UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA
KATEDRA MATEMATICKÉ ANALÝZY A APLIKACÍ MATEMATIKY

DIPLOMOVÁ PRÁCE

Odhady parametrů smíšených distribučních funkcí náhodného vektoru

Vedoucí diplomové práce:
Prof. RNDr. Ing. Lubomír Kubáček, DrSc., Dr.h.c.
Rok odevzdání: 2010

Vypracoval:
Bc. Tomáš Hanzlíček
AME, II. ročník

Prohlášení

Prohlašuji, že jsem vytvořil tuto diplomovou práci samostatně za vedení a pomoci profesora Kubáčka a že jsem v seznamu použité literatury uvedl všechny zdroje použité při zpracování práce.

V Olomouci dne 10. března 2010

Poděkování

Rád bych na tomto místě poděkoval profesoru Kubáčkovi jako mému vedoucímu diplomové práce za obětavou spolupráci i za čas, který mi věnoval při konzultacích. Dále si zasluží poděkování můj počítač, že vydržel moje pracovní tempo, typografický systém TEX, kterým je práce vysázena a program Matlab, se kterým jsem zpracovával data.

Obsah

Úvodní slovo

V dnešní době je statistika a jí příbuzné disciplíny nedílnou součástí našich životů. Ať už si to uvědomujeme či ne, setkáváme se s jejími důsledky a aplikacemi prakticky každý den. Při studiu, v zaměstnání, při sledování televize, při čtení novin nebo v konverzaci s ostatními lidmi. Proto je určitě přínosné zajímat se ze statistického hlediska hlouběji vztahy a zákonitostmi, které se vyskytují kolem nás. Pomáhá nám to pochopit různé přírodní či společenské jevy. Aplikace statistiky lze najít snad ve všech vědních odvětvích počínaje matematikou, ekonomií přes medicínu, techniku, astronomii a psychologií konče. Vzhledem k rozmanitosti, která je typická tomuto světu, se statistické metody rozvinuly do obrovské šíře a lze předpokládat další rozvoj. Zde se budeme zabývat smíšenými distribucemi.

Smíšené distribuce jsou poměrně mladou oblastí matematické statistiky. Největší rozvoj zažily teprve až ve druhé polovině 20. století. V současnosti díky možnostem výpočetní techniky je možno smíšené distribuce pohodlně využívat při řešení rozmanitých problémů. V této práci se seznámíme se smíšenými distribucemi jako s užitečným nástrojem modelování různých jevů a dat. Ukážeme si základní definici smíšené distribuce, její různé typy a tvary a zaměříme se na metody odhadů parametrů smíšených distribucí. Na závěr si uvedeme příklad použití smíšené distribuce v medicínské praxi.

Cílem práce je seznámit čtenáře se základními vztahy v oblasti smíšených distribucí, s metodami odhadů jejich parametrů a ukázat jejich praktickou aplikaci.

Práce je zaměřena především na aplikaci ne zcela standardního algoritmu na reálná data s medicínsky smysluplnou interpretací. Zároveň se snaží být i jakýmsi vodítkem či návodem pro čtenáře, který při řešení svých úkolů potřebuje pomoc.

1 Úvod do problematiky smíšených distribucí

1.1 Historický náhled

Jednu z prvních významnějších analýz, které zahrnovaly použití smíšených distribucí, provedl na konci 19. století známý anglický biometrik Karl Pearson. Ve své práci modeloval smíšenou distribuci ze dvou hustot normálních rozdělení se středními hodnotami μ_1 a μ_2 a rozptyly σ_1^2 a σ_2^2 v proporcích π_1 a π_2 z dat, které mu poskytl jeho kolega W.F.R. Weldon. Později se ukázalo, že jeho práce byla první, která prosazovala statistickou analýzu jako primární metodu studia biologických problémů. Data, která Pearson analyzoval, se skládala z výsledků měření poměru čela k tělesné výšce člověka. K dispozici měl 1000 měření z oblasti neapolského zálivu. Měření bylo rozděleno do 29 intervalů a zobrazeno v histogramu, ve kterém se projevila určitá šikmost. Z tohoto pozorování Weldon usuzoval, zda není možné, že měřená populace se vyvíjí ve dva nové poddruhy. Na základě tohoto zjištění se pak obrátil na Pearsona pro pomoc s analýzou dat.

