že $L(\boldsymbol{\psi}^{(0)}) \leq L(\boldsymbol{\psi}^{(1)})$. Postupujeme dále až do té doby, než dostaneme $\boldsymbol{\psi}^{(m)}$ takové, že

$$|L(\boldsymbol{\psi}^{(m)}) - L(\boldsymbol{\psi}^{(m-1)})| \leq \alpha \tag{3.2}$$

pro $\alpha \in (0, \infty)$ předem dané dostatečně malé.

Definujeme funkci $Q(\boldsymbol{\psi},\boldsymbol{\psi}^{(p)})$ jako

$$Q(\boldsymbol{\psi},\boldsymbol{\psi}^{(p)}) = \mathsf{E}_{\boldsymbol{\psi}^{(p)}}[\ell_c(\boldsymbol{\psi})|\boldsymbol{y}].$$

V E-kroku hodnotu této funkce spočteme a v M-kroku maximalizujeme $Q(\boldsymbol{\psi},\boldsymbol{\psi}^{(p)})$ přes $\boldsymbol{\Psi}$.

Konkrétně je hodnota funkce $Q(\boldsymbol{\psi},\boldsymbol{\psi}^{(p)})$ je dle druhé části lemmatu 4 rovna

$$\sum_{k=1}^{g}\sum_{j=1}^{n} \tau_k(\boldsymbol{y}_j;\boldsymbol{\psi}^{(p)}) \left(\log w_k - \frac{1}{2}\log(2\pi)^p - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}(\boldsymbol{y}_j - \boldsymbol{\mu}_k)^\top(\boldsymbol{\Sigma})^{-1}(\boldsymbol{y}_j - \boldsymbol{\mu}_k)\right)$$

V M-kroku chceme najít řešení rovnice $\boldsymbol{\psi}^{(p+1)} = \arg\max_{\boldsymbol{\psi}\in\Psi} Q(\boldsymbol{\psi},\boldsymbol{\psi}^{(p)})$.

Dle třetí části lemmatu 4 jsou odhady vah rovny

$$w_k^{(p+1)} = \frac{1}{n}\sum_{j=1}^{n} \tau_k(\boldsymbol{y}_j;\boldsymbol{\psi}^{(p)}), \quad k = 1,\ldots,g.$$

To lze interpretovat tak, že každé pozorování přispívá k odhadu w_k takovou vahou, jaká je jeho pravděpodobnost, že do skupiny k patří.

Dále odhady průměrů jsou rovny

$$\boldsymbol{\mu}_k^{(p+1)} = \frac{\sum_{j=1}^{n} \tau_k(\boldsymbol{y}_j;\boldsymbol{\psi}^{(p)})\boldsymbol{y}_j}{\sum_{j=1}^{n} \tau_k(\boldsymbol{y}_j;\boldsymbol{\psi}^{(p)})}, \quad k = 1,\ldots,g,$$

což je vážený průměr pozorování příslušících ke skupině k.

A nakonec odhad varianční matice je

$$\boldsymbol{\Sigma}^{(p+1)} = \frac{1}{n}\sum_{k=1}^{g}\sum_{j=1}^{n} \tau_k(\boldsymbol{y}_j;\boldsymbol{\psi}^{(p)})(\boldsymbol{y}_j - \boldsymbol{\mu}_k^{(p+1)})(\boldsymbol{y}_j - \boldsymbol{\mu}_k^{(p+1)})^\top,$$

což je analogií vážené výběrové varianční matice.

Za zmínku stojí, že pro odhady vah $w_k^{(p+1)}$ bychom se k tomuto výsledku dostali i kdybychom uvažovali obecnější případ, tedy libovolnou konečnou směs. Naopak ale pro odhady středních hodnot $\boldsymbol{\mu}_k^{(p+1)}$ a varianční matice $\boldsymbol{\Sigma}^{(p+1)}$ bychom v obecném případě konečných směsí nijak derivaci funkce $T(\boldsymbol{\xi},\boldsymbol{\psi}^{(p)})$ upravit nemohli, a tedy ani dostat explicitní odhady. Případ normálních směsí je jedním z mála, kde zvládneme explicitně určit odhady všech neznámých parametrů.

