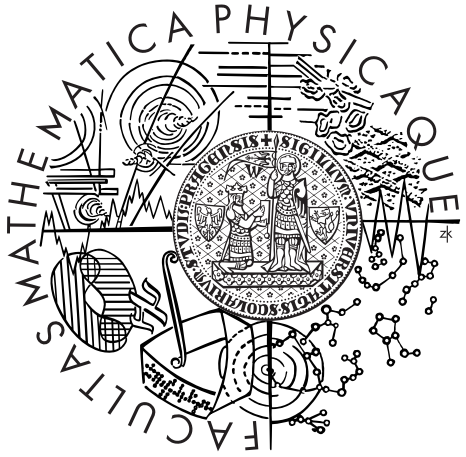
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Tomáš Rusý

Modely konečných směsí

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce:	doc. RNDr. Arnošt Komárek, Ph.D.
Studijní program:	Matematika
Studijní obor:	Obecná matematika

Praha 2015

Poděkování

Rád bych na tomto místě poděkoval mému panu vedoucímu doc. RNDr. Arnoštu Komárkovi, PhD. za cenné rady, věcné připomínky a drobné korektury, které významně pozvedly úroveň této bakalářské práce. Děkuji také panu RNDr. Jaromíru Běláčkovi, CSc. za poskytnutí dat k ilustraci modelu směsi v mé bakalářské práci a za uvedení do problematiky, jichž se data týkají.

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

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 v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ..... Podpis autora

Název práce: Modely konečných směsí

Autor: Tomáš Rusý

Katedra pravděpodobnosti a matematické statistiky

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

Abstrakt: Tato práce se věnuje konečným směsím a klade si za cíl představit čtenáři použití metody maximální věrohodnosti pro odhad modelu směsi. K tomu využije EM algoritmus, který je v práci detailně zaveden a odvozen. Pro obecný model směsi popíše oba dva kroky algoritmu, E-kroku a M-kroku, a také výpočet nových odhadů části parametrů reprezentující model. Pro specifickou rodinu normálních směsí jsou pak odvozeny explicitní vzorce pro iteraci EM algoritmu. Popsaná teorie je pak aplikována na modelu produkce cytokinu interleukin 10 při napadení člověka paradontózou, díky kterému je názorně ukázána využitelnost modelu v praxi. Na závěr je pak na odhad modelu směsi navázáno teorií shlukování, zabývající se rozřazování prvků do skupin. Podobně jako předchozí teorie je také ilustrována na výše zmíněném modelu produkce cytokinu IL10.

Klíčová slova: konečná směs, normální směs, EM algoritmus, shlukování

Title: Finite Mixture Models

Author: Tomáš Rusý

Department of Probability and Mathematical Statistics

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

Abstract: This work focuses on finite mixture models and aims to introduce the maximum likelihood method as an approach of fitting finite mixtures. For that purpose the EM algorithm is adopted and derived in detail. Both the E and M steps of the EM algorithm are presented and performed for general finite mixture model. We derive new estimates for some of parameters defining the model. All updated estimates of the iteration of the EM algorithm are derived explicitly for the specific family of normal mixtures. The described theory is then applied to a model of the production of cytokine interleukin 10 in human periodontitis attack, which clearly demonstrates an application of the model in practice. Finally, we discuss the theory of clustering, which is based on our previous results. Like the previous theory, this one is also illustrated in the aforementioned model of cytokine IL10 production.

Keywords: finite mixture, normal mixture, EM algorithm, clustering

Obsah

Použité značení

$\mathbb{N}$	množina přirozených čísel
$\mathbb{R}$	množina reálných čísel
$\mathbb{R}^p$	vektorový prostor dimenze p nad reálnými čísly
$\boldsymbol{u}$	sloupcový vektor $\boldsymbol{u}$
$\boldsymbol{A}^\top$	transponovaná matice $\boldsymbol{A}$
$\boldsymbol{A} \otimes \boldsymbol{B}$	Kroneckerův součin matic $\boldsymbol{A}$ a $\boldsymbol{B}$
$\mathrm{int}(\Psi)$	vnitřek množiny Ψ
$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	normální rozdělení se střední hodnotou $\boldsymbol{\mu}$ a rozptylovou maticí $\boldsymbol{\Sigma}$
$\varphi(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$	hustota normálního rozdělení s vektorem středních hodnot $\boldsymbol{\mu}$ a rozptylovou maticí $\boldsymbol{\Sigma}$.
$\mathsf{P}(\cdot\|\boldsymbol{y})$	$\mathsf{P}(\cdot\|\boldsymbol{Y} = \boldsymbol{y})$
$\mathsf{E}\,[\cdot\|\boldsymbol{y}]$	$\mathsf{E}\,[\cdot\|\boldsymbol{Y} = \boldsymbol{y}]$
$\mathsf{E}\,[g(\boldsymbol{Y})\|\boldsymbol{\psi}']$	střední hodnota z funkce $g(\boldsymbol{Y})$, když hustota náhodného vektoru $\boldsymbol{Y}$ je určena hodnotou parametrického vektoru $\boldsymbol{\psi}'$
MMV	metoda maximální věrohodnosti
IL10	cytokin interleukin 10
IL10_Pi	logaritmizované hodnoty produkce cytokinu IL10, při stimulování kulturou bakterie *prevotella intermedia*
IL10_nestim	logaritmizované hodnoty produkce cytokinu IL10, nestimulované

Úvod

Konečné směsi představují stále se vyvíjející a často používanou metodu statistické analýzy, jež nalezla širšího uplatnění zejména díky vývoji výpočetní techniky v minulém století. Úlohy, jejichž řešení bylo dříve nemyslitelné, se se vznikem počítačů staly překonatelnou překážkou, nad kterou bylo možno vystavit teorii zabývající se touto metodou. Za cíl této bakalářské práce si klademe model konečné směsi názorně představit a ilustrovat jeho použití na vhodném praktickém případě.

V úvodní části je představen cíl a struktura práce a zmíníme se i o prvním publikovaném použití modelu směsi ve statistické analýze. Na to navážeme v první kapitole, která je věnována představení modelu směsi. Její součástí je jak rigorózní definice objektu, kterým model směsi nazýváme, tak i abstraktní a praktická ilustrace, jenž usnadňují pochopení modelu. Mimo jiné se v této části zmíníme i o reálné interpretaci tohoto teoretického pojmu a situacím, kdy na něj můžeme v obyčejném světě narazit.

Protože ve statistické analýze je zásadním problémem uvažované modely odhadnout a zkoumat jejich vlastnosti, i v našem textu se podobné technice budeme věnovat. Konkrétně se ve druhé kapitole budeme zabývat využitím metody maximální věrohodnosti jako cesty, kterou lze obecný model směsi odhadnout. Zjistíme, že samotnému nalezení odhadu předchází řada komplikací, které se budeme snažit vyřešit. K nalezení odhadů pak budeme používat EM algoritmus, který pro směsi podrobně zavedeme a detailně si vysvětlíme princip, na kterém je založen. Na závěr této kapitoly pak odvodíme v maximální obecnosti vzorce pro část parametrů modelu směsi jedné iterace EM algoritmu.

Odvodit i zbylé vzorce pro ostatní parametry modelu se ukazuje jako nemožné, a proto se ve třetí kapitole omezíme pouze na normální směsi, tj. směsi, jejichž složky mají normální rozdělení. Ukážeme si, že pro tuto specifickou rodinu lze vzorce pro nové odhady parametrů EM algoritmu vyjádřit explicitně, což značně zjednodušuje jeho použití. Tím zakončíme teoretickou část práce, věnující se odhadování pomocí metody maximální věrohodnosti a ukážeme si použití námi odvozené techniky v praxi. K tomu nám poslouží model, jenž je představen v první kapitole, zabývající se produkcí obrané látky interleukin 10 při napadení člověka onemocněním zubů - paradontózou.

Nakonec se budeme ve čtvrté kapitole věnovat odpovědi na jednu přirozenou otázku, která v kontextu naší interpretace směsi vyvstává. Tou je, že v případě, kdy se nějaká vlastnost chová jinak ve více skupinách lidí, zda-li nemůžeme

na základě pozorování té vlastnosti u nějakého člověka určit, ze které skupiny pochází. Odborně této technice říkáme shlukování a v rámci kapitoly si představíme jedno pravidlo, kterým se v tomto případě můžeme řídit. Opět pro lepší pochopení ukážeme použití představené látky na modelové situaci uvedené v první kapitole.

Věřím, že celkově bakalářská práce představuje podrobný popis problematiky konečných směsích a komplexní odvození velké většiny vyvozených závěrů. Díky tomu by měla být pochopitelná i širšímu matematickému publiku, jelikož staví pouze na základech teorie pravděpodobnosti.

Motivace

První zmínka o problematice konečných směsí se objevila na konci 19-tého století v práci anglického matematika a biostatistika Karla Pearsona. Pearson (1894) se v ní zabýval daty populace krabů v zálivu u města Neapol. Ty mu byly dodány od biologa Weldona, který analyzoval poměr velikosti jejich čela ku délce těla. Na základě toho pak vyslovil hypotézu, že naměřená data mohou být signálem vzniku nového druhu krabů v zálivu. Tato domněnka byla podpořena faktem, že rozdělení testovaných dat neodpovídalo normálnímu rozdělení.

Pearson se snažil odhadnout toto rozdělení směsí dvou normálních rozdělení momentovou metodou, což při pěti proměnných vedlo na polynomiální rovnici devátého stupně. A ačkoliv se mu tuto úlohu nakonec podařilo vyřešit, následovníků, kteří se vydali v jeho stopách, mnoho nebylo. To proto, že řešení podobných či složitějších rovnic bylo v té době velmi obtížné, pracné, případně nemožné.

Díky pokroku techniky v posledních desetiletích se ale řešení takovýchto úloh stalo snadným, což zapříčinilo značný rozvoj teorie odhadování modelů konečných směsí. Ta posléze našla uplatnění v mnoha odvětvích. Problematika shlukování a vlastnost normálních konečných směsí, že velmi dobře dokáží aproximovat širokou třídu rozdělení, patří mezi ty nejužívanější. Zejména ta druhá je ale už natolik pokročilá, že se jí v této práci věnovat nebudeme.

Kapitola 1

Model konečné směsi

V této kapitole si představíme, co pod konečnou směsí rozumíme a to včetně parametrického vyjádření, které budeme používat v dalších částech práce. Krom toho se navíc zmíníme o interpretaci, kterou tento abstraktní pojem má, a jak je možno na něj nahlížet. Pro lepší představu pak uvedeme jeden praktický příklad, kde náhodný vektor vystihující diskutovaný problém bude mít rozdělení, které je možno modelem směsi popsat.

1.1 Konečná směs

Definice 1 (Konečná směs)**.** *Nechť* $\boldsymbol{Y}$ *je* p*-rozměrný náhodný vektor,* $g \in \mathbb{N}$, $\pi_j \in (0,1)$, $j \in \{1, \ldots, g\}$*, tak, že*

$$\sum_{j=1}^{g} \pi_j = 1$$

a nechť $f_j(\boldsymbol{y})$, $j \in \{1, \ldots, g\}$*, jsou hustoty*[1] *na* $\mathbb{R}^p$*. Nechť* $\boldsymbol{Y}$ *má hustotu*

$$f(\boldsymbol{y}) = \sum_{j=1}^{g} \pi_j f_j(\boldsymbol{y}), \qquad \boldsymbol{y} \in \mathbb{R}^p. \tag{1.1}$$

Potom řekneme, že rozdělení náhodného vektoru $\boldsymbol{Y}$ *je* g-složková konečná směs. *Hustotě* (1.1) *budeme říkat* hustota směsi s g složkami.

Přirozené číslo g představuje počet složek ve směsi, a ačkoliv v definici připouštíme i $g = 1$, ve skutečnosti tento případ není v definici možný, z důvodu podmínek, které klademe na π_j. Nás beztak ale zajímá pouze netriviální případ kdy $g > 1$. Hodnoty π_j vyjadřují poměr zastoupení j-té složky ve směsi a uvažujeme je v otevřeném intervalu (0,1). To proto, že pokud by pro nějaké j platilo $\pi_j = 1$, pak by šlo o 1-složkovou konečnou směs, což je případ, který nás nezajímá. Naopak, v případě kdy by $\pi_j = 0$, pak by se jednalo o $g - 1$-složkovou směs, kterou popíšeme $g - 1$ složkami.

[1] Hustoty v této práci uvažujeme vzhledem k p-rozměrné Lebesgueově míře.

1.2 Interpretace směsi

Nyní si uvedeme jednu z možných interpretací g-složkové konečné směsi. Uvažme náhodnou veličinu U nabývající hodnot $1, \ldots, g$ s pravděpodobnostmi

$$\mathsf{P}(U = j) = \pi_j, \qquad j = 1, \ldots, g.$$

Dále předpokládejme, že podmíněná hustota náhodného vektoru $\boldsymbol{Y}$ při jevu $U = j$ je $f_j(\boldsymbol{y})$. Pak nepodmíněná hustota náhodného vektoru $\boldsymbol{Y}$ je

$$f(\boldsymbol{y}) = \sum_{j=1}^{g} f(\boldsymbol{y}|U = j)\,\mathsf{P}(U = j) = \sum_{j=1}^{g} \pi_j f_j(\boldsymbol{y}),$$

což je přesně hustota (1.1). Tedy rozdělení konečné směsi $\boldsymbol{Y}$, jenž je hustotou (1.1) reprezentováno, lze interpretovat jako chování sledované vlastnosti v g skupinách, kde v každé skupině se vlastnost chová jinak, neboli má jiné rozdělení dané hustotami $f_1, \ldots, f_g$.

Podobně lze místo náhodné veličiny U uvážit g-rozměrný náhodný vektor $\boldsymbol{Z} = (Z_1, \ldots, Z_g)^\top$, jehož složky nabývají pouze nul a jedné jedničky, přitom jedničky právě v té složce, jejíž index odpovídá skupině, ze které sledovaný prvek pochází. Tj.

$$\boldsymbol{Z} \sim \mathrm{Mult}(1, \pi_1, \ldots, \pi_g).$$

Předpokládejme, že pro podmíněnou hustotu $\boldsymbol{Y}$ při $\boldsymbol{Z} = \boldsymbol{z}$ platí

$$f(\boldsymbol{y}|\boldsymbol{z}) = \sum_{j=1}^{g} z_j f_j(\boldsymbol{y}) = \prod_{j=1}^{g} \left(f_j(\boldsymbol{y})\right)^{z_j}.$$

Pak nepodmíněná hustota $\boldsymbol{Y}$, analogicky s předchozím příkladem, je dána hustotou (1.1).