Pearson modeloval data metodou momentů. Dnes se používá vhodnější metoda maximální věrohodnosti, kterou si přiblížíme i v této práci. V jeho modelu se ukázalo, že tato smíšená distribuce dvou normálních rozdělení přesně splňuje Pearsonův záměr, jenž byl modelovat zjevnou šikmost projevenou v histogramu, kterou nelze adekvátně modelovat symetrickým normálním rozdělením. Odhady parametrů svého modelu ($\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$ a π_1; π_2) vypočítal jako kořeny polynomu stupně 9, což na konci 19. století nebylo jednoduchým úkolem. Není překvapivé, že další vědci se během let pokoušeli zjednodušit Pearsonův postup, například C.V.L. Charlier. S nástupem počítačů se pozornost přesunula na odhadování parametrů smíšených distribucí metodou maximální věrohodnosti, kterou se zabývali Jeffreys, Rao, Hasselblad, Dempster či Aitkin. V posledních letech se smíšenými distribucemi zabývali například Lindsay Böhning nebo Wedel a Kamakura, kteří se zabývali aplikací smíšených distribucí v marketingu. Vznikly také nové metody odhadování parametrů. My se seznámíme s tzv. *EM algoritmem*. Další podrobnosti lze nalézt například v [8].

1.2 Základní definice

Smíšené distribuce poskytly matematicky založený přístup ke statistickému modelování široké škály různých jevů. Díky užitečnosti, která je dána značnou flexibilitou při modelování, získaly modely smíšených distribucí značnou pozornost nejen z praktického, ale i z teoretického úhlu pohledu. V posledních letech se rozsah a potenciál aplikací smíšených distribucí široce rozšířil. Aplikace smíšených distribucí se uplatnily například v astronomii, biologii, genetice, medicíně, psychiatrii, ekonomii či marketingu. Teorie v této kapitole je založena především na knihách [6, 8].

Definice 1.1 Nechť $f_i(\boldsymbol{y}; \boldsymbol{\theta}_i)$ je hustota p-rozměrného náhodného vektoru $\boldsymbol{Y}_i$ a $\pi_i > 0, \quad \forall i = 1, 2, \ldots, c,$ nechť platí $\pi_1 + \pi_2 + \ldots + \pi_c = 1$. Pak hustotu

$$f(\boldsymbol{y}) = \sum_{i=1}^{c} \pi_i f_i(\boldsymbol{y}; \boldsymbol{\theta}_i) \tag{1}$$

nazýváme *smíšenou distribucí.*

Parametry π_i reprezentují váhu i-té složky smíšené distribuce náhodného vektoru $\boldsymbol{Y}_i$, $f_i(\boldsymbol{y}; \boldsymbol{\theta}_i)$ je hustota složky s parametry zastoupenými vektorem $\boldsymbol{\theta}_i$ a c je počet složek modelu.

Funkční hodnotu $f(\boldsymbol{y})$ pak spočítáme tak, že spočítáme funkční hodnoty $f_i(\boldsymbol{y}; \boldsymbol{\theta}_i)$ v bodě $\boldsymbol{y}$, vynásobíme je příslušnými vahami a sečteme je.

V této formulaci modelu je počet složek modelu stanoven. Může se ale stát, že je počet složek neznámý, a pak je nutno jej zjistit z dat spolu s dalšími parametry.