4. Praktické použití

V poslední části práce aplikujeme teorii popsanou v prvních třech kapitolách na klasifikaci reálných dat.

Máme k dispozici archeologické popisy lebek z článku Howellse (Howells, 1996). Konkrétně se jedná o 156 ženských lebek, u kterých jsou známy jejich největší délky mozkovny v milimetrech (*gol*) a týlní úhly ve stupních (*oca*). Dále máme informaci o místě jejich nálezu: Austrálie, Berg v Rakousku nebo Burjati na Sibiři.

Pro ilustraci klasifikace zapomeneme, že známe rozdělení lebek dle místa nálezu a budeme se je snažit správně zařadit do skupin pouze na základě dat o největších délkách mozkovny a týlních úhlech. Následně porovnáme naše výsledky se skutečností.

4.1 Popis dat

Nejprve se podíváme na základní popisné statistiky proměnných *gol* a *oca* jak pro skupinu všch lebek, tak pro lebky v jednotlivých skupinách. Tyto statistiky jsou shrnuty v tabulkách 4.2 a 4.1.

Pro lepší přehlednost rozdílů mezi skupinami jsme vytvořili boxploty pro proměnné *gol* a *oca*, které jsou prezentovány v grafech 4.1 a 4.2.

Jak můžeme vidět, skupina Austrálie má výrazně větší hodnoty největší délky mozkovny než skupina Berg a Burjati, které mají hodnoty podobné. U týlního úhlu pozorujeme menší rozdíly mezi skupinami, přičemž skupina Austrálie vykazuje nižší hodnoty než skupiny Berg a Burjati.

Můžeme tedy předpokládat, že se nám nejlépe podaří oddělit skupinu Austrálie od zbytku, ale mezi skupiny Berg a Burjati se nám správně klasifikovat nejspíš nepovede.

V neposlední řadě si vykreslíme graf 4.3 ukazující vztah mezi *gol* a *oca* a barevně odděluje příslušníky jednotlivých skupin.

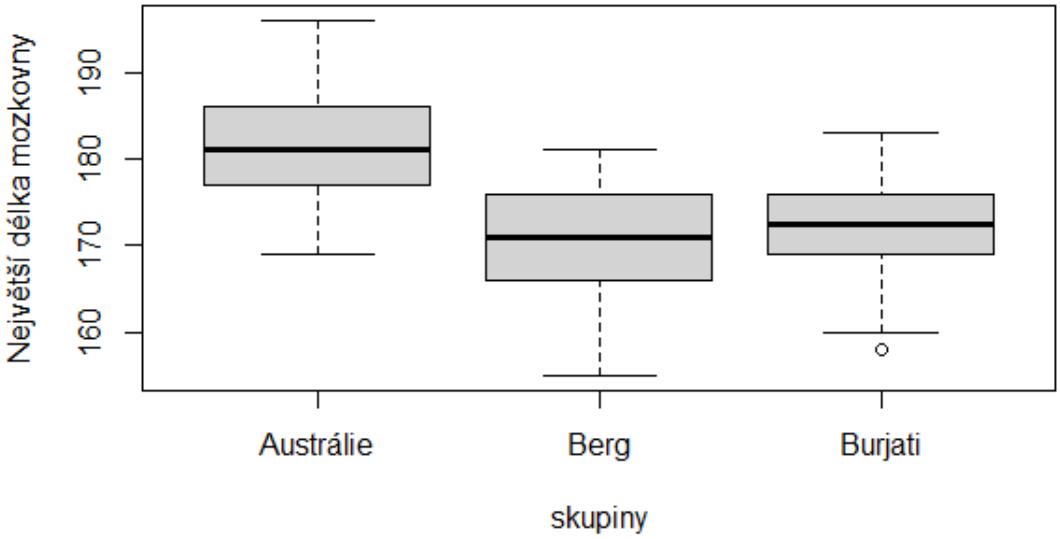
Na základě směrodatných odchylek a kovariancí mezi proměnnými v jednotlivých skupinách můžeme předpokládat homoskedasticitu.