Typicky ale náhodný vektor $\boldsymbol{Z}$, respektive hodnotu náhodné veličiny U, které popisují, z jaké skupiny sledované populace sledovaný objekt s pozorovanou hodnotou $\boldsymbol{Y} = \boldsymbol{y}$ pocházel, nepozorujeme. S využitím Bayesovy věty však zpětně můžeme počítat pravděpodobnost toho, že sledovaný objekt náležel do j-té skupiny, poté co byla pozorována hodnota $\boldsymbol{Y} = \boldsymbol{y}$.

$$\begin{aligned} p_j(\boldsymbol{y}) &= \mathsf{P}(\text{objekt náležel do } j\text{-té složky}|\boldsymbol{y}) \\ &= \mathsf{P}(Z_j = 1|\boldsymbol{y}) \\ &= \pi_j f_j(\boldsymbol{y})/f(\boldsymbol{y}), \qquad j = 1, \ldots, g. \end{aligned} \tag{1.2}$$

1.3 Parametrizace směsi

Dále se budeme zabývat odhadováním parametrů modelu konečné směsi, a k tomu budeme potřebovat hustotu směsi v řeči nějakých parametrů vyjádřit. Ačkoliv je možné, aby se sledovaná vlastnost chovala v různých skupinách úplně odlišně, my se v této práci omezíme jen na konečné směsi, jejíž složky pocházejí ze stejné parametrické rodiny.

Definice 2 (Parametrická směs)**.** *Nechť* $\boldsymbol{Y}$ *je* p*-rozměrný náhodný vektor,* $g \in \mathbb{N}$, $\pi_j \in (0,1)$, $j \in \{1,\ldots,g\}$*, tak, že*

$$\sum_{j=1}^{g} \pi_j = 1$$

a nechť $f(\boldsymbol{y};\boldsymbol{\theta}_j), j \in \{1,\ldots,g\}$*, jsou hustoty na* $\mathbb{R}^p$ *s neznámými parametry obsaženými ve vektoru* $\boldsymbol{\theta}_j \in \Theta_j \subset \mathbb{R}^{p_j}$*. Nechť* $\boldsymbol{Y}$ *má hustotu*

$$f(\boldsymbol{y};\boldsymbol{\psi}) = \sum_{j=1}^{g} \pi_j f(\boldsymbol{y};\boldsymbol{\theta}_j). \tag{1.3}$$

Potom řekneme, že rozdělení náhodného vektoru $\boldsymbol{Y}$ *je* g-složková parametrická směs. *Hustotě* (1.3) *budeme říkat* parametrická hustota směsi s g složkami.

Vektor $\boldsymbol{\psi}$ nechť označuje vektor obsahující všechny neznámé parametry modelu. Symbolem Ψ budeme značit parametrický prostor pro $\boldsymbol{\psi}$. Označíme $\boldsymbol{\xi}$ vektor obsahující všechny parametry v $\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_g$, o kterých připouštíme, že mohou být různé. Pak lze psát

$$\boldsymbol{\psi} = \left(\pi_1,\ldots,\pi_{g-1},\boldsymbol{\xi}^\top\right)^\top.$$

V zápise je záměrně opomenut poměr zastoupení π_g směsi g. To proto, že součet π_j je jedna, a tudíž je jeden z nich nadbytečný.

Značení. Hustotu p-dimenzionálního normálního náhodného vektoru $\boldsymbol{Y}$ se střední hodnotou $\boldsymbol{\mu}$ a kovarianční maticí $\boldsymbol{\Sigma}$ budeme značit $\varphi(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma})$. Analogicky symbolem $\varphi(y;\mu,\sigma^2)$ budeme značit hustotu jednorozměrného normálního rozdělení se střední hodnotou μ a rozptylem σ^2.

Definice 3 (Normální směs)**.** *Nechť* $\boldsymbol{Y}$ *je* p*-rozměrný náhodný vektor,* $g \in \mathbb{N}$, $\pi_j \in (0,1)$, $j \in \{1,\ldots,g\}$*, tak, že*

$$\sum_{j=1}^{g} \pi_j = 1$$

a nechť $\varphi(\boldsymbol{y};\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)$, $j \in \{1,\ldots,g\}$*, jsou hustoty* p*-rozměrného normálního rozdělení s parametry* $\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j$*. Nechť* $\boldsymbol{Y}$ *má hustotu*

$$f(\boldsymbol{y};\boldsymbol{\psi}) = \sum_{j=1}^{g} \pi_j \varphi(\boldsymbol{y};\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j). \tag{1.4}$$

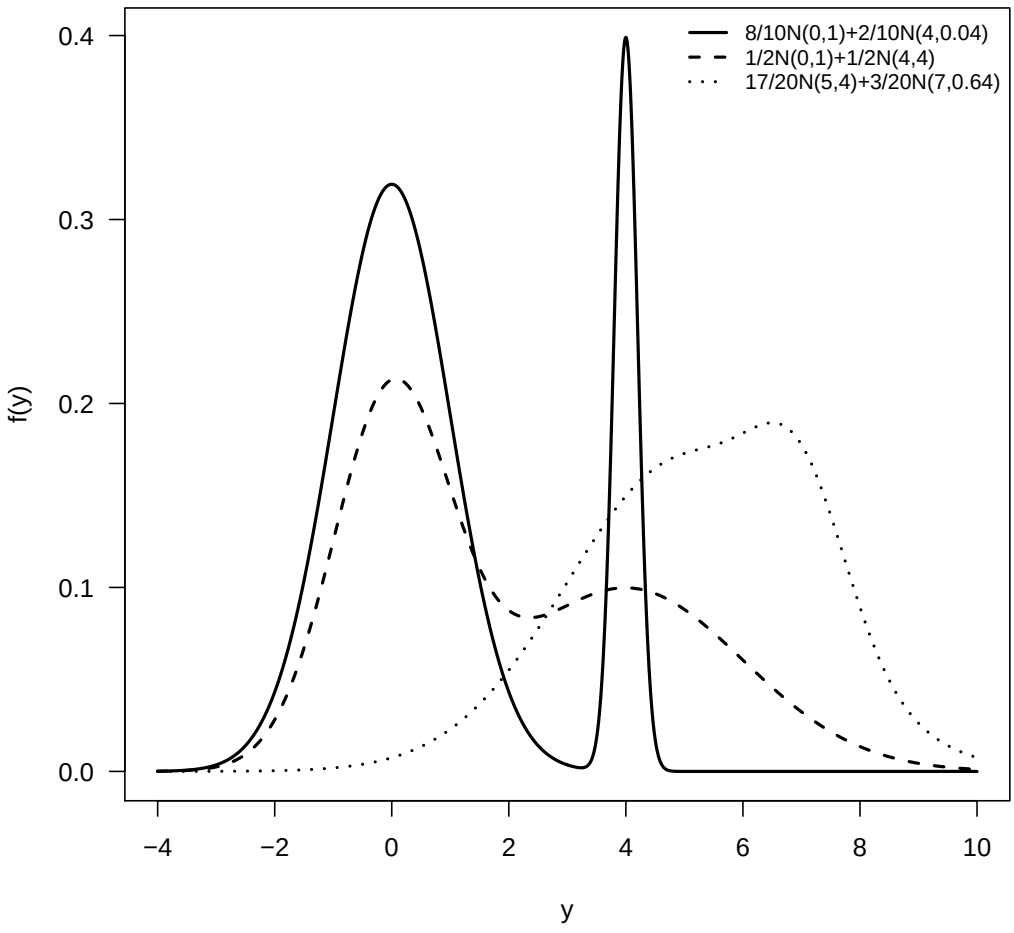
Potom řekneme, že rozdělení náhodného vektoru $\boldsymbol{Y}$ *je* g-složková normální směs. *Hustotě* (1.4) *budeme říkat* parametrická hustota normální směsi s g složkami. *V případě, kdy* $p = 1$ *budeme mluvit o* g-složkové jednorozměrné normální směsi.

Uvědomíme si, že pokud použijeme pro normální směs zápis 1.3, pak prvky parametrického vektoru $\boldsymbol{\theta}_j$ jsou všechny parametry normálního rozdělení, u kterých připouštíme, že jsou různé. Typicky se bude jednat zejména o následující dva typy.

Příklad. Heteroskedastickou normální směsí myslíme směs, ve které připouštíme, že rozptylové matice složek mohou být různé. Pak v $\boldsymbol{\theta}_j$ jsou obsaženy parametry vektoru středních hodnot $\boldsymbol{\mu}_j$ a prvky matice $\boldsymbol{\Sigma}_j$ obsažené v jejím horním trojúhelníku. Parametrický vektor $\boldsymbol{\psi}$ pak čítá všechny tyto parametry dohromady pro $j = 1, \ldots, g$ společně s poměry složek π_j.

Příklad. Homoskedastická normální směs je taková, ve které naopak požadujeme, aby rozptylové matice složek byly shodné. Potom v $\boldsymbol{\theta}_j$ jsou společně s prvky parametry vektoru středních hodnot $\boldsymbol{\mu}_j$ i prvky určující společnou rozptylovou matici $\boldsymbol{\Sigma}$. Vektor $\boldsymbol{\psi}$ pak obsahuje krom prvků $\boldsymbol{\mu}_j$ a poměrů složek $\pi_j, j = 1, \ldots, g$ i prvky matice $\boldsymbol{\Sigma}$, které jsou v jejím horním trojúhelníku, a to právě jednou.

Pouze poznamenejme, že pomocí normálních směsí lze kromě jiného aproximovat širokou škálu rozdělení, jak je empiricky ukázáno na obrázku 1.1, kde jsou navíc uvedeny pouze dvousložkové jednorozměrné normální směsi. Všechny se ale od sebe znatelně liší. V legendě obrázku symbolem $\pi N(\mu_1,\sigma_1^2) + (1-\pi)N(\mu_2,\sigma_2^2)$ myslíme směs, kde jsou normální rozdělení definovaná podle příslušných parametrů μ_1, σ_1^2 resp. μ_2, σ_2^2 s poměry zastoupení složek π a $1 - \pi$.



Obrázek 1.1: Příklady několika normálních směsí.

1.4 Příklad onemocnění paradontózou

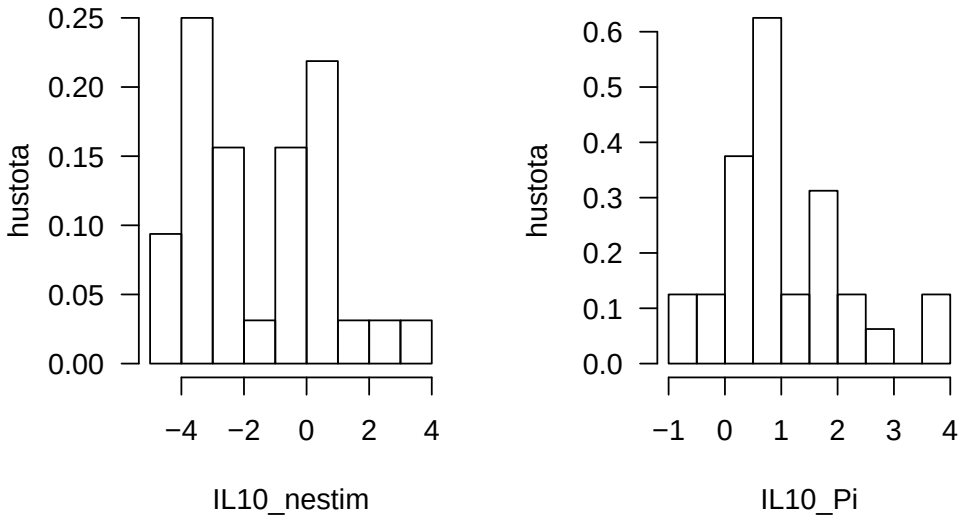
Pro lepší pochopení modelu směsi si zde uvedeme následující příklad, na který budeme v dalších částech práce navazovat. Uvažme populaci lidí, ve které je rozšířena paradontóza, odborně paradontitida. Toto mezi lidmi relativně běžné zánětlivé onemocnění může skončit až ztrátou zubů. Jeho včasná detekce pak často zachrání mnoha lidem jak chrup, tak i velké množství peněz, které musí být na případnou náhradu vynaloženy.

Data diskutovaná v této bakalářské práci jsou výsledkem měření produkce cytokinu *interleukin 10* (dále IL10), což je protein, který se významně účastní v imunitní odpovědi těla na případný útok nějaké nemoci. Jeho zvýšená produkce tedy může být signálem nastupujícího onemocnění. Při měření pak kultura obsahující tento cytokin v prvním případě nebyla nijak stimulována a byla měřena jeho samotná ničím nezrychlená produkce a v druhém případě byla stimulována přítomností bakterie *prevotella intermedia*. To je jedna z těch, které se běžně vyskytují v dutině ústní, a má významný podíl na vzniku paradontitidy.

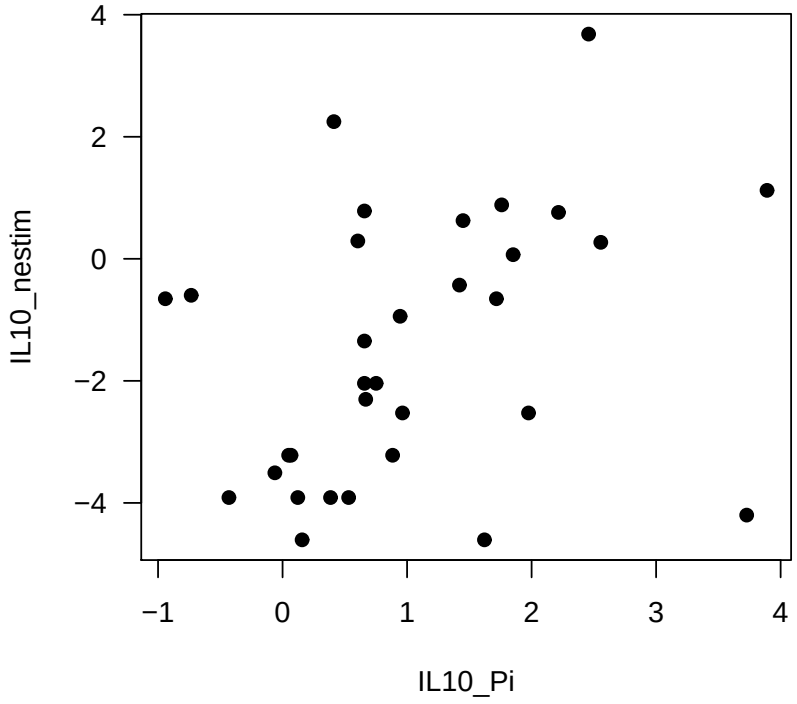
V rámci statistické analýzy nás bude zajímat zejména odpověď na otázku, zda-li je produkce cytokinu IL10 u pacientů s paradontózou vyšší než u lidí, kteří touto nemocí netrpí. Budeme předpokládat, že produkce se v každé z těchto dvou skupin řídí jiným rozdělením, které se pokusíme odhadnout, a na závěr pak budeme zkoumat, zda-li data odpovídají patřičné směsi.

Označíme $\boldsymbol{X}$ dvourozměrný náhodný vektor, jehož první složka X_1 reprezentuje nestimulovanou produkci cytokinu IL10 a druhá složka X_2 produkci IL10 za přítomnosti bakterie *prevotella intermedia*. Pokud budeme předpokládat, že růstové kultury se v každé ze dvou uvažovaných skupin řídí log-normálním rozdělením, pak zřejmě náhodný vektor $\boldsymbol{Y} = \log \boldsymbol{X} = (\log X_1, \log X_2)^\top$, bude mít při platnosti naší hypotézy rozdělení dvousložkové dvourozměrné normální směsi.