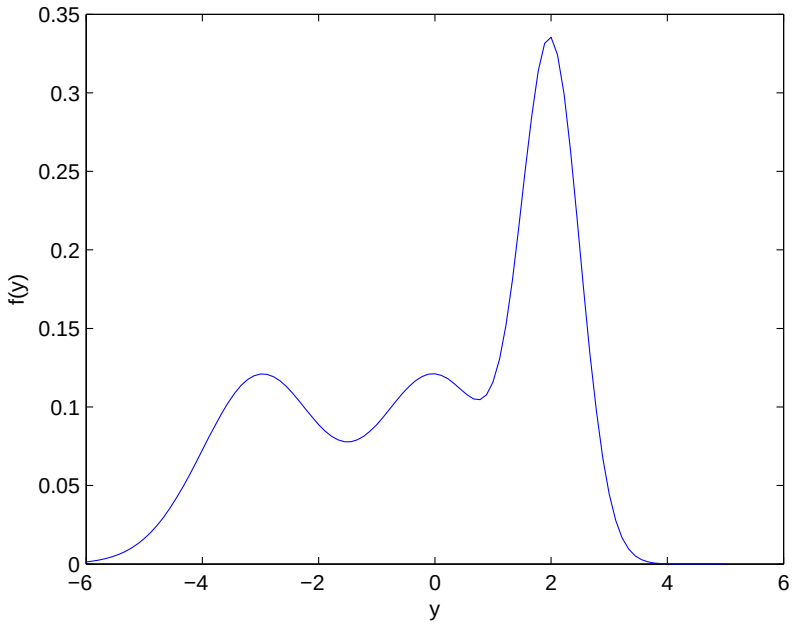
V další části této kapitoly se nejprve budeme krátce věnovat smíšeným distribucím vytvořených z jednorozměrných náhodných veličin a pak se blíže seznámíme s vícerozměrnými smíšenými distribucemi

Pro názornost si uvedeme příklad.

Příklad 1.1 Nakreslete graf smíšené distribuce dané vztahem

$$f(y) = 0.3 \times \phi(y; -3, 1) + 0.3 \times \phi(y; 0, 1) + 0.4 \times \phi(y; 2, 0.5), \tag{2}$$

kde $\phi(y; \mu, \sigma^2)$ reprezentuje hodnotu hustoty normálního rozdělení v bodě y s danou střední hodnotou μ a rozptylem σ^2.



Obrázek 1: Hustota smíšené distribuce složené ze 3 komponent

Z modelu je patrné, že váhy jednotlivých složek splňují podmínky, a že hustoty složek jsou vycentrovány v bodech -3, 0, resp. 2. Graf této distribuce je zobrazen na obrázku 1. Kód pro vytvoření grafu v Matlabu je uveden v příloze A1.

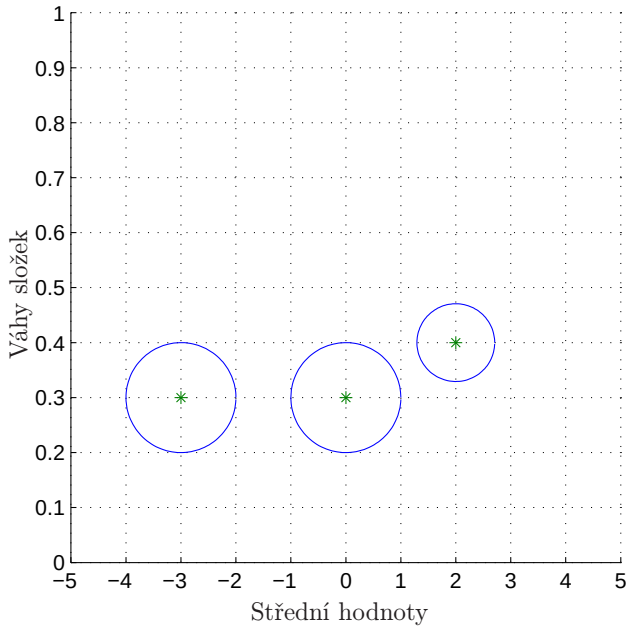
Obecně složky smíšených distribucí mohou být jakéhokoliv typu. Tzn. spojité nebo diskrétní. Zde se budeme zabývat pouze případy složek se spojitým rozdělením, speciálně s normálním rozdělením.