	kovariance
všichni	-19,95
Austrálie	-12,76
Berg	-22,44
Burjati	-10,07

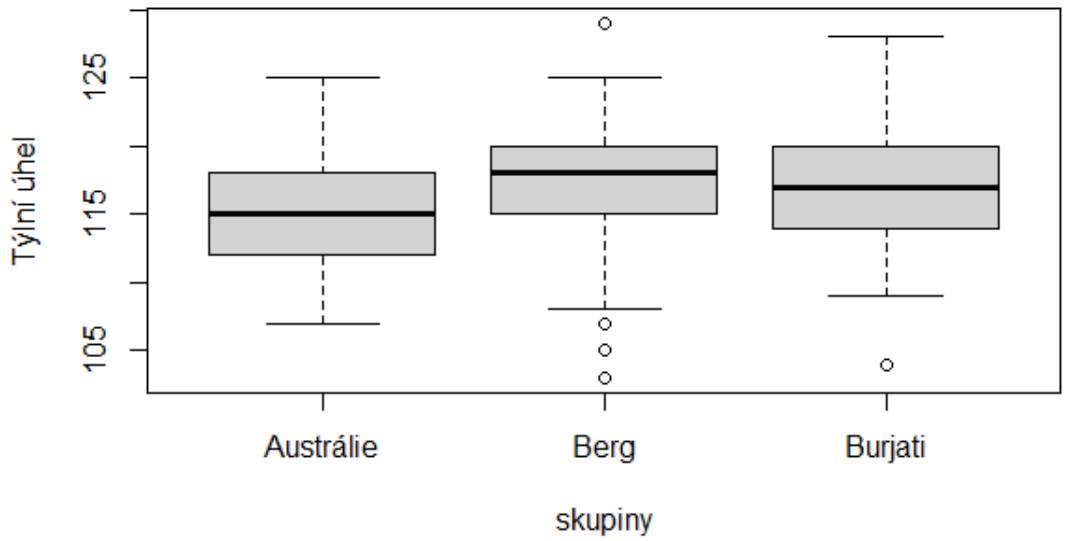
Tabulka 4.1: Kovariance mezi *gol* a *oca*.

	minimum	medián	výb. průměr	maximum	směr. odchylka
			Největší délka mozkovny [mm]		
všichni	155	175,0	174,3	196	7,64
Austrálie	169	181,0	181,1	196	6,36
Berg	155	171,0	170,5	181	6,51
Burjati	158	172,5	171,8	183	5,43
			Týlní úhel [stupně]		
všichni	103	117,0	116,3	129	4,89
Austrálie	107	115,0	114,7	125	4,06
Berg	103	118,0	117,0	129	5,56
Burjati	104	117,0	117,0	128	4,61