Pro přehlednost a ilustraci si zde uvedeme ještě histogramy logaritmizovaných hodnot jak nestimulované, tak stimulované produkce cytokinu IL10, (viz obrázek 1.2). Popiskem *IL10_nestim* je označen histogram hodnot nestimulované produkce cytokinu IL10, podobně *IL10_Pi* jsou označeny hodnoty produkce IL10 při přítomnosti bakterie *prevotella intermedia*. Pro celkový náhled, jak vypadají naše data dvojrozměrně, ještě uvádíme bodový graf naměřených hodnot (viz obrázek 1.3). Dva shluky okolo bodů o souřadnicích $[0, -4]$ a [2,0] v tomto grafu naznačují, že rozdělení by mohlo odpovídat dvousložkové normální směsi. Otázkou zůstává, jaké parametry tato směs má. Jeden z možných způsobů jejich odhadnutí ukážeme v následující kapitole.



Obrázek 1.2: Histogramy hodnot logaritmizované produkce stimulované i nestimulované IL10.



Obrázek 1.3: Bodový zákres dat produkce cytokinu IL10.

Kapitola 2

Odhad metodou maximální věrohodnosti

Existuje několik možností, kterých lze použít pro odhadování parametrů modelu s rozdělením určeným hustotou (1.3). Zmínit můžeme například momentovou metodu, kterou použil Karl Pearson, případně třeba metodu nejmenší vzdálenosti. Těmi se ale v tomto textu zabývat nebudeme a místo toho se budeme soustředit na odhadování pomocí metody maximální věrohodnosti, která je obecně jednou z nejvyužívanějších metod odhadu. Ta spočívá v sestavení věrohodnostní funkce a její následné maximalizaci. V této kapitole mimo jiné rozebereme i úskalí, která na nás při použití této metody v kontextu směsí čekají.

Pro potřeby této práce budeme značit $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ náhodný výběr o rozsahu n ze spojitého rozdělení s pravděpodobnostní hustotou $f(\boldsymbol{y})$, $\boldsymbol{y} \in \mathbb{R}^p$. Pak tedy je $\boldsymbol{Y}_i$, $i \in \{1, \ldots, n\}$, p-rozměrný náhodný vektor. Navíc označíme $\boldsymbol{Y}_a = \left(\boldsymbol{Y}_1^\top, \ldots, \boldsymbol{Y}_n^\top\right)^\top$ náhodný vektor reprezentující celý výběr. $\boldsymbol{Y}_a$ je tedy $(n{\cdot}p)$-rozměrný náhodný vektor. Realizace náhodných vektorů budeme značit odpovídajícím malým písmenem. Tedy symbolem $\boldsymbol{y}_1$ budeme značit realizaci náhodného vektoru $\boldsymbol{Y}_1$, podobně symbolem $\boldsymbol{y}_a$ budeme značit realizaci celého náhodného výběru.

V kontextu konečných směsí a při zachování značení z kapitoly 1 označíme pro náhodný vektor $\boldsymbol{Y}_i$ jeho příslušný náhodný vektor $\boldsymbol{Z}_i$ určující, ze které skupiny měřený prvek pochází. Navíc ještě zavedeme značení pro následující náhodné vektory $\boldsymbol{Z}_a = \left(\boldsymbol{Z}_1^\top, \ldots, \boldsymbol{Z}_n^\top\right)^\top$ a $\boldsymbol{Y}_c = \left(\boldsymbol{Y}_a^\top, \boldsymbol{Z}_a^\top\right)^\top$.

2.1 Věrohodnostní funkce

Pro náhodný výběr $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ z rozdělení s hustotou (1.3) sestavíme věrohodnostní funkci jakožto sdruženou hustotu tohoto náhodného výběru. Potom pro $\boldsymbol{\psi} \in \Psi$ má věrohodnostní funkce tvar

$$L(\boldsymbol{\psi}) = \prod_{i=1}^{n} f(\boldsymbol{Y}_i; \boldsymbol{\psi}) = \prod_{i=1}^{n} \left(\sum_{j=1}^{g} \pi_j f(\boldsymbol{Y}_i; \boldsymbol{\theta}_j) \right).$$

Analogicky můžeme sestavit i log-věrohodnostní funkci

$$\ell(\boldsymbol{\psi}) = \log L(\boldsymbol{\psi}) = \sum_{i=1}^{n} \log f(\boldsymbol{Y}_i; \boldsymbol{\psi}) = \sum_{i=1}^{n} \log \left(\sum_{j=1}^{g} \pi_j f(\boldsymbol{Y}_i; \boldsymbol{\theta_j}) \right). \tag{2.1}$$

Odhad metodou maximální věrohodnosti (dále budeme značit MMV) je potom prvek $\hat{\boldsymbol{\psi}}_{\mathrm{MLE}} \in \Psi$ splňující

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

Je zřejmé, že aby tato rovnice měla řešení, musí být věrohodnostní funkce omezená. Tento požadavek ale, jak ukazuje následující příklad, není v případě směsí vždy splněn.

Příklad. Uvažme jednorozměrnou normální směs s dvěma komponentami. Nechť tedy

$$f(y; \boldsymbol{\psi}) = \pi_1 \varphi(y; \mu_1, \sigma_1^2) + (1 - \pi_1)\varphi(y; \mu_2, \sigma_2^2).$$

Pak $\boldsymbol{\psi} = (\pi_1, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)^\top$. Položme $\Psi = (0,1)\times\mathbb{R}\times(0,\infty)\times\mathbb{R}\times(0,\infty)$. Nechť $Y_1, \ldots, Y_n$ je náhodný výběr z rozdělení s touto hustotu. Pak log-věrohodnostní funkce pro odhad parametrického vektoru $\boldsymbol{\psi}$ při realizovaném náhodném výběru $Y_1 = y_1, \ldots, Y_n = y_n$ je tvaru

$$\ell(\boldsymbol{\psi}) = \sum_{i=1}^{n} \log\Big(\pi_1 \varphi(y_i; \mu_1, \sigma_1^2) + (1 - \pi_1)\varphi(y_i; \mu_2, \sigma_2^2) \Big). \tag{2.2}$$

Položme $\boldsymbol{\psi}_\delta = (\frac{1}{2}, y_1, \delta^2, 0, 1)^\top, \delta > 0$. Pak $\forall \delta > 0 : \boldsymbol{\psi}_\delta \in \Psi$. Pro toto $\boldsymbol{\psi}_\delta$ nyní vyjádříme log-věrohodnostní funkci (2.2)

$$\ell(\boldsymbol{\psi}_\delta) = \sum_{i=1}^{n} \log \left(\frac{1}{2} \frac{1}{\sqrt{2\pi}\delta} \exp\left(-\frac{(y_i - y_1)^2}{2\delta^2} \right) + \frac{1}{2} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y_i^2}{2} \right) \right). \tag{2.3}$$

Nyní hodnotu $\ell(\boldsymbol{\psi}_\delta)$ odhadneme zespoda tak, že pro $i = 1$ vynecháme druhý člen součtu v argumentu logaritmu a pro ostatní i vynecháme první člen. Tím hodnotu na pravé straně (2.3) snížíme, protože oba sčítance jsou vždy kladné. Tedy

$$\begin{aligned}
\ell(\boldsymbol{\psi}_\delta) &> \log \left(\frac{1}{2} \frac{1}{\sqrt{2\pi}\delta} \exp\left(-\frac{(y_1 - y_1)^2}{2\delta^2} \right) \right) + \sum_{i=2}^{n} \log \left(\frac{1}{2} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-y_i^2}{2} \right) \right) \\
&> -\log\left(2\sqrt{2\pi}\delta \right) - (n-1)\log\left(2\sqrt{2\pi} \right) + \sum_{i=2}^{n} \frac{-y_i^2}{2} \\
&> -\log(\delta) - n\log\left(2\sqrt{2\pi} \right) - \sum_{i=2}^{n} \frac{y_i^2}{2}.
\end{aligned} \tag{2.4}$$

Nyní je z (2.4) patrné, jak se $\ell(\boldsymbol{\psi}_\delta)$ chová když $\delta \to 0$. Jinými slovy

$$\lim_{\delta\to 0} \ell(\boldsymbol{\psi}_\delta) > \lim_{\delta\to 0} \left(-\log(\delta) - n\log\left(2\sqrt{2\pi} \right) - \sum_{i=2}^{n} \frac{y_i^2}{2} \right) = \infty.$$

Z toho vyplývá, že věrohodnostní funkce této směsi není omezená, a tedy neexistuje odhad metodou maximální věrohodnosti. Navíc tento postup lze zobecnit pro vícerozměrné, vícesložkové i mnohé další nenormální směsi prostým umístěním střední hodnoty jedné složky do jednoho pozorování a uvážením rozptylu blížícího se k nule, případně rozptylové matice k singulární. Maximálně věrohodné odhady tedy v podstatě odpovídají směsi s degenerovanými komponentami.

❑

Hustota složená z nekonečných výběžků v místech pozorovaných dat by ale nepředstavovala jejich žádně zjednodušení, které od modelu čekáme. Aitkin (2001, str. 289) navrhuje způsob, kterým lze problém neomezené věrohodnosti řešit. Stačilo by omezit parametrický prostor Ψ tak, aby byl uzavřený a omezený. Toho se dá docílit například určením dolní hranice δ pro směrodatné odchylky složek směsi. Nicméně takovéto opatření způsobí že výsledný odhad bude na δ závislý, což může vadit zejména v případě mnoha odlehlých pozorování.

Na druhou stranu v případě modelu jednorozměrné homoskedastické normální směsi, tj. směsi, jejichž všechny složky mají stejný rozptyl, lze ukázat, že věrohodnostní funkce, v případě dostatečně velkého náhodného výběru, omezená je. To bude předmětem následující věty.

Věta 1. *Nechť $Y_1, \ldots, Y_n$ je náhodný výběr obsahující alespoň g+1 různých pozorování z jednorozměrné normální směsi s hustotou*

$$f\left(y; \boldsymbol{\psi}\right) = \sum_{j=1}^{g} \pi_j \varphi(y; \mu_j, \sigma^2). \tag{2.5}$$

Pak je věrohodnostní funkce $L(\boldsymbol{\psi})$ shora omezená.

Důkaz. Uvažme maximální možný parametrický prostor který pro normální směs (2.5) připadá v úvahu, tj $\mu_j \in \mathbb{R}$, $\sigma^2 \in (0{,}\infty)$, $\pi_j \in (0{,}1)$, $j = 1, \ldots, g$. Pro pozorovaný náhodný výběr $Y_1 = y_1, \ldots, Y_n = y_n$ vyjádříme a shora odhadneme věrohodnostní funkci náhodného výběru:

$$\begin{aligned} L(\boldsymbol{\psi}) &= \prod_{i=1}^{n} \left(\sum_{j=1}^{g} \pi_j \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2} (y_i - \mu_j)^2 \right) \right) \\ &\leq \prod_{i=1}^{n} \left(\sum_{j=1}^{g} \frac{1}{\sigma} \exp\left(-\frac{1}{2\sigma^2} (y_i - \mu_j)^2 \right) \right). \end{aligned} \tag{2.6}$$

Dále využijeme faktu, že máme více různých pozorování než složek směsi, tudíž

$$\exists \varepsilon > 0, \exists i_0 \in \{1, \ldots, n\} : \forall j \in \{1, \ldots, g\} \ \ |y_{i_0} - \mu_j| > \varepsilon.$$

Bez újmy na obecnosti položme $i_0 = 1$. Dále roztrhneme součin v (2.6) na dvě části, nejprve odhadneme příspěvek do věrohodnostní funkce pro $i = 1$, a poté příspěvek součinu zbytku,

$$\sum_{j=1}^{g} \frac{1}{\sigma} \exp\left(-\frac{1}{2\sigma^2}(y_i - \mu_j)^2\right) < \sum_{j=1}^{g} \frac{1}{\sigma} \exp\left(-\frac{1}{2\sigma^2}\varepsilon^2\right) = \frac{g}{\sigma} \exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right),$$
$$\prod_{i=2}^{n} \left(\sum_{j=1}^{g} \frac{1}{\sigma} \exp\left(-\frac{(y_i - \mu_j)^2}{2\sigma^2}\right)\right) < \prod_{i=2}^{n} \left(\sum_{j=1}^{g} \frac{1}{\sigma} \exp\left(-\frac{0^2}{2\sigma^2}\right)\right) = \left(\frac{g}{\sigma}\right)^{n-1}. \tag{2.7}$$

Nyní dosadíme rovnice (2.7) do (2.6) a vyjádříme:

$$L(\boldsymbol{\psi}) < \frac{g}{\sigma} \exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right) \cdot \left(\frac{g}{\sigma}\right)^{n-1} = \left(\frac{g}{\sigma}\right)^{n} \exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right). \tag{2.8}$$

Položme

$$f(\sigma) = \left(\frac{g}{\sigma}\right)^{n} \exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right), \qquad \sigma \in (0,\infty).$$

Nyní si stačí uvědomit, že funkce $f(\sigma)$ je spojitá na $(0,\infty)$, a navíc

$$\lim_{\sigma\to\infty} f(\sigma) = 0 \qquad \& \qquad \lim_{\sigma\to 0} f(\sigma) = 0,$$

z čehož je patrné, že $f(\sigma)$ je shora omezená, a tudíž z (2.8) je i věrohodnostní funkce $L(\boldsymbol{\psi})$ shora omezená.

❑

2.2 Věrohodnostní rovnice

Je patrné, že odhad metodou maximální věrohodnosti často nemusí existovat a i v případě, kdy řešení existuje, nemusí být vyhovující. Místo maximalizace věrohodnostní funkce se však můžeme zaměřit na hledání řešení rovnice

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

případně ekvivalentně na hledání řešení rovnice

$$\partial \ell(\boldsymbol{\psi})/\partial\boldsymbol{\psi} = \mathbf{0}. \tag{2.9}$$

Rovnici (2.9) říkáme věrohodnostní rovnice. Je zřejmé, že pokud existuje řešení $\hat{\boldsymbol{\psi}}_{\mathrm{MLE}} \in \mathrm{int}(\Psi)$, pak je řešením (2.9). Naopak pokud najdeme prvek $\hat{\boldsymbol{\psi}} \in \Psi$ splňující (2.9), které je lokálním maximem věrohodnostní funkce (2.1), můžeme ho považovat za kandidáta na odhad metodou maximální věrohodnosti. Dokonce ho má smysl za jistých okolnostní jako odhad MMV přijmout, i přesto, že není globálním maximem. Lze totiž ukázat, že má podobné vlastnosti jako odhad MMV. My se touto teorií v práci zabývat nebudeme, více na toto téma je řečeno například v McLachlan a Peel (2000, kapitola 2.2).