1.3 Vizualizace smíšených distribucí

Ukážme si nejprve, jakým způsobem lze zobrazit podkladovou strukturu smíšených distribucí. Strukturou máme na mysli počet složek ve spojení se středními hodnotami a rozptyly. V podstatě se snažíme vizualizovat vícerozměrný parametrický prostor ($\mu_1, \ldots, \mu_c, \sigma_1^2, \ldots, \sigma_c^2$ a $\pi_1, \ldots, \pi_{c-1}$, tzn. $3c-1$ parametrů) ve dvojrozměrné reprezentaci. Způsob, který si zde uvedeme se anglicky nazývá dF plot.

Ve vizualizaci pomocí dF plotu je každá složka znázorněna kruhem, který je umístěn v bodě se souřadnicemi $[\mu_i, \pi_i]$ a jehož poloměr je dán standardní odchylkou, jejíž velikost zjišťujeme ve směru osy střeních hodnot.

Nyní si uveďme příklad dF plotu (viz obrázek 2), který odpovídá modelu z příkladu 1.1. První kruh zleva koresponduje se složkou s váhou $\pi_1 = 0.3$ a



Obrázek 2: dF plot modelu třísložkové smíšené distribuce

střední hodnotou $\mu_1 = -3$. Jeho poloměr je roven 1. Podobně střední (resp. pravý) kruh reprezentují druhou (resp. třetí) složku. Všimněme si, že vizualizace nám pomáhá zjistit už na první pohled, které složky mají nejvetší váhu a kde jsou umístěny. Zdrojový kód obrázku viz příloha A2.

1.4 Smíšené distribuce dvou normálních rozdělení se stejným rozptylem

V této části si ukážeme, jak vypadají některé hustoty dvousložkových smíšených distribucí normálních rozdělení v proporcích π_1 a π_2, uvažujeme-li stejný

rozptyl σ^2 a střední hodnoty μ_1 a μ_2. Odpovídající hustota má tvar

$$f(y) = \pi_1\phi(y; \mu_1, \sigma^2) + \pi_2\phi(y; \mu_2, \sigma^2), \tag{3}$$

kde

$$\phi(y; \mu, \sigma^2) = (2\pi)^{-\frac{1}{2}}\sigma^{-1}\exp\left[-\frac{1}{2}(y-\mu)^2/\sigma^2\right] \tag{4}$$

je hustota normálního rozdělení se střední hodnotou μ a rozptylem σ^2.

Pokud se střední hodnoty od sebe dostatečně liší, lze očekávat, že hustota $f(y)$ se bude podobat dvěma hustotám položeným vedle sebe. Taková hustota je bimodální. V případě, že jsou si střední hodnoty blízko, je hustotu unimodální. Na ukázku si uvedeme pár grafů hustot pro různé hodnoty Δ v případě, že $\mu_1 = 0$, $\mu_2 = \Delta$, $\sigma = 1$ a váhy složek jsou si rovny, tj. $\pi_1 = \pi_2 = 0.5$ (viz obrázek 3, kód příloha A3). Na této ukázce lze vidět, že s rostoucím Δ se tvar hustoty mění z unimodálního na bimodální. Hranice této změny je v našem případě znatelná pro $\Delta = 3$.