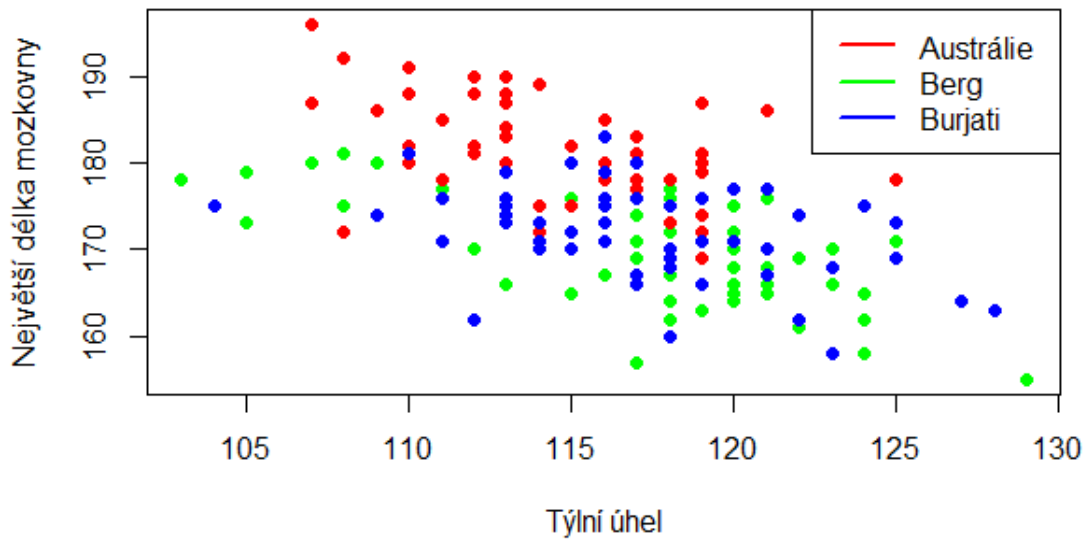
Tabulka 4.2: Základní popisné statistiky *gol* a *oca*.



Obrázek 4.1: Největší délka mozkovny v jednotlivých skupinách.



Obrázek 4.2: Týlní úhel v jednotlivých skupinách.



Obrázek 4.3: Vztah mezi *gol* a *oca*.

4.2 Klasifikace

Nyní zapomeneme, že skutečnou příslušnost lebek do skupin známe a pokusíme se ji určit pomocí klasifikační metody popsané v prvních třech kapitolách.

Nejprve potřebujeme zvolit počáteční hodnoty neznámých parametrů. Za počáteční odhady vah vezmeme pro všechny skupiny stejnou hodnotu, tedy

$$w_1^{(0)} = w_2^{(0)} = w_3^{(0)} = \frac{1}{3}.$$

Jak jsme uvedli v kapitole 4.1, u proměnné *gol* se pohybujeme mezi 155mm a 196mm a u proměnné *oca* mezi 103° a 129°. Tyto rozsahy rovnoměrně rozdělíme a zaokrouhlíme na celá čísla, tím získáme hodnoty

$$\boldsymbol{\mu}_1^{(0)} = \begin{pmatrix} 186 \\ 123 \end{pmatrix}, \quad \boldsymbol{\mu}_2^{(0)} = \begin{pmatrix} 176 \\ 116 \end{pmatrix}, \quad \boldsymbol{\mu}_3^{(0)} = \begin{pmatrix} 165 \\ 109 \end{pmatrix},$$

jako počáteční odhady středních hodnot. Konečně za počáteční hodnotu varianční matice vezmeme

$$\boldsymbol{\Sigma}^{(0)} = \begin{pmatrix} 58{,}37 & -19{,}95 \\ -19{,}95 & 23{,}91 \end{pmatrix},$$

kterou jsme sestavili pomocí kvadrátů směroatných odchylek a kovariance pro všechna pozorování.

Pro tyto hodnoty spustíme EM algoritmus s $\alpha = 10^{-6}$ z (3.2) a obdržíme konečné odhady neznámých parametrů.

Tabulka 4.3 porovnává odhady se skutečnými hodnotami neznámých parametrů, kde odhad skutečné varianční matice jsme spočítali následovně:

$$\frac{1}{n-3}\left(\sum_{i=1}^{n_1}(\boldsymbol{Y}_{1i}-\boldsymbol{\mu}_1)(\boldsymbol{Y}_{1i}-\boldsymbol{\mu}_1)^\top + \sum_{i=1}^{n_2}(\boldsymbol{Y}_{2i}-\boldsymbol{\mu}_2)(\boldsymbol{Y}_{2i}-\boldsymbol{\mu}_2)^\top + \sum_{i=1}^{n_3}(\boldsymbol{Y}_{3i}-\boldsymbol{\mu}_3)(\boldsymbol{Y}_{3i}-\boldsymbol{\mu}_3)^\top\right),$$

kde $\boldsymbol{Y}_{ki}$ je i-té pozorování příslušící k-té skupině a n_k je počet pozorování v k-té skupině, pro $k = 1, 2, 3$.