2.3 Identifikovatelnost rozdělení směsi

Nyní se budeme věnovat jinému problému, který vyvstává při odhadování modelů konečných směsi, a to je problém identifikovatelnosti. Jinak řečeno, chceme odhadnout rozdělení náhodné veličiny, které je reprezentováno její hustotou závisející na parametrickém vektoru. Má tedy smysl si klást podmínku, aby různé hodnoty parametrických vektorů v našem parametrickém prostoru určovaly jiná rozdělení. Přesněji řečeno, aby platilo $\forall \boldsymbol{\psi}_1, \boldsymbol{\psi}_2 \in \Psi$:

$$f(\boldsymbol{y}; \boldsymbol{\psi}_1) = f(\boldsymbol{y}; \boldsymbol{\psi}_2) \text{ s.v.} \Leftrightarrow \boldsymbol{\psi}_1 = \boldsymbol{\psi}_2.$$

Jak se ale ukazuje, v kontextu směsí toto nemusí být pravda.

Příklad. Uvažme homoskedastickou normální směs o dvou složkách tak, že

$$f(y; \boldsymbol{\psi}) = \pi_1 \varphi(y; \mu_1, \sigma^2) + (1 - \pi_1)\varphi(y; \mu_2, \sigma^2).$$

Pak $\boldsymbol{\psi} = (\pi_1, \mu_1, \mu_2, \sigma^2)^\top$. Při přirozené volbě $\Psi = (0,1) \times \mathbb{R}^2 \times (0, \infty)$, položme $\boldsymbol{\psi}_1 = (0.3, 1, 5, 1)^\top$ a $\boldsymbol{\psi}_2 = (0.7, 5, 1, 1)^\top$. Zřejmě $\boldsymbol{\psi}_1, \boldsymbol{\psi}_2 \in \Psi, \boldsymbol{\psi}_1 \neq \boldsymbol{\psi}_2$ a přitom pro s.v. $y \in \mathbb{R}$

$$f(y; \boldsymbol{\psi}_1) = 0.3 \cdot \varphi(y; 1, 1) + 0.7 \cdot \varphi(y; 5, 1) = 0.7 \cdot \varphi(y; 5, 1) + 0.3 \cdot \varphi(y; 1, 1) = f(y; \boldsymbol{\psi}_2).$$

Při takovéto volbě parametrů došlo pouze k prohození první a druhé složky, což výslednou hustotu neovlivní.

❑

Zřejmě v případě g složek existuje $g!$ permutací indexů složek a tudíž i $g!$ parametrických vektorů $\boldsymbol{\psi}$, které reprezentují totožné rozdělení (za předpokladu že všechny jsou prvkem parametrického prostoru).

Definice 4 (Identifikovatelnost rozdělení v modelu směsi)**.** *Řekneme, že model směsi* $\{f(\boldsymbol{y}; \boldsymbol{\psi}) : \boldsymbol{\psi} \in \Psi\}$ *je* identifikovatelný*, pokud* $\forall \boldsymbol{\psi}, \boldsymbol{\psi}' \in \Psi$ *takové, že*

$$f(\boldsymbol{y}; \boldsymbol{\psi}) = \sum_{j=1}^{g} \pi_j f(\boldsymbol{y}; \boldsymbol{\theta}_j), \qquad f(\boldsymbol{y}; \boldsymbol{\psi}') = \sum_{j=1}^{g'} \pi'_j f(\boldsymbol{y}; \boldsymbol{\theta}'_j)$$

a

$$f(\boldsymbol{y}; \boldsymbol{\psi}) = f(\boldsymbol{y}; \boldsymbol{\psi}') \text{ s.v.}$$

platí, že $g = g'$ *a existuje permutace indexů* τ *složek směsi tak, že*

$$\pi_j = \pi'_{\tau(j)}, \qquad \boldsymbol{\theta}_j = \boldsymbol{\theta}'_{\tau(j)}, \qquad j = 1, \ldots, g.$$

V případě, kdy budeme diskutovat výsledky nějakých parametrických odhadů, budeme diskutovat případ pouze pro jedno uspořádání směsi. Kde to bude možné, budeme směsi řadit buď dle velikosti jejich poměru zastoupení případně podle velikosti první složky střední hodnoty. Na závěr bych chtěl ještě dodat, že tento fenomén nám nebude činit žádné obtíže při počítání odhadu MMV EM algoritmem, který se pro výpočet modelu směsi běžně používá, a bude představen v následující části.

2.4 EM algoritmus

Jedna z možností, jak hledat odhady metodou maximální věrohodnosti, je pomocí EM algoritmu. V této sekci se nejdříve budeme věnovat zavedení EM algoritmu pro směsi, vyjasníme si, proč a jak funguje, a dokážeme si základní vlastnost, která zaručuje, že po jedné iteraci EM algoritmu dostaneme „lepší“ odhad parametrů, než jsme doposud měli. V další části pak pro obecný model směsi krok algoritmu pro část parametrů vyjádříme.

Nechť $g \in \mathbb{N}, \pi_j \in (0,1)$, $j \in \{1,\dots,g\}$, tak, že

$$\sum_{j=1}^{g} \pi_j = 1$$

a nechť $f(\boldsymbol{y};\boldsymbol{\theta}_j)$, $j \in \{1,\dots,g\}$, jsou hustoty na $\mathbb{R}^p$ s neznámými parametry obsaženými ve vektoru $\boldsymbol{\theta}_j$. Mějme náhodný výběr $\boldsymbol{Y}_1,\dots,\boldsymbol{Y}_n$ z rozdělení s hustotou

$$f(\boldsymbol{y};\boldsymbol{\psi}) = \sum_{j=1}^{g} \pi_j f(\boldsymbol{y};\boldsymbol{\theta}_j).$$

Nechť

$$\boldsymbol{Z}_i \sim \mathrm{Mult}(1,\pi_1,\dots,\pi_g), \qquad i = 1,\dots,n,$$

jsou náhodné vektory určující, z jaké skupiny populace objekt pochází. Připomeňme si, že značíme $\boldsymbol{Y}_a$ celý náhodný výběr, podobně také $\boldsymbol{Z}_a = \left(\boldsymbol{Z}_1^\top,\dots,\boldsymbol{Z}_n^\top\right)^\top$ a $\boldsymbol{Y}_c = \left(\boldsymbol{Y}_a^\top,\boldsymbol{Z}_a^\top\right)^\top$.

Myšlenka, proč se EM algoritmus využívá pro počítání odhadů MMV v případech směsí, je jednoduchá. V zásadě je totiž velmi obtížné hledat maxima log-věrohodnostní funkce (2.1), kterou navíc ani při speciálních případech konkrétních hustot nelze nějak jednoduše maximalizovat. EM algoritmus tento problém tzv. nekompletních dat $\boldsymbol{Y}_a$, kdy neznáme příslušné vektory určující, ze kterých skupin naměřené objekty pocházely, převádí na hledání maxima věrohodnostní funkce odpovídající tzv. kompletním datům $\boldsymbol{Y}_c$.

Uvažme nyní hustoty těchto náhodných vektorů:

$$f(\boldsymbol{Y}_c;\boldsymbol{\psi}) = f(\boldsymbol{Y}_a,\boldsymbol{Z}_a;\boldsymbol{\psi}) = f(\boldsymbol{Y}_c|\boldsymbol{Y}_a;\boldsymbol{\psi}) f(\boldsymbol{Y}_a;\boldsymbol{\psi}),$$

$$\log f(\boldsymbol{Y}_a;\boldsymbol{\psi}) = \log f(\boldsymbol{Y}_c;\boldsymbol{\psi}) - \log f(\boldsymbol{Y}_c|\boldsymbol{Y}_a;\boldsymbol{\psi}). \tag{2.10}$$

Pokud vyjádříme

$$f(\boldsymbol{Y}_c;\boldsymbol{\psi}) = \prod_{i=1}^{n}\prod_{j=1}^{g}\Big(\pi_j f(\boldsymbol{Y}_i;\boldsymbol{\theta}_j)\Big)^{Z_{ij}}, \qquad f(\boldsymbol{Y}_a;\boldsymbol{\psi}) = \prod_{i=1}^{n}\sum_{j=1}^{g}\pi_j f(\boldsymbol{Y}_i;\boldsymbol{\theta}_j)$$

a dosadíme do (2.10) získáme rovnici ve tvaru

$$\ell(\boldsymbol{\psi}) = \ell_c(\boldsymbol{\psi}) - \log f(\boldsymbol{Y}_c|\boldsymbol{Y}_a;\boldsymbol{\psi}), \tag{2.11}$$

kde $\ell_c(\boldsymbol{\psi})$ je věrohodnostní funkce odpovídající kompletním datům $\boldsymbol{Y}_c$, tj.

$$\ell_c(\boldsymbol{\psi}) = \log\left(\prod_{i=1}^{n}\prod_{j=1}^{g}\Big(\pi_j f(\boldsymbol{Y}_i;\boldsymbol{\theta}_j)\Big)^{Z_{ij}}\right) = \sum_{i=1}^{n}\sum_{j=1}^{g} Z_{ij}\Big(\log\pi_j + \log f(\boldsymbol{Y}_i,\boldsymbol{\theta}_j)\Big). \tag{2.12}$$

Nyní uvažme pozorovaný náhodný výběr $\boldsymbol{Y}_a = \boldsymbol{y}_a$ a spočtěme střední hodnoty v rovnosti (2.11) za podmínky, že známe $\boldsymbol{y}_a$ a $\boldsymbol{\psi}'$. Pak s využitím, že

$$\mathsf{E}\left[\ell(\boldsymbol{\psi})|\boldsymbol{y}_a,\boldsymbol{\psi}'\right] = \mathsf{E}\left[\sum_{i=1}^{n}\log f(\boldsymbol{y}_i;\boldsymbol{\psi})\middle|\boldsymbol{y}_a,\boldsymbol{\psi}'\right] = \sum_{i=1}^{n}\log f(\boldsymbol{y}_i;\boldsymbol{\psi})\,\mathsf{E}\left[1|\boldsymbol{y}_a,\boldsymbol{\psi}'\right] = \ell(\boldsymbol{\psi}),$$

získáme

$$\ell(\boldsymbol{\psi}) = Q(\boldsymbol{\psi}|\boldsymbol{\psi}') - H(\boldsymbol{\psi}|\boldsymbol{\psi}'), \tag{2.13}$$

kde

$$Q(\boldsymbol{\psi}|\boldsymbol{\psi}') = \mathsf{E}\left[\ell_c(\boldsymbol{\psi})|\boldsymbol{y}_a,\boldsymbol{\psi}'\right], \qquad H(\boldsymbol{\psi}|\boldsymbol{\psi}') = \mathsf{E}\left[\log f(\boldsymbol{y}_a,\boldsymbol{Z}_a;\boldsymbol{\psi})|\boldsymbol{y}_a,\boldsymbol{\psi}'\right].$$

Lemma 2 (Dempster a kol., 1977, str. 6)**.** *Pro všechny dvojice* $(\boldsymbol{\psi},\boldsymbol{\psi}') \in \Psi\times\Psi$ *platí* $H(\boldsymbol{\psi}|\boldsymbol{\psi}') \le H(\boldsymbol{\psi}'|\boldsymbol{\psi}')$.

Důkaz. Zvol pevné $\boldsymbol{\psi}' \in \Psi$, pak pro libovolné $\boldsymbol{\psi} \in \Psi$ platí

$$\begin{aligned}
H(\boldsymbol{\psi}|\boldsymbol{\psi}') - H(\boldsymbol{\psi}'|\boldsymbol{\psi}') &= \\
&= \mathsf{E}\left[\log f(\boldsymbol{y}_a,\boldsymbol{Z}_a;\boldsymbol{\psi})|\boldsymbol{y}_a,\boldsymbol{\psi}'\right] - \mathsf{E}\left[\log f(\boldsymbol{y}_a,\boldsymbol{Z}_a;\boldsymbol{\psi}')|\boldsymbol{y}_a,\boldsymbol{\psi}'\right] \\
&= \mathsf{E}\left[\log\frac{f(\boldsymbol{y}_a,\boldsymbol{Z}_a;\boldsymbol{\psi})}{f(\boldsymbol{y}_a,\boldsymbol{Z}_a;\boldsymbol{\psi}')}\middle|\boldsymbol{y}_a,\boldsymbol{\psi}'\right] \\
&\le \mathsf{E}\left[\frac{f(\boldsymbol{y}_a,\boldsymbol{Z}_a;\boldsymbol{\psi})}{f(\boldsymbol{y}_a,\boldsymbol{Z}_a;\boldsymbol{\psi}')} - 1\middle|\boldsymbol{y}_a,\boldsymbol{\psi}'\right] \\
&= \mathsf{E}\left[1|\boldsymbol{y}_a,\boldsymbol{\psi}\right] - \mathsf{E}\left[1|\boldsymbol{y}_a,\boldsymbol{\psi}'\right] = 1-1 = 0.
\end{aligned}$$

Ve třetím řádku v nerovnosti bylo využito, že $\forall x \in (0,\infty), \log x \le x-1$. Stojí za zmínku, že rovnost nastává právě tehdy, když $f(\boldsymbol{y}_a,\boldsymbol{z}_a|\boldsymbol{y}_a;\boldsymbol{\psi}) = f(\boldsymbol{y}_a,\boldsymbol{z}_a|\boldsymbol{y}_a;\boldsymbol{\psi}')$ skoro všude.

❑

Z (2.13) je patrné, že pokud pro pevné $\boldsymbol{\psi}' \in \Psi$ najdeme prvek $\boldsymbol{\psi} \in \Psi$ splňující nerovnost $Q(\boldsymbol{\psi}|\boldsymbol{\psi}') \ge Q(\boldsymbol{\psi}'|\boldsymbol{\psi}')$, pak

$$\ell(\boldsymbol{\psi}) \ge \ell(\boldsymbol{\psi}'). \tag{2.14}$$

To proto, že

$$\ell(\boldsymbol{\psi}) = Q(\boldsymbol{\psi}|\boldsymbol{\psi}') - H(\boldsymbol{\psi}|\boldsymbol{\psi}') \ge Q(\boldsymbol{\psi}'|\boldsymbol{\psi}') - H(\boldsymbol{\psi}'|\boldsymbol{\psi}') = \ell(\boldsymbol{\psi}').$$

Nyní už je zřejmé, jak bude algoritmus postupovat. Pro libovolnou počáteční hodnotu $\boldsymbol{\psi}^{(0)} \in \Psi$ budeme hledat $\boldsymbol{\psi}^{(1)} \in \Psi$ takové, že $Q(\boldsymbol{\psi}^{(1)}|\boldsymbol{\psi}^{(0)}) \ge Q(\boldsymbol{\psi}^{(0)}|\boldsymbol{\psi}^{(0)})$, a pak $\boldsymbol{\psi}^{(2)}$ a tak dále. To nám zaručí, že $\ell(\boldsymbol{\psi}^{(2)}) \ge \ell(\boldsymbol{\psi}^{(1)}) \ge \ell(\boldsymbol{\psi}^{(0)})$ a budeme tak získávat věrohodnější odhady parametrů. Takovému postupu říkáme zobecněný EM algoritmus. Samotný EM algoritmus spočívá v hledání maxima funkce $Q(\boldsymbol{\psi}|\boldsymbol{\psi}'), \boldsymbol{\psi} \in \Psi$. Přesněji vše popíšeme v následující definici.

Definice 5. *Nechť M je zobrazení, $M : \Psi \to \Psi$, takové, že* $\forall \boldsymbol{\psi}' \in \Psi$

$$Q(M(\boldsymbol{\psi}')|\boldsymbol{\psi}') \geq Q(\boldsymbol{\psi}'|\boldsymbol{\psi}').$$

Pak algoritmus se zobrazením M nazveme zobecněný EM algoritmus. *Pokud navíc platí, že* $\forall \boldsymbol{\psi}' \in \Psi$

$$M(\boldsymbol{\psi}') = \arg\max_{\boldsymbol{\psi} \in \Psi} Q(\boldsymbol{\psi}|\boldsymbol{\psi}'),$$

řekneme, že algoritmus se zobrazením M je EM algoritmus.

Zobrazení M nám po k iteracích vytvoří pro libovolnou počáteční hodnotu parametru prvek $\boldsymbol{\psi}^{(k)}$. Z (2.14) víme, že $\ell(\boldsymbol{\psi}^{(k+1)}) \geq \ell(\boldsymbol{\psi}^{(k)})$. Tudíž, v případě, že je věrohodnostní funkce omezená, musí posloupnost $\ell(\boldsymbol{\psi}^{(k)})$ konvergovat k nějakému reálnému číslu. Konvergence může ale nastat i v případě, že věrohodnostní funkce omezená není. Dempster a kol., 1977 ukázali, že za jistých podmínek pro výše uvedené funkce, zejména jejich diferencovatelnosti, pak

$$\ell^* = \lim_{k \to \infty} \ell(\boldsymbol{\psi}^{(k)})$$

bude lokální maximum log-věrohodnostní funkce $\ell(\boldsymbol{\psi})$, případně nějaký sedlový bod. Samotný algoritmus ukončíme v okamžiku kdy rozdíl $\ell(\boldsymbol{\psi}^{(k+1)}) - \ell(\boldsymbol{\psi}^{(k)})$ bude menší než nějaká předem určená mez.

Pro použití EM algoritmu tedy bude první nutné zvolit několik různých počátečních hodnot a z nich EM algoritmus spustit. Pro všechny tyto pozice pak algoritmus dokonverguje k nějakému hodnotě funkce $\ell(\boldsymbol{\psi})$, které v drtivé většině případů bude jejím lokálním maximem. Jejich počet pak bude záviset na tvaru log-věrohodnostní funkce. Parametrické vektory, které tyto maxima nabývají, pak pro nás budou kandidáty na hledaný odhad. O jeho rozdělení, případně jiných vlastnostech ale nebudeme schopni nic říci, neboť ten nejspíše nebude klasickým odhadem metodou maximální věrohodnosti a zkoumání jeho vlastností by bylo, jak jsme si uvedli v sekci 2.2, nad rámec této bakalářské práce.

V následující části se budeme snažit pro konečné směsi takový algoritmus najít. Nechť $\boldsymbol{\psi}^{(k)} \in \Psi$, $k \in \mathbb{N}_0$, budeme hledat $\boldsymbol{\psi}^{(k+1)} \in \Psi$ takové, že

$$\boldsymbol{\psi}^{(k+1)} = \arg\max_{\boldsymbol{\psi} \in \Psi} Q(\boldsymbol{\psi}|\boldsymbol{\psi}^{(k)}).$$

Hledání tohoto prvku rozdělíme do dvou kroků. V prvním, v takzvaném E-kroku, spočítáme střední hodnotu $\mathsf{E}\left[\ell_c(\boldsymbol{\psi})|\boldsymbol{y}_a, \boldsymbol{\psi}^{(k)}\right]$ v závislosti na parametru $\boldsymbol{\psi}$ a poté, ve druhém kroku zvaném M-krok budeme tuto funkci maximalizovat pro $\boldsymbol{\psi} \in \Psi$.

2.4.1 E-krok

Nyní tedy spočítáme střední hodnotu $\mathsf{E}\left[\ell_c(\boldsymbol{\psi})|\boldsymbol{y}_a, \boldsymbol{\psi}^{(k)}\right]$. Za $\ell_c(\boldsymbol{\psi})$ dosadíme z (2.12) a vyjádříme:

$$\begin{aligned} Q(\boldsymbol{\psi}|\boldsymbol{\psi}^{(k)}) &= \mathsf{E}\left[\ell_c(\boldsymbol{\psi})|\boldsymbol{y}_a, \boldsymbol{\psi}^{(k)}\right] \\ &= \mathsf{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{g} Z_{ij}\Big(\log \pi_j + \log f(\boldsymbol{y}_i,\boldsymbol{\theta}_j)\Big)\middle| \boldsymbol{y}_a, \boldsymbol{\psi}^{(k)}\right] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{g} \mathsf{E}\left[Z_{ij}|\boldsymbol{y}_a, \boldsymbol{\psi}^{(k)}\right]\Big(\log \pi_j + \log f(\boldsymbol{y}_i,\boldsymbol{\theta}_j)\Big) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{g} p_j(\boldsymbol{y}_i;\boldsymbol{\psi}^{(k)})\Big(\log \pi_j + \log f(\boldsymbol{y}_i,\boldsymbol{\theta}_j)\Big). \end{aligned} \tag{2.15}$$

V poslední rovnosti jsme využili toho, že

$$\mathsf{E}\left[Z_{ij}|\boldsymbol{y}_a, \boldsymbol{\psi}^{(k)}\right] = P\left(Z_{ij}=1|\boldsymbol{y}_a, \boldsymbol{\psi}^{(k)}\right) = p_j(\boldsymbol{y}_i;\boldsymbol{\psi}^{(k)}).$$

Zde symbolem $p_j(\boldsymbol{y}_i;\boldsymbol{\psi}^{(k)})$ značíme zpětnou pravděpodobnost toho, že $\boldsymbol{y}_i$ náleželo do j-té skupiny, pokud hustota směsi odpovídá parametru $\boldsymbol{\psi}^{(k)}$. Výpočet byl proveden v (1.2). Tedy

$$p_j(\boldsymbol{y}_i;\boldsymbol{\psi}^{(k)}) = \frac{\pi_j^{(k)} f(\boldsymbol{y}_i;\boldsymbol{\theta}_j^{(k)})}{f(\boldsymbol{y}_i;\boldsymbol{\psi}^{(k)})}. \tag{2.16}$$

Dále budeme značit symbolem

$$p_{ij}^{(k)} = p_j(\boldsymbol{y}_i;\boldsymbol{\psi}^{(k)}).$$

2.4.2 M-krok

Teď bude naším cílem najít řešení rovnice $\boldsymbol{\psi}^{(k+1)} = \arg\max_{\boldsymbol{\psi}\in\Psi} Q(\boldsymbol{\psi}|\boldsymbol{\psi}^{(k)})$. Označíme-li

$$Q_1(\boldsymbol{\psi}|\boldsymbol{\psi}^{(k)}) = \sum_{i=1}^{n}\sum_{j=1}^{g} p_{ij}^{(k)} \log \pi_j, \quad Q_2(\boldsymbol{\psi}|\boldsymbol{\psi}^{(k)}) = \sum_{i=1}^{n}\sum_{j=1}^{g} p_{ij}^{(k)} \log f(\boldsymbol{y}_i,\boldsymbol{\theta}_j),$$

z (2.15) lze ověřit, že $Q(\boldsymbol{\psi}|\boldsymbol{\psi}^{(k)}) = Q_1(\boldsymbol{\psi}|\boldsymbol{\psi}^{(k)}) + Q_2(\boldsymbol{\psi}|\boldsymbol{\psi}^{(k)})$. Navíc v $Q_1(\boldsymbol{\psi}|\boldsymbol{\psi}^{(k)})$ se z parametrů vyskytují pouze poměry složek π_j a v $Q_2(\boldsymbol{\psi}|\boldsymbol{\psi}^{(k)})$ naopak pouze parametry hustot $\boldsymbol{\theta}_j$. To nám umožní maximalizaci rozdělit do dvou částí. V první, pomocí maximalizace $Q_1(\boldsymbol{\psi}|\boldsymbol{\psi}^{(k)})$, najdeme nové odhady $\pi_j^{(k+1)}$ a ve druhé se poté zaměříme na hledání nových odhadů $\boldsymbol{\theta}_j^{(k+1)}$.

Nyní se tedy budeme věnovat hledání $\pi_j^{(k+1)}$. Označme

$$\boldsymbol{\pi} = (\pi_1, \dots \pi_g)^\top, \quad P = \left(\boldsymbol{\pi} \in (0,1)^g : \sum_{j=1}^{g} \pi_j = 1\right), \quad P_{1j}^{(k)} = \sum_{i=1}^{n} p_{ij}^{(k)}.$$

Pak P je parametrický prostor pro $\boldsymbol{\pi}$. Pomocí Lagrangeovy věty o multiplikátoru nyní vyřešíme rovnici

$$\boldsymbol{\pi}^{(k+1)} = \arg\max_{\boldsymbol{\pi}\in P} \sum_{i=1}^{n}\sum_{j=1}^{g} p_{ij}^{(k)} \log \pi_j = \arg\max_{\boldsymbol{\pi}\in P} \sum_{j=1}^{g} P_{1j}^{(k)} \log \pi_j, \tag{2.17}$$

Protože P není kompaktní prostor, musíme ho nějak uzavřít. To ale nebude obtížné, protože uzavřený není jen když $\pi_j \to 0$ pro nějaké j a v tomto případě se hodnota sumy blíží k $-\infty$. Formálně položme $c = \min_{j=1,\dots,g} P_{1j}^{(k)}, c > 0$ a uvažme $\boldsymbol{\pi} \in P$ takové, že $\exists j_0 : \pi_{j_0} \leq \left(\frac{c}{n}\right)^{gn/c}$. Potom, protože sčítance sumy jsou záporné

$$\sum_{j=1}^{g} P_{1j}^{(k)} \log \pi_j < P_{1j_0}^{(k)} \log \pi_{j_0} \leq c \log \left(\frac{c}{n}\right)^{gn/c} = gn \log \frac{c}{n}.$$

Nyní je zřejmé, že pokud maximum sumy bude větší než $gn \log \frac{c}{n}$, musí být splněna následující podmínka:

$$\frac{\partial}{\partial \pi_j}\left(\sum_{j=1}^{g} P_{1j}^{(k)} \log \pi_j + \lambda\left(\sum_{j=1}^{g} \pi_j - 1\right)\right)(\pi_j^{(k+1)}) = 0, \quad j = 1,\dots,g,$$

kde λ je Lagrangeův multiplikátor. Z čehož máme:

$$P_{1j}^{(k)} \frac{1}{\pi_j^{(k+1)}} + \lambda = 0 \quad \Leftrightarrow \quad P_{1j}^{(k)} + \lambda \pi_j^{(k+1)} = 0, \quad j = 1,\dots,g. \tag{2.18}$$

Sečtením rovnic (2.18) pro všechna j a s využitím, že

$$\sum_{j=1}^{g} \pi_j^{(k+1)} = 1 \quad \& \quad \sum_{j=1}^{g} P_{1j}^{(k)} = n,$$

dostaneme $\lambda = -n$. Dosazením do (2.18) získáme, že

$$\pi_j^{(k+1)} = \frac{P_{1j}^{(k)}}{n}, \qquad j = 1,\dots,g.$$

Poslední, co zbývá ukázat je:

$$\sum_{j=1}^{g} P_{1j}^{(k)} \log \frac{P_{1j}^{(k)}}{n} > \sum_{j=1}^{g} n \log \frac{c}{n} = gn \log \frac{c}{n}.$$

Tedy můžeme říci, že

$$\pi_j^{(k+1)} = \frac{P_{1j}^{(k)}}{n} = \frac{1}{n}\sum_{i=1}^{n} p_{ij}^{(k)}, \qquad j = 1,\dots,g, \tag{2.19}$$

maximalizuje (2.17). Zajímavá je interpretace těchto vzorců. Každé pozorování přispívá do odhadu π_j velikostí pravděpodobnosti (při aktuální hodnotě $\boldsymbol{\psi}^{(k)}$), se kterou do j-té komponenty patří, což je velmi přímočaré a odpovídá intuici.

Zbývá nám maximalizovat $Q_2(\boldsymbol{\psi}|\boldsymbol{\psi}^{(k)})$. Tedy najít nový odhad $\boldsymbol{\xi}^{(k+1)}$ parametru $\boldsymbol{\xi}$. To se v takovéto obecnosti ukazuje jako nemožné. Jedna z možných cest je hledání řešení rovnice

$$\frac{\partial Q_2(\boldsymbol{\psi}|\boldsymbol{\psi}^{(k)})}{\partial \boldsymbol{\xi}} = \sum_{i=1}^{n}\sum_{j=1}^{g} p_{ij}^{(k)} \frac{\partial \log f(\boldsymbol{y}_i,\boldsymbol{\theta}_j)}{\partial \boldsymbol{\xi}} = \mathbf{0}.$$

Tyto rovnice však už v obecném tvaru nejdou dále nějak pěkně upravit. Ukážeme si ale, že v případě normálních směsí lze najít explicitní řešení. To bude předmětem následující kapitoly.

Kapitola 3

Normální směsi

V této kapitola se zaměříme na rodinu směsí, které v jejím samotném kontextu hrají prominentní roli. Těmi jsou normální směsi. V předchozí kapitole jsme odvodili pouze část vzorců pro iteraci EM algoritmu a nyní si ukážeme, že v případě normálních směsí lze explicitně odvodit vzorce i pro ostatní parametry modelu.

Mějme tedy p-rozměrnou normální směs ve tvaru (1.4), tj. směs jejíž hustoty složek jsou tvaru

$$f\left(\boldsymbol{y};\boldsymbol{\theta}_j\right) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_j|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu}_j)^\top \boldsymbol{\Sigma}_j^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_j)\right), \quad j=1,\ldots,g,$$

kde vektor $\boldsymbol{\theta}_j$ obsahuje vektor středních hodnot $\boldsymbol{\mu}_j$ a prvky matice $\boldsymbol{\Sigma}_j$, které jsou v horním trojúhelníku. Ty budeme značit $\mathrm{vec}(\boldsymbol{\Sigma}_j)$. Tj $\boldsymbol{\theta}_j = (\boldsymbol{\mu}_j^\top, \mathrm{vec}(\boldsymbol{\Sigma}_j)^\top)^\top$.

3.1 EM algoritmus pro normální směsi

Nechť $\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_n$ je náhodný výběr z p-rozměrné g-složkové normální směsi $f(\boldsymbol{y},\boldsymbol{\psi})$, kde vektor $\boldsymbol{\psi}$ obsahuje všechny neznámé parametry modelu. Z minulé kapitoly víme, že naším cílem je maximalizovat funkci

$$Q(\boldsymbol{\psi}|\boldsymbol{\psi}^{(k)}) = \mathsf{E}\left[\ell_c(\boldsymbol{\psi})|\boldsymbol{y}_a,\boldsymbol{\psi}^{(k)}\right]$$

pro předem dané $\boldsymbol{\psi}^{(k)}$. Označili jsme si

$$\boldsymbol{\psi}^{(k+1)} = \arg\max_{\boldsymbol{\psi}\in\Psi} Q(\boldsymbol{\psi}|\boldsymbol{\psi}^{(k)}) = \arg\max_{\boldsymbol{\psi}\in\Psi}\Big(Q_1(\boldsymbol{\psi}|\boldsymbol{\psi}^{(k)}) + Q_2(\boldsymbol{\psi}|\boldsymbol{\psi}^{(k)})\Big)$$

a pomocí maximalizace funkce $Q_1(\boldsymbol{\psi}|\boldsymbol{\psi}^{(k)})$ jsme získali nový odhad (2.19) poměrů složek směsi

$$\pi_j^{(k+1)} = \frac{P_{1j}^{(k)}}{n}, \qquad j=1,\ldots,g.$$

Zbývá nám dopočítat parametry vektoru $\boldsymbol{\xi}$, což v případě námi uvažovaných normálních směsí znamená nové odhady středních hodnot a rozptylových matic.

Vyjádříme-li $Q_2(\boldsymbol{\psi}|\boldsymbol{\psi}^{(k)})$ pro normální směs, dostaneme

$$\begin{aligned}\boldsymbol{\xi}^{(k+1)} &= \arg\max \sum_{i=1}^{n}\sum_{j=1}^{g} p_{ij}^{(k)} \left(-\frac{1}{2}\log\left((2\pi)^k|\boldsymbol{\Sigma}_j|\right) - \frac{1}{2}(\boldsymbol{y}_i-\boldsymbol{\mu}_j)^\top\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{y}_i-\boldsymbol{\mu}_j)\right) \\ &= \arg\max \sum_{i=1}^{n}\sum_{j=1}^{g} p_{ij}^{(k)} \left(\log\left(|\boldsymbol{\Sigma}_j^{-1}|\right) - (\boldsymbol{y}_i-\boldsymbol{\mu}_j)^\top\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{y}_i-\boldsymbol{\mu}_j)\right). \qquad (3.1)\end{aligned}$$

Uvědomíme si, že nalezení nových středních hodnot je jednoduché. To z toho důvodu, že (3.1) je kvadratickou funkcí $\boldsymbol{\mu}_j$, navíc se záporným koeficientem u druhé mocniny. Tudíž pro $\boldsymbol{\mu}_j^{(k+1)}$ nutně musí platit a je i postačující, že

$$\frac{\partial}{\partial\boldsymbol{\mu}_j}\left(\sum_{i=1}^{n}\sum_{j=1}^{g} p_{ij}^{(k)} \left(\log\left(|\boldsymbol{\Sigma}_j^{-1}|\right) - (\boldsymbol{y}_i-\boldsymbol{\mu}_j)^\top\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{y}_i-\boldsymbol{\mu}_j)\right)\right)(\boldsymbol{\mu}_j^{(k+1)}) = \mathbf{0}.$$

Derivováním dle $\boldsymbol{\mu}_j$, využitím, že $\boldsymbol{\Sigma}_j^{-1}$ je regulární matice a následnou úpravou potom dostaneme, že

$$\begin{aligned}\sum_{i=1}^{n} p_{ij}^{(k)} 2\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{y}_i-\boldsymbol{\mu}_j^{(k+1)}) &= \mathbf{0}, \\ \sum_{i=1}^{n} p_{ij}^{(k)}\boldsymbol{y}_i &= \sum_{i=1}^{n} p_{ij}^{(k)}\boldsymbol{\mu}_j^{(k+1)}, \\ \boldsymbol{\mu}_j^{(k+1)} &= \frac{\sum_{i=1}^{n} p_{ij}^{(k)}\boldsymbol{y}_i}{\sum_{i=1}^{n} p_{ij}^{(k)}}. \qquad (3.2)\end{aligned}$$

Podobně jako vzorec pro výpočet poměrů složek směsi, vzorec (3.2) má také přirozenou interpretaci. Každé měření přispívá do odhadu dané střední hodnoty proporcí, se kterou si v daném okamžiku myslíme, že do té složky patří. Tento součet pak vydělíme předpokládanou velikostí složky ve směsi. Jde tudíž o analogii váženého výběrového průměru.

Konečně nový odhad rozptylové matice $\boldsymbol{\Sigma}_j$ spočítáme obdobně. Vyjádříme

$$\frac{\partial}{\partial\boldsymbol{\Sigma}_j^{-1}}\left(\sum_{i=1}^{n}\sum_{j=1}^{g} p_{ij}^{(k)}\left(\log\left(|\boldsymbol{\Sigma}_j^{-1}|\right) - (\boldsymbol{y}_i-\boldsymbol{\mu}_j)^\top\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{y}_i-\boldsymbol{\mu}_j)\right)\right)(\boldsymbol{\Sigma}_j^{(k+1)}) = \mathbf{0}.$$

Odsud dostaneme, že je nutné řešit rovnici

$$\sum_{i=1}^{n} p_{ij}^{(k)}\left(\boldsymbol{\Sigma}_j^{(k+1)} - (\boldsymbol{y}_i-\boldsymbol{\mu}_j^{(k+1)})(\boldsymbol{y}_i-\boldsymbol{\mu}_j^{(k+1)})^\top\right) = \mathbf{0}.$$

Jednoduchou úpravou poté dostaneme výsledný vzorec

$$\boldsymbol{\Sigma}_j^{(k+1)} = \frac{\sum_{i=1}^{n} p_{ij}^{(k)}(\boldsymbol{y}_i-\boldsymbol{\mu}_j^{(k+1)})(\boldsymbol{y}_i-\boldsymbol{\mu}_j^{(k+1)})^\top}{\sum_{i=1}^{n} p_{ij}^{(k)}}, \qquad j=1,\ldots,g. \qquad (3.3)$$

Zbývá ukázat, že se opravdu jedná o maximum funkce. K tomu nám bude stačit zjištění, že se jedná o konkávní funkci matice $\boldsymbol{\Sigma}_j^{-1}$. S využitím pravidel maticového kalkulu o derivaci inverzní matice můžeme počítat

$$\frac{\partial^2}{\partial(\boldsymbol{\Sigma}_j^{-1})^2} Q_2(\boldsymbol{\psi}|\boldsymbol{\psi}^{(k)}) = \sum_{i=1}^{n} -p_{ij}^{(k)}(\boldsymbol{\Sigma}_j \otimes \boldsymbol{\Sigma}_j) = -P_{1j}^{(k)}(\boldsymbol{\Sigma}_j \otimes \boldsymbol{\Sigma}_j), \tag{3.4}$$

kde symbol $\otimes$ značí Kroneckerův maticový součin. Potřebujeme, aby matice na pravé straně v (3.4) byla negativně definitní pro všechna $\boldsymbol{\Sigma}_j^{-1}$, k čemuž stačí ukázat, že matice $\boldsymbol{\Sigma}_j \otimes \boldsymbol{\Sigma}_j$ je pozitivně definitní. Pro naše účely tuto vlastnost ukážeme pouze pro matice $\boldsymbol{\Sigma}_j$ rozměru 2×2. Předpokládejme, že prvky matice $\boldsymbol{\Sigma}_j$ jsou postupně a_{11}, a_{12}, a_{21} a a_{22}. Ze Schurovy komplementární podmínky pro pozitivní definitnost matice (Dattorro, 2015, strana 527) nám stačí ukázat, že matice $a_{11}\boldsymbol{\Sigma}_j$ a $a_{22}\boldsymbol{\Sigma}_j - a_{12}\boldsymbol{\Sigma}_j \cdot \frac{1}{a_{11}}\boldsymbol{\Sigma}_j^{-1} \cdot a_{21}\boldsymbol{\Sigma}_j$ jsou pozitivně definitní. To je ale přímý důsledek toho, že $\boldsymbol{\Sigma}_j$ je pozitivně definitní díky čemuž ze Sylvestrova kritéria víme, že $a_{11} > 0$ a $a_{11}a_{22} > a_{12}a_{21}$.

Znovu se jedná o analogii výpočtu výběrové rozptylové matice, které jsou navíc váženy odhadovanou pravděpodobností, že daný prvek náleží do j-té skupiny.

Označíme-li navíc

$$\boldsymbol{P}_{2j}^{(k)} = \sum_{i=1}^{n} p_{ij}^{(k)} \boldsymbol{y}_i \quad \& \quad \boldsymbol{P}_{3j}^{(k)} = \sum_{i=1}^{n} p_{ij}^{(k)} \boldsymbol{y}_i \boldsymbol{y}_i^\top,$$

můžeme všechny parametry pro $(k+1)$ní iteraci EM algoritmu normální směsi odvozené v (2.19),(3.2) a (3.3) vyjádřit jako

$$\begin{aligned}
\pi_j^{(k+1)} &= \frac{P_{1j}^{(k)}}{n}, \qquad \boldsymbol{\mu}_j^{(k+1)} = \frac{\boldsymbol{P}_{2j}^{(k)}}{P_{1j}^{(k)}}, \\
\boldsymbol{\Sigma}_j^{(k+1)} &= \frac{\boldsymbol{P}_{3j}^{(k)} - \left(P_{1j}^{(k)}\right)^{-1} \boldsymbol{P}_{2j}^{(k)} \left(\boldsymbol{P}_{2j}^{(k)}\right)^\top}{P_{1j}^{(k)}}, \qquad (j = 1, \ldots, g).
\end{aligned} \tag{3.5}$$

Zobrazení definované vzorci (3.5) pro $\boldsymbol{\psi} \in \Psi$ je tedy EM algoritmem alespoň pro dvourozměrné normální směsi. Obecně v literatuře, která se odvozením vzorců EM algoritmu pro normální směsi zabývá, bylo zvykem odvodit toto zobrazení pouze z nutných podmínek existence extrému. To, zda-li tyto parametry skutečně maximalizují (2.15), nebylo, pokud nám je známo, nikde detailně ukázáno a tento fakt se tiše předpokládal.

V případě homoskedastické směsi, když předpokládáme, že rozptylové matice jsou stejné, můžeme postupovat obdobně pomocí derivace funkce $Q_2(\boldsymbol{\psi}|\boldsymbol{\psi}^{(k)})$. Po jednoduchých úpravách získáme vážený průměr rozptylových matic (3.3).

$$\begin{aligned}
\boldsymbol{\Sigma}^{(k+1)} &= \frac{\sum_{j=1}^{g} P_{1j}^{(k)} \boldsymbol{\Sigma}_j^{(k+1)}}{n} = \frac{\sum_{j=1}^{g} \sum_{i=1}^{n} p_{ij}^{(k)} (\boldsymbol{y}_i - \boldsymbol{\mu}_j^{(k+1)})(\boldsymbol{y}_i - \boldsymbol{\mu}_j^{(k+1)})^\top}{n} \\
&= \frac{1}{n} \sum_{j=1}^{g} \left(\boldsymbol{P}_{3j}^{(k)} - \left(P_{1j}^{(k)}\right)^{-1} \boldsymbol{P}_{2j}^{(k)} \left(\boldsymbol{P}_{2j}^{(k)}\right)^\top \right).
\end{aligned}$$

3.2 Odhad modelu produkce cytokinu IL10

Nyní navážeme na model produkce cytokinu IL10, který jsme uvedli v sekci 1.4. V předchozích kapitolách jsme popsali způsob odhadování modelu pomocí metody maximální věrohodnosti s využitím EM algoritmu, jenž nyní aplikujeme. Jakým způsobem analýzu provedeme, popíšeme na následujících řádcích.

K analýze použijeme software R a volně dostupný balík *EMCluster* (Chen a kol., 2012), v němž je již EM algoritmus naprogramován. Cílem analýzy bude najít vhodný model rozdělení produkce cytokinu, který odpovídá jak naměřeným datům tak i lidské intuici. Uvažujme tudíž dvě skupiny, jednu nemocnou a druhou zdravou, čemuž odpovídá volba $g = 2$. Pro spuštění EM algoritmu bude potřeba vybrat počáteční podmínky pro algoritmus. Ty se pokusíme zvolit tak, abychom objevili co možná všechna lokální maxima log-věrohodnostní funkce (2.1). Zvolme tedy 9 středních hodnot rovnoměrně rozdělených po ploše, na které jsou umístěna data, a dále 9 různých rozptylových matic. Samotný algoritmus spustíme z odlišných kombinací těchto středních hodnot a rozptylových matic při shodném poměru složek směsi 0,5. Tímto postupem vytvoříme celkem 2916 různých počátečních pozic, ze kterých postupně algoritmus odstartujeme. Ukázalo se, ze všech těchto počátečních pozic algoritmus dokonvergoval vždy k jednomu ze sedmi parametrických vektorů, určujících rozdělení směsí uvedených v tabulce 3.1.

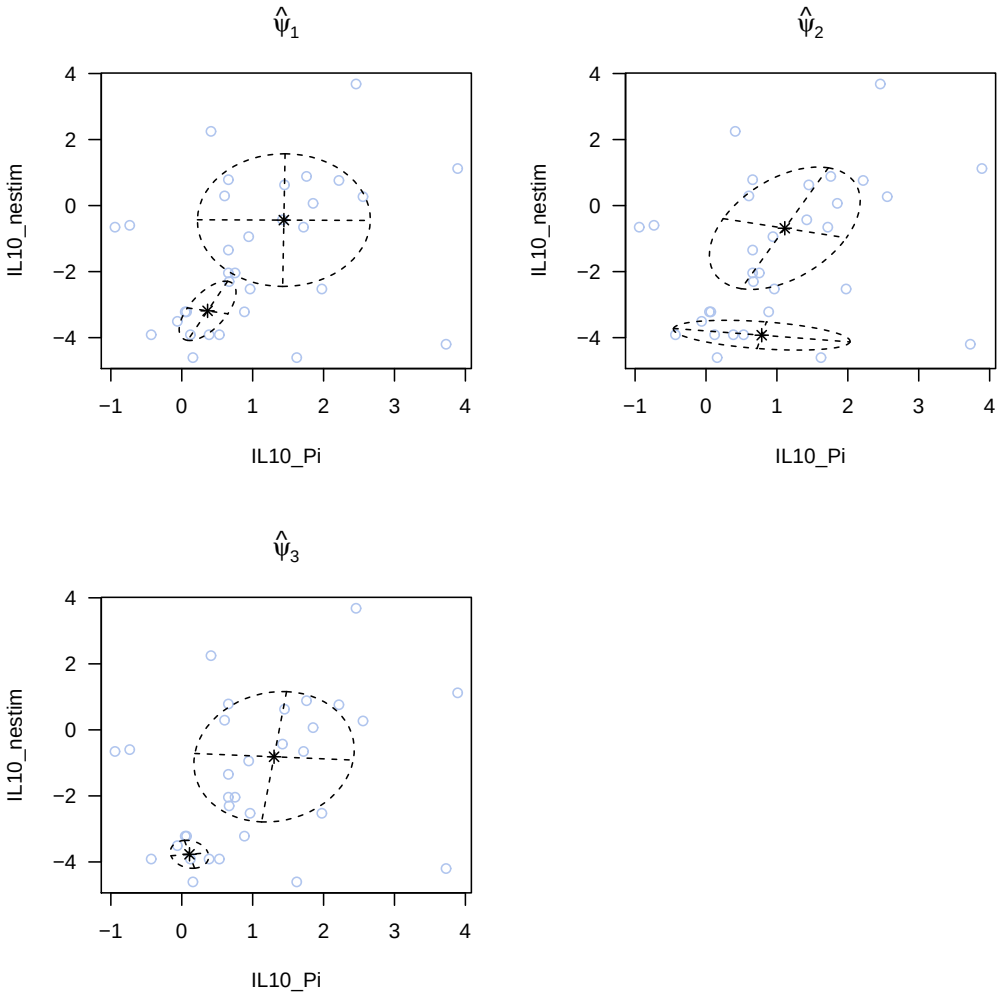
	π	μ_1	μ_2	σ_1	σ_2	ρ	$\ell(\boldsymbol{\psi})$
$\hat{\boldsymbol{\psi}}_1$	0,62	1,44	−0,44	1,22	2,01	0,01	−108,7
	0,38	0,36	−3,19	0,40	0,90	0,63	
$\hat{\boldsymbol{\psi}}_2$	0,75	1,11	−0,69	1,06	1,85	0,44	−111,6
	0,25	0,79	−3,93	1,25	0,45	−0,41	
$\hat{\boldsymbol{\psi}}_3$	0,77	1,30	−0,81	1,13	1,98	0,11	−109,2
	0,23	0,11	−3,77	0,27	0,42	−0,16	
$\hat{\boldsymbol{\psi}}_4$	0,94	0,85	−1,48	0,89	2,10	0,45	41,4
	0,06	3,81	−1,54	0,08	2,66	1,00	
$\hat{\boldsymbol{\psi}}_5$	0,90	1,06	−1,44	0,95	2,19	0,59	−109,9
	0,10	0,76	−1,89	2,05	1,60	−1,00	
$\hat{\boldsymbol{\psi}}_6$	0,84	0,97	−1,83	0,94	2,17	0,30	−113,5
	0,16	1,34	0,26	1,73	0,67	0,93	
$\hat{\boldsymbol{\psi}}_7$	0,94	0,96	−1,52	1,04	2,05	0,55	12,7
	0,06	2,07	−0,98	1,66	3,22	−1,00	

Pro $\hat{\boldsymbol{\psi}}_i$ tabulka ukazuje odhad parametrů dvousložkové normální směsi jemu odpovídající. π je poměr zastoupení složky ve směsi a $\mu_1, \mu_2, \sigma_1, \sigma_2, \rho$ pak její parametry. (V každé směsi jsou dvě složky.)

Tabulka 3.1: Parametry rozdělení směsi určené lokálními maximizátory věrohodnostní funkce.

Je zajímavé, že hodnoty log-věrohodnostní funkce v získaných parametrických vektorech jsou od sebe dosti odlišné. Zejména hodnoty $\ell(\hat{\boldsymbol{\psi}}_4)$ a $\ell(\hat{\boldsymbol{\psi}}_7)$ jsou překvapivě vysoké a znatelně se liší od ostatních, které jsou všechny blízko hodnotě -110, jak se můžeme přesvědčit v tabulce 3.1. Pokud se ale podrobněji podíváme na korelační koeficient $\hat{\boldsymbol{\psi}}_4$ a $\hat{\boldsymbol{\psi}}_7$ druhé složky v této tabulce, zjistíme, že je 1, resp. -1. To znamená, že rozptylová matice této složky je singulární. O tom, že podobná situace může nastat, jsme diskutovali v sekci 2.1. Tato degenerovaná rozdělení určená vektory $\hat{\boldsymbol{\psi}}_4$ a $\hat{\boldsymbol{\psi}}_7$ nám zřejmě neposkytnou vhodný model.

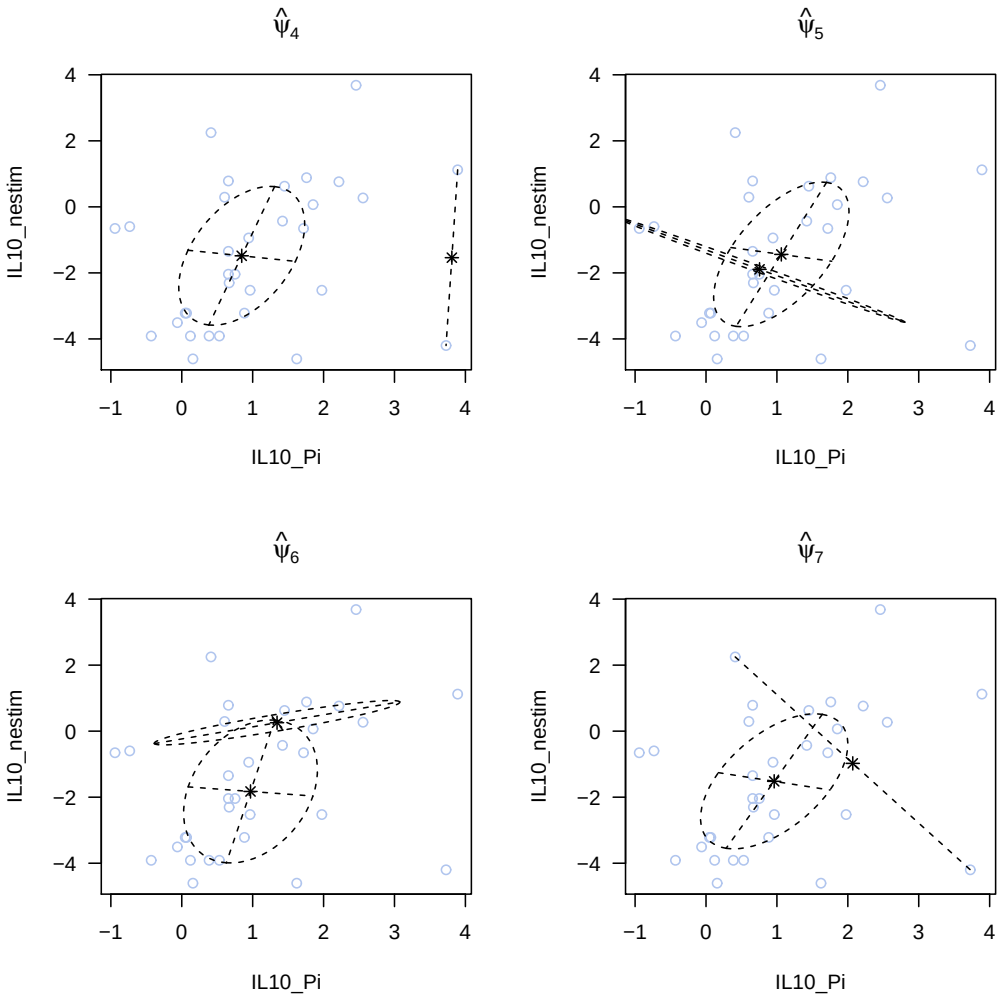
Vidíme, že ani jeden ze zbylých odhadů parametru $\boldsymbol{\psi}$ nebude odhadem metodou maximální věrohodnosti, protože nemaximalizuje věrohodnostní funkci. Avšak jak je z obrázků 3.1 a 3.2, kde jsou rozdělení směsí naznačena, patrné, některá získaná rozdělení by mohla přijatelně vystihovat popisovaný model. Protože jsme v průběhu této bakalářské práce nezmínili žádné možné testy, kterými lze vhodnost modelu testovat, v případě že bychom chtěli vybrat jedno rozdělení, které „sedí nejlépe“, nezbývá nám nic jiného než použít intuici a heuristiku.



Obrázek 3.1: Znázornění prvních tří rozdělení, jejichž odpovídající parametrické vektory lokálně maximalizují věrohodnostní funkci.

Jedním z vodítek, které z rozdělení by mohlo být vhodné, může být informace o poměru podílu složky π_1. Data, která máme k dispozici, totiž obsahovala i údaj, zda-li měřený pacient trpěl paradontózou či ne. Díky tomu víme, že zkoumáme měření 17 nemocných a 15 zdravých pacientů. Proto lze očekávat, že poměr π_1 by měl být okolo hodnoty 0,55. Pokud se podíváme na parametrické vektory $\hat{\boldsymbol{\psi}}_5$ a $\hat{\boldsymbol{\psi}}_6$, hodnota π_1 je u nich větší než 0,83, což je výrazně více, než je očekáváno. Navíc složka s menším poměrem je u obou rozdělení do značné míry singulární, jak je patrné z obrázku 3.2 i hodnot v tabulce 3.1, a z medicínského hlediska není jediný důvod, proč by měla být produkce cytokinu u jedné takové skupiny takto svázána. Proto se rozděleními reprezentovanými těmito parametrickými vektory nebudeme dále zabývat.

Zbylé 3 rozdělení, definované parametrickými vektory $\hat{\boldsymbol{\psi}}_1, \hat{\boldsymbol{\psi}}_2$ a $\hat{\boldsymbol{\psi}}_3$ zatím ponecháme jako realistické a budeme je zkoumat podrobněji v další sekci, kde budeme porovnávat správnost určení jednotlivých pacientů do daných skupin. Zjistíme, jak moc modely odpovídají naměřeným hodnotám, a zkusíme určit, který z nich nejlépe popisuje produkci cytokinu IL10, tak jak byla zjištěna.



Obrázek 3.2: Znázornění zbylých čtyř rozdělení, jejichž odpovídající parametrické vektory lokálně maximalizují věrohodnostní funkci.

Kapitola 4

Shlukování

V některých aplikacích modelů konečných směsí může být cílem zjištění, do kterých skupin pozorované prvky patří. Je přirozené se ptát, pokud předpokládáme, že studovaná populace se dělí řekněme na dvě skupiny, ze které skupiny daný objekt pochází. Dost možná je toto totiž i hlavní cíl, proč model směsi odhadujeme.

Mějme tedy g různých skupin nějaké populace a sledujme na ní rozdělení náhodného vektoru $\boldsymbol{Y}$. Nechť tato veličina má v j-té skupině rozdělení definované hustotou $f(\boldsymbol{y};\boldsymbol{\theta}_j)$. Pak, jak jsme si ukázali, má náhodný vektor $\boldsymbol{Y}$ rozdělení dáno hustotou (1.3), tedy hustotou

$$f\left(\boldsymbol{y};\boldsymbol{\psi}\right)=\sum_{j=1}^{g}\pi_j f(\boldsymbol{y};\boldsymbol{\theta}_j),$$

kde $\boldsymbol{\theta}_j$ a π_j, $j=1,\ldots,g$, $\sum\pi_j=1$, jsou neznámé parametry obsaženy ve vektoru $\boldsymbol{\psi}$. Naším cílem je na základě pozorování $\boldsymbol{Y}$ určit, ze které skupiny pozorovaný objekt pochází. Jinými slovy to znamená určit pro $\boldsymbol{y}$ jeho odpovídající vektor $\boldsymbol{z}$, jenž daný jev přesně popisuje.

Abychom tak ale mohli učinit, je první nutné odhadnout samotné rozdělení směsi $\boldsymbol{Y}$. To uděláme na základě pozorovaného náhodného výběru $\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n$, s hustotou (1.3), který je pro odhadnutí modelu nezbytný. Odhad lze pak provést například pomocí dříve uvedené metody maximální věrohodnosti a EM algoritmu. Předpokládejme tedy, že jsme pomocí jisté metody získali odhad $\hat{\boldsymbol{\psi}}\in\Psi$ parametrického vektoru $\boldsymbol{\psi}$. Predikce hodnot vektorů $\boldsymbol{z}_i$ pak lze provést na základě velikosti zpětné pravděpodobnosti, že pozorovaný vektor $\boldsymbol{y}_i$ patří do nějaké skupiny. Ty jsou dány pravděpodobnostmi $p_1(\boldsymbol{y}_i;\hat{\boldsymbol{\psi}}),\ldots,p_g(\boldsymbol{y}_i;\hat{\boldsymbol{\psi}})$, kde používáme značení $p_j(\boldsymbol{y}_i;\hat{\boldsymbol{\psi}})=\mathsf{P}(Z_{ij}=1|\boldsymbol{Y}_i=\boldsymbol{y}_i;\hat{\boldsymbol{\psi}})$, $j=1,\ldots,g$, viz (2.16).

Přesněji, hodnotu vektoru $\boldsymbol{z}_i$ budeme predikovat vektorem $\hat{\boldsymbol{z}}_i$, jehož prvek $\hat{z}_{ij}$ definujeme jako

$$\begin{aligned}\hat{z}_{ij}&=1, \qquad \text{když} \quad j=\arg\max_{l=1,\ldots,g} p_l(\boldsymbol{y}_i;\hat{\boldsymbol{\psi}}),\\ &=0, \qquad \text{jinak,}\end{aligned} \tag{4.1}$$

pro $i=1,\ldots,n$ a $j=1,\ldots,g$. Je zřejmé, že v případě rovnosti některých z těchto pravděpodobností by se mohlo stát, že vektor $\hat{\boldsymbol{z}}_i$ obsahuje více než jednu jedničku.

I v tomto případě bychom pozorování přiřadili pouze jedné skupině, a to libovolné z těch, u kterých je dosažena nejvyšší hodnota $p_j(\boldsymbol{y}_i; \hat{\boldsymbol{\psi}})$. Označme tento odhad $\hat{\boldsymbol{z}}_i = r_B(\boldsymbol{y}_i; \hat{\boldsymbol{\psi}})$.

V případě, že náš odhad $\hat{\boldsymbol{\psi}}$ přesně odpovídá skutečné hodnotě parametrického vektoru $\boldsymbol{\psi}$, nazývá se toto pravidlo přiřazování (4.1) Bayesovo pravidlo. Protože navíc všechny tyto pravděpodobnosti mají společný jmenovatel $f(\boldsymbol{y}_i)$, lze se pochopitelně rozhodovat pouze na základě relativních velikostí hustot složek vážených podle poměru zastoupení ve směsi, tj. na základě velikosti $\pi_j f(\boldsymbol{y}_i; \hat{\boldsymbol{\theta}}_j)$.

Bude užitečné se i zamyslet, jak bude fungovat toto pravidlo v případě, kdy získáme špatný odhad $\hat{\boldsymbol{\psi}}$ parametru $\boldsymbol{\psi}$. Je možné, že i tak budou $\hat{\boldsymbol{z}}_i$ stále dobrými odhady $\boldsymbol{z}_i$ čímž myslíme, že ve většině případů bude platit $r_B(\boldsymbol{y}_i; \boldsymbol{\psi}) = r_B(\boldsymbol{y}_i; \hat{\boldsymbol{\psi}})$. Ze způsobu určování těchto odhadů je zřejmé, že je podstatné, aby model velmi dobře odpovídal zejména na hranicích určujících různé skupiny, tedy v okolí množiny

$$\{\boldsymbol{y} \quad : \quad \pi_j f(\boldsymbol{y}; \boldsymbol{\theta}_j) = \pi_l f(\boldsymbol{y}; \boldsymbol{\theta}_l), \qquad j < l = 2, \ldots, g\}. \tag{4.2}$$

Tudíž například pro vzájemně odlehlejší skupiny je důležité jen, aby model přesně odpovídal zejména na krajích rozdělení jednotlivých skupin. Přesnost určení, odkud měření pochází, je tedy dána v podstatě vlastností jak úspěšně dokáže $r_B(\boldsymbol{y}_i; \hat{\boldsymbol{\psi}})$ přiřazovat do správných skupin pozorování nejasného původu.

4.1 O nezdravých hodnotách zdravých lidí

V této poslední části mé bakalářské práce práce se ještě jednou vrátíme k modelu produkce cytokinu IL10 u lidí trpících paradontózou. Jak jsme si uvedli v sekci 3.2, metodou maximální věrohodnosti s pomocí EM algoritmu jsme získali 7 parametrických vektorů které lokálně maximalizují věrohodnostní funkci (2.1). Po porovnání vlastností, které mají jimi definovaná rozdělení s přirozenou lékařskou představou produkce cytokinu, nám zůstaly tři parametrické vektory $\hat{\boldsymbol{\psi}}_1, \hat{\boldsymbol{\psi}}_2, \hat{\boldsymbol{\psi}}_3$, které by mohly vhodně popisovat náš model.

Každé z těchto rozdělení má dvě dvourozměrné normální složky. Je přirozené předpokládat, že rozdělení, kterým se řídí produkce cytokinu IL10 u zdravých lidí bude to, jehož střední hodnota je v obou složkách menší. V případě rozdělení definované vektory $\hat{\boldsymbol{\psi}}_1$, $\hat{\boldsymbol{\psi}}_2$ a $\hat{\boldsymbol{\psi}}_3$ toto vždy splňuje složka s menším poměrem ve směsi, jak je patrné například z obrázku 3.1. Nyní navážeme na problematiku diskutovanou na počátku této kapitoly a budeme různým měřením přiřazovat skupinu, do které zkoumaní pacienti patří. To budeme provádět na základě pravidla $r_B(\boldsymbol{y}_i; \hat{\boldsymbol{\psi}}_i)$, $i = 1, 2, 3$.

Protože víme, kteří pacienti trpěli paradontózou, můžeme pro porovnání uvážit i odhad parametrického vektoru $\hat{\boldsymbol{\psi}}_c$, který byl vypočten jako směs normálních rozdělení, které byly odhadnuty metodou maximální věrohodnosti pro každou skupinu zvlášť. Tedy jsme odhadli přímo parametry složek $\boldsymbol{\theta}_j$ a jejich poměry pak byly dopočteny jako poměr zastoupení nemocných/zdravých v datech.

Nejmenší počet šesti chyb zaznamenalo přiřazovací pravidlo $r_B(\boldsymbol{y}_i; \hat{\boldsymbol{\psi}}_1)$, které bylo těsně následováno pravidlem $r_B(\boldsymbol{y}_i; \hat{\boldsymbol{\psi}}_3)$, jak je patrné z tabulky 4.1. Třetí v pořadí skončilo $r_B(\boldsymbol{y}_i; \hat{\boldsymbol{\psi}}_c)$ a nejhorší s více jak třetinou špatně určených pacientů bylo $r_B(\boldsymbol{y}_i; \hat{\boldsymbol{\psi}}_2)$. Je zajímavé, že $r_B(\boldsymbol{y}_i; \boldsymbol{\psi}_c)$ skončilo v porovnání s dalšími pravidly o tolik hůře, ačkoliv na výpočet parametrického vektoru určující toto pravidlo bylo použito více informací než pro ostatní odhady. To je ale dáno zejména čtyřmi odlehlými pozorováními „zdravých“ lidí, které ve všech případech byly zařazeny do skupiny nemocných. Pro $r_B(\boldsymbol{y}_i; \hat{\boldsymbol{\psi}}_c)$ měla ale tato pozorování důsledek zvětšení oblasti pro zdravé pacienty, do které se pak vešly i hodnoty pacientů nemocných.

Překvapující je i nula v kolonce špatně určených zdravých lidí u pravidla daného vektorem $\hat{\boldsymbol{\psi}}_3$. To znamená, že každý pacient, u kterého bylo odhadnuto, že je zdravý, byl zdravý. Pokud se podíváme na třetí graf v obrázku 3.1, můžeme si všimnout, že menší složka ze směsi má velmi malý rozptyl a je koncentrovaná u několika málo bodů v levém dolním rohu grafu. V té oblasti se ale nachází pouze 8 měření zdravých pacientů.

Jak bylo zmíněno, cytokin IL10, stejně jako další látky ze skupiny interleukinů jsou produkovány jako imunologická reakce při napadení těla. V našich datech za nemocného pacienta považujeme pacienta trpícího paradontózou. Pochopitelně ale i pro nás zdravý pacient může mít zvýšenou produkci IL10 z důvodu jiného onemocnění, což lze přijmout i za důvod čtyř velmi odlehlých pozorování zdravých pacientů. To by mohlo vést k utvoření hypotézy, že v těchto hodnotách se pohybují čísla úplně zdravých pacientů a že menší složka ze směsi určena odhadem $\hat{\boldsymbol{\psi}}_3$ je odhadem produkce IL10 u této skupiny.

Zmínili jsme, že přiřazovací pravidlo $r_B(\boldsymbol{y}_i; \hat{\boldsymbol{\psi}}_1)$ je ještě o trochu lepší než výše diskutované $r_B(\boldsymbol{y}_i; \hat{\boldsymbol{\psi}}_3)$. Jak je z obrázku 4.1 patrné, výsledné predikované vektory pro jednotlivá pozorování se liší pouze u pěti pacientů. Jejich měření leží relativně někde mezi skupinami s největší koncentrací zdravých a nemocných pacientů, tj. právě v okolí množiny (4.2). Jedná se tedy o měření nejasného původu a přiřazovací pravidla se vlastně liší jen v tom, že jednou tuto malou skupinku přiřadí k nemocným a jednou ke zdravým. Protože v ní jsou ale 3 hodnoty zdravých pacientů, tak pravidlo $r_B(\boldsymbol{y}_i; \hat{\boldsymbol{\psi}}_1)$ vychází v hodnocení lépe.

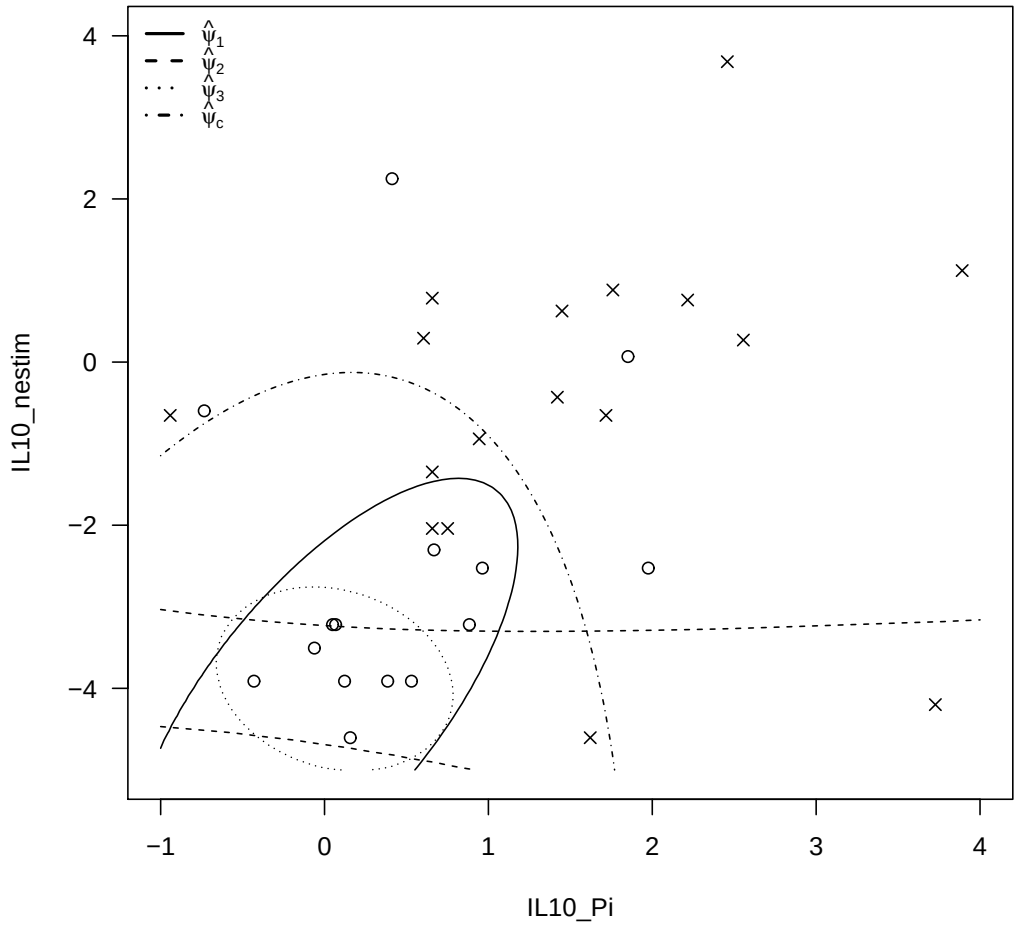
Bylo by předčasné dělat závěry ohledně výsledného tvaru směsi, ty by totiž měly být provedeny až na základě nějakých seriózních testů, které jsme v této práci neodvozovali ani neprováděli. Na druhou stranu si lze na základě naší

Pravidlo	Správně určení		Špatně určení		Počet chyb
	Zdraví	Nemocní	Zdraví	Nemocní	
$\boldsymbol{\psi}_1$	11	15	2	4	6
$\boldsymbol{\psi}_2$	6	15	2	9	11
$\boldsymbol{\psi}_3$	8	17	0	7	7
$\boldsymbol{\psi}_c$	11	12	5	4	9

Tabulka 4.1: Chybovost určování do skupin.

analýzy vytvořit hlubší představu o modelu a povaze pozorování. Pro skutečné použití našeho modelu v praxi by pak navíc byl potřeba odhad na základě náhodného výběru z celé populace, což výběr, se kterým jsme pracovali nesplňoval. Data totiž jsou primárně určena pro ukázaní vztahu zvýšené produkce interleukinů u pacientů trpící paradontózou a v zájmu výzkumných cílů paní doktorky Bártlové, která data naměřila, bylo získat hodnoty produkce u jistého počtu lidí v každé skupině. Proto je podíl nemocných a zdravých lidí v našem datovém souboru podobný, což pochopitelně neodpovídá realitě. Na druhou stranu by tento problém šel vyřešit prostým upravením složek směsi tak, aby u směsi reprezentující skupinu nemocných byl poměr roven velikosti rozšíření nemoci v populaci.

I přesto věřím, že se nám na tomto příkladě podařilo ilustrovat použití modelu směsi jako způsob analýzy výskytu nějaké vlastnosti v populaci.



Legenda: křížek znázorňuje hodnotu pacienta s paradontózou, kolečko zdravého.

Obrázek 4.1: Hranice přiřazovacího pravidla pro rozdělení definované uvažovanými parametrickými vektory $\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \boldsymbol{\psi}_3$ a $\boldsymbol{\psi}_c$.

Závěr

Jsme na konci bakalářské práce, která měla za cíl představit model konečné směsi a ukázat cestu, jakým se dá následně v praxi použít. Zejména jsme se věnovali odhadování modelu pomocí metody maximální věrohodnosti, při které ale může nastat množství komplikací. Problematická se nám ukázala hlavně vlastnost, že věrohodnostní funkce náhodného výběru nemusí být omezená, kvůli čemuž jsme se zaměřili na hledání lokálních maxim věrohodnostní funkce pomocí EM algoritmu. Ačkoliv tyto odhady pak nemaximalizují hodnotu věrohodnostní funkce, lze ukázat, že mají podobné vlastnosti jako odhady metodou maximální věrohodnosti. Bohužel už nám nezbyl prostor se touto velmi důležitou teorií zabývat, čímž by se použitelnost konečných směsí ukázala v plné síle. Zde vidím prostor, kam by se tato práce mohla dále rozvíjet.

V průběhu této práce jsme se detailně věnovali EM algoritmu, znalost jehož principu je pro pochopení funkcionality stěžejní. V rámci druhé kapitoly, která se odvození algoritmu věnuje, je zachycena i elegantní myšlenka, kterou algoritmus využívá. V dalších částech textu jsou pak odvozeny vzorce pro iteraci EM algoritmu aplikovaného na nejběžněji používaný model normálních směsí. V rámci toho jsme kladli důraz na splnění postačujících podmínek pro maximum funkce, když veškerá literatura, která se podobnému odvození věnovala a se kterou autor přišel do styku, hledala řešení pouze podmínek nutných. Funkčnost EM algoritmu se nám pak podařilo dokázat pouze pro nejvýše dvourozměrné normální směsi. Ve vyšší dimenzi se o ni můžeme pouze domnívat. To ale bylo postačující s ohledem na naši následnou analýzu v první kapitole uvedeného modelu, kde jsme zkoumali dvourozměrnou normální směs.

Při výběru zadání bakalářské práce autor kladl důraz na možnost prakticky aplikovat vypsané téma, které v ní mělo být náležitě vystiženo. Jsem přesvědčen, že model produkce interleukinu 10 jakožto obranné reakce lidského těla při napadení nějakou nemocí názorně ilustroval jedno z možných použití směsi. Souvislost produkce interleukinů při napadení paradontózou je aktuálně předmětem bádání paní doktorky Bártlové a pana doktora Běláčka z 1. lékařské fakulty Univerzity Karlovy a já věřím, že model směsi jim umožní trochu jiný náhled na jimi zkoumaný problém. Použití modelu směsi k ukázání jistých vztahů mezi produkcí cytokinu a onemocnění paradontózou už by pak byla taková třešnička na dortu této práce. Je ale zřejmé, že pro případné publikování výsledků by bylo nutné v analýze pokračovat, což součástí této práce nebude.

Pevně věřím, že tato práce poskytne všem čtenářům jednoduše pochopitelný přehled o problematice konečných směsí a jejím možném využití a zanechá v nich přesvědčení, že se jí vyplatí, s ohledem na její praktičnost i eleganci, věnovat.

Literatura

Aitkin, M. (2001). Likelihood and bayesian analysis of mixtures. *Statistical Modelling*, **1**, 287–304.

Chen, W.-C., Maitra, R. a Melnykov, V. (2012). EMCluster: EM algorithm for model-based clustering of finite mixture gaussian distribution. R Package, URL http://cran.r-project.org/package=EMCluster.

Dattorro, J. (2015). *Convex Optimization, Euclidean Distance Geometry*. Meboo Publishing USA. ISBN 978-0-578-16140-2.

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.

McLachlan, G. a Peel, D. (2000). *Finite Mixture Model*. John Wiley & Sons, Inc., New York. ISBN 0-471-00626-2.

Pearson, K. (1894). Contributions to the theory of mathematical evolution. *Philosophical Transactions of the Royal Society of London A*, **185**, 71–110.

Seznam obrázků

Seznam tabulek