Obecně Δ definujeme vztahem

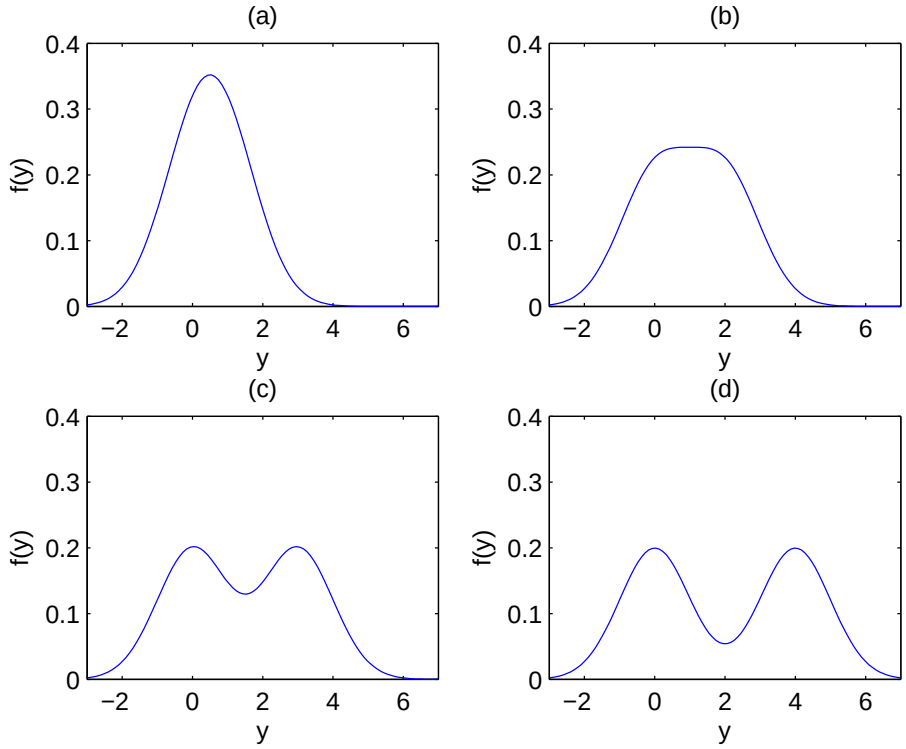
$$\Delta = |\mu_1 - \mu_2| / \sigma \tag{5}$$

a nazýváme Mahalanobisova vzdálenost mezi složkami smíšené distribuce normálních rozdělení se stejným rozptylem.

Jestliže jsou si střední hodnoty v modelu blízké, tak překrytí hustot složek inklinuje k zastínění rozdílu mezi nimi (viz obrázky 3 (a),(b)). Obecně výsledkem nebude symetrická hustota, pokud ovšem si váhy složek nejsou rovny (viz obrázek 3). Tento fakt je zobrazen na obrázku 4 (kód analogicky jako v příloze A3), kde jsme použili stejné střední hodnoty a rozptyl, ale změnili jsme hodnoty vah složek na $\pi_1 = 0.75$ a $\pi_2 = 0.25$.

V modelu lze také určit přesně vyjádření šikmosti γ_1 a špičatosti γ_2 hustoty $f(y)$ a to následovně (viz [8]),

$$\gamma_1 = \frac{a(a-1)\Delta^3}{\{a\Delta^2 + (a+1)^2\}^{3/2}} \tag{6}$$



Obrázek 3: Srovnání grafů hustot dvousložkových smíšených distribucí se stejnými vahami složek, stejným rozptylem $\sigma = 1$ a středními hodnotami $\mu_1 = 0$, $\mu_2 = \Delta$ v různý variantách Δ: a) $\Delta = 1$, b) $\Delta = 2$, c) $\Delta = 3$, d) $\Delta = 4$

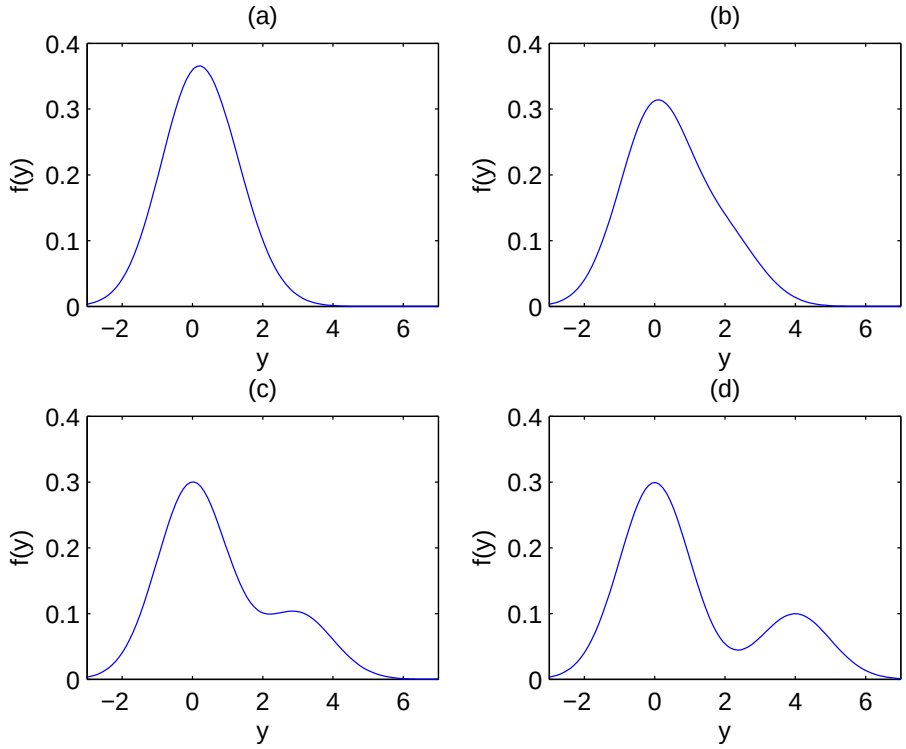
a

$$\gamma_2 = \frac{a(a^2 - 4a + 1)\Delta^4}{\{a\Delta^2 + (a+1)^2\}^2}, \tag{7}$$

kde Δ je Mahalanobisova vzdálenost a a je poměr větší váhy složky k menší. Situace se poněkud zkomplikuje, když budeme uvažovat více složek v modelu a nestejné rozptyly. To bude předmětem další sekce.

1.5 Smíšené distribuce normálních rozdělení s nestejnými rozptyly

Užitečnou publikaci charakterizující tvary hustot smíšených distribucí dvou normálních rozdělení s nestejnými rozptyly napsal I. Eisenberger (viz [2]). Napří-



Obrázek 4: Srovnání grafů hustot dvousložkových smíšených distribucí s vahami složek $\pi_1 = 0.75$ a $\pi_2 = 0.25$, stejným rozptylem $\sigma = 1$ a středními hodnotami $\mu_1 = 0$, $\mu_2 = \Delta$ v různý variantách Δ: a) $\Delta = 1$, b) $\Delta = 2$, c) $\Delta = 3$, d) $\Delta = 4$

klad určil, kdy hustota bude bimodální a kdy ne. Platí-li nerovnost

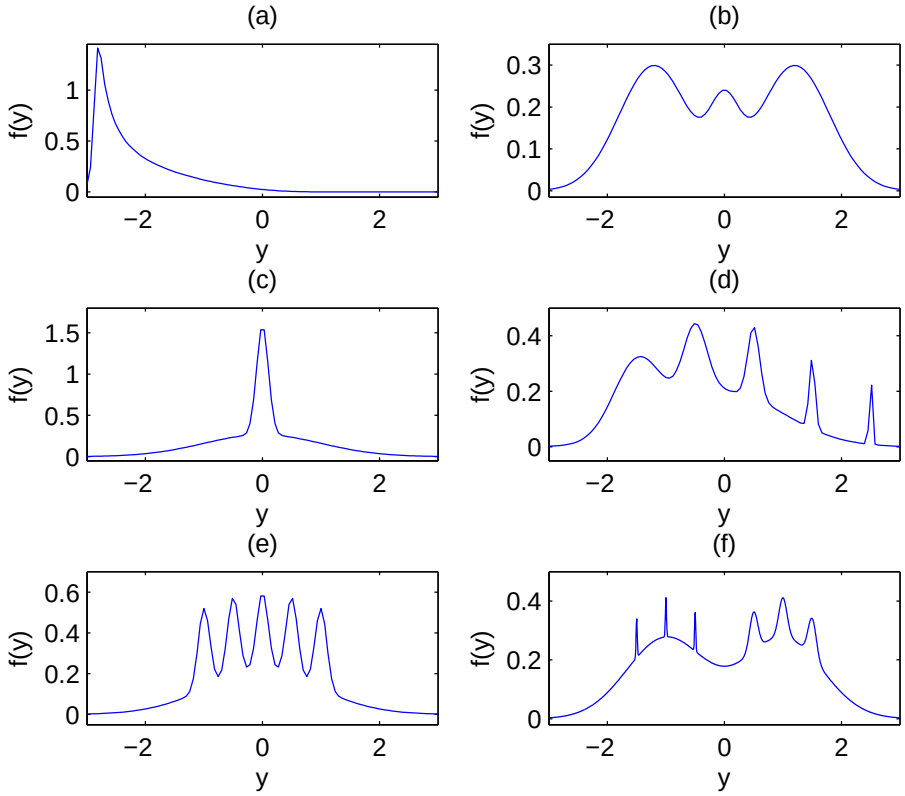
$$\Delta^2 < (27\sigma_2^2)/\left\{4(1+k)\right\}, \tag{8}$$

kde $k = \sigma_2^2/\sigma_1^2$, pak hustota $f(y)$ nemůže být bimodální. Naopak, platí-li opačná nerovnost

$$\Delta^2 > (27\sigma_2^2)/\left\{4(1+k)\right\}, \tag{9}$$

pak existuje hodnota π_1, pro kterou je hustota $f(y)$ bimodální.

Problematika c-složkových smíšených distribucí normálních rozdělení je velmi široká. Abychom si ukázali proměnlivost a flexibilitu, uvedeme si pár příkladů grafů hustot pro různé počty složek, středních hodnot a rozptylů (viz obrázek 5 a tabulka 1, kód v Matlabu vytvoříme analogicky jako v příloze A3).



Obrázek 5: Grafy hustot různých smíšených distribucí: a) Silně šikmá, b) Trimodální, c) Špičatá unimodální, d) Asymetrický dráp , e) Dráp, f) Asymetrický dvojitý dráp

Tabulka 1: Smíšené distribuce

Hustota	$f(y)$
a) Silně šikmá	$\sum_{i=0}^{7} \frac{1}{8} N(3\left\{(\frac{2}{3})^i - 1\right\}, (\frac{2}{3})^{2i})$
b) Trimodální	$\frac{9}{20}N(-\frac{6}{5}, (\frac{3}{5})^2) + \frac{9}{20}N(\frac{6}{5}, (\frac{3}{5})^2) + \frac{1}{10}N(0, (\frac{1}{4})^2)$
c) Špičatá unimodální	$\frac{2}{3}N(0,1) + \frac{1}{3}N(0, (\frac{1}{10})^2)$
d) Asymetrický dráp	$\frac{1}{2}N(0,1) + \sum_{i=-2}^{2}(2^{1-i}/31)N(i + \frac{1}{2}, (2^{-i}/10)^2)$
e) Dráp	$\frac{1}{2}N(0,1) + \sum_{i=0}^{4} \frac{1}{10} N(i/2 - 1, (\frac{1}{10})^2)$
f) Asymetrický dvojitý dráp	$\sum_{i=0}^{1} \frac{46}{100} N(2i - 1, (\frac{2}{3})^2) +$ $\sum_{i=1}^{3} \frac{1}{300} N(-i/2, (\frac{1}{100})^2) +$ $+ \sum_{i=1}^{3} \frac{7}{300} N(i/2, (\frac{7}{100})^2)$

1.6 Vícerozměrné smíšené distribuce

V této části si ukážeme jak vypadají smíšené distribuce složené z hustot náhodných vektorů z normálních rozdělení. Obecný tvar vypadá následovně

$$f(\boldsymbol{y}) = \pi_1 f_1(\boldsymbol{y}; \boldsymbol{\mu}_1, \Sigma_1) + \pi_2 f_2(\boldsymbol{y}; \boldsymbol{\mu}_2, \Sigma_2) + \ldots + \pi_c f_c(\boldsymbol{y}; \boldsymbol{\mu}_c, \Sigma_c), \quad \boldsymbol{y} \in \mathbb{R}^p. \quad (10)$$

za podmínek $\sum_{i=1}^{c} \pi_i = 1, \pi_i > 0$. Vektor $\boldsymbol{\mu}_i$ symbolizuje vektor středních hodnot i-té složky distribuce a matice Σ_i je její varianční matice.

Příklad 1.2 Ukažme si jak vypadá smíšená distribuce dvou složek z dvourozměrného normálního rozdělení. Obecný předpis této smíšené distribuce je

$$f(\boldsymbol{y}) = \pi\phi(\boldsymbol{y}; \boldsymbol{\mu}_1, \Sigma_1) + (1 - \pi)\phi(\boldsymbol{y}; \boldsymbol{\mu}_2, \Sigma_2), \quad (11)$$

kde

$$\phi(\boldsymbol{y}; \boldsymbol{\mu}, \Sigma) = (2\pi)^{-\frac{p}{2}} |\Sigma|^{-\frac{1}{2}} exp\left\{-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{y} - \boldsymbol{\mu})\right\} \quad (12)$$

a parametry mají tvar

$$\boldsymbol{\mu}_1 = \begin{pmatrix} \mu_{11} \\ \mu_{12} \end{pmatrix}, \boldsymbol{\mu}_2 = \begin{pmatrix} \mu_{21} \\ \mu_{22} \end{pmatrix}, \Sigma_1 = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}, \Sigma_2 = \begin{pmatrix} \sigma'_{11} & \sigma'_{12} \\ \sigma'_{21} & \sigma'_{22} \end{pmatrix}, \pi \in (0, 1).$$

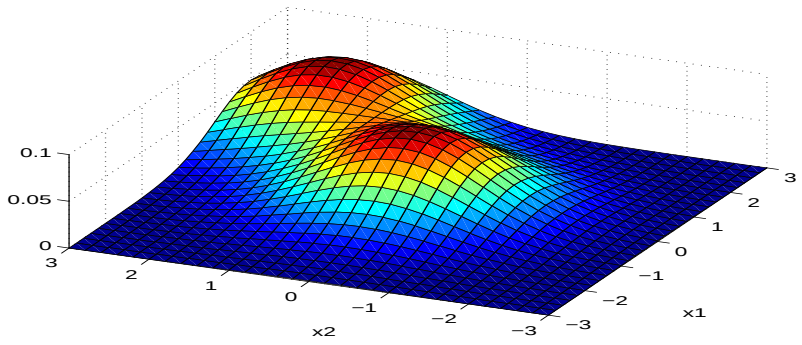
Poznámka: V rovnici (11) je π parametr, ve vztahu (12) je π konstanta.

Ukažme si tedy, jak bude vypadat smíšená distribuce s těmito paramety (viz obrázek 6, kód příloha A4):

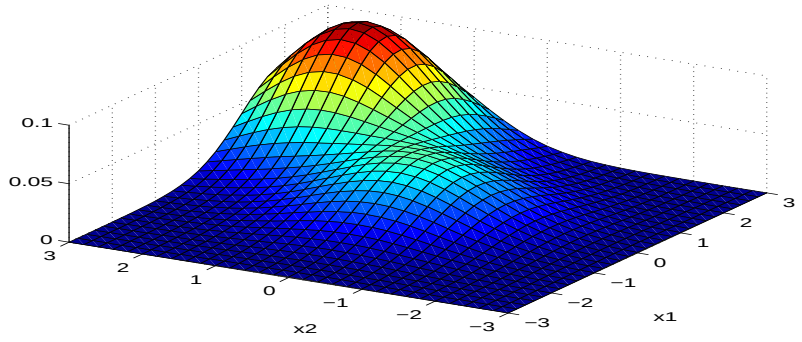
$$\boldsymbol{\mu}_1 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \boldsymbol{\mu}_2 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \pi = 0,5.$$

Vidíme, že obě složky se graficky projeví jako pravidelné „kopce“, které se liší pouze svým umístěním. Pravidelnost je dána volbou variačních matic, ve kterých je nulová kovariance proměnných.

Teď změníme váhy složek tak, že první bude mít váhu 0,7 a druhá 0,3. Ostatní parametry zatím ponecháme. Pozměněný graf smíšené distribuce vidíme na obrázku 7 (kód analogicky jako A4).



Obrázek 6: Dvousložková smíšená distribuce dvourozměrného nekorelovaného náhodného vektoru se stejnými vahami složek