parametr	**odhad**	**skutečnost**
w_1	0,75	0,31
w_2	0,08	0,34
w_3	0,17	0,35
$\boldsymbol{\mu}_1$	$\begin{pmatrix} 174{,}63 \\ 116{,}02 \end{pmatrix}$	$\begin{pmatrix} 181{,}10 \\ 114{,}70 \end{pmatrix}$
$\boldsymbol{\mu}_2$	$\begin{pmatrix} 173{,}97 \\ 108{,}43 \end{pmatrix}$	$\begin{pmatrix} 170{,}50 \\ 117{,}00 \end{pmatrix}$
$\boldsymbol{\mu}_3$	$\begin{pmatrix} 173{,}00 \\ 121{,}21 \end{pmatrix}$	$\begin{pmatrix} 171{,}80 \\ 117{,}00 \end{pmatrix}$
$\boldsymbol{\Sigma}$	$\begin{pmatrix} 57{,}69 & -18{,}90 \\ -18{,}90 & 14{,}58 \end{pmatrix}$	$\begin{pmatrix} 37{,}32 & -15{,}12 \\ -15{,}12 & 23{,}05 \end{pmatrix}$

Tabulka 4.3: Porovnání odhadů se skutečností.

Nakonec se podívejme na to, jak se nám podařilo zařadit jednotlivá pozorování do skupin.

Vzhledem k tomu, že metoda je schopna pouze rozdělit populaci na tři skupiny a nikoli je rozřadit do „pravých“ skupin, zvolili jsme pojmenování jednotlivých skupin jako takovou kombinaci odhadu a skutečnosti, která maximalizuje počet správně zařazených.

Tabulce 4.4 zachycuje kolik lebek se nám podařilo zařadit správně. Z celkových 156 lebek byla metoda úspěšná u 58 z nich.

Skutečná skupina / Odhadnutá skupina	Austrálie	Berg	Burjati
Austrálie	44	42	43
Berg	1	6	3
Burjati	4	5	8

Tabulka 4.4: Porovnání zařazení se skutečností.

Závěr

V této práci jsme se věnovali tomu, jak lze konečné směsi využít ke klasifikaci populace do skupin. Důraz jsme kladli na odhadování neznámých parametrů, které jsou ke klasifikaci stěžejní. Proto jsme popsali metodu maximální věrohodnosti, a to i s problémy, které zde mohou nastat. Zjistili jsme také, že hledat maximálně věrohodný odhad pro směsové modely není vůbec snadné, a proto jsme použili EM algoritmus, který je zde popsán. Zde by se mohla práce dále rozvinout o určité vlastnosti tohoto algoritmu, které by nám objasnily, proč je tak výhodné ho v této situaci použít.

S pomocí EM algoritmu jsme se snažili roztřídit lebky na základě jejich největší délky mozkovny a týlních úhlů do správných skupin podle místa nálezu. Výsledky ukázaly, že tato metoda je velmi přínosná pro klasifikaci určitého typu dat.

Věřím, že tato práce čtenáře obohatila a usnadnila mu vhled do problematiky využití konečných směsí ke klasifikaci.

Seznam použité literatury

Dempster, A. P., Laird, N. M. a Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **39**(1), 1–38.

Howells, W. W. (1996). Howells' craniometric data on the internet. *American journal od physical anthropology*, **101**, 441–442.

McLachlan, G. J. a Peel, D. (2000). *Finite mixture models*. Wiley Series in Probability and Statistics, New York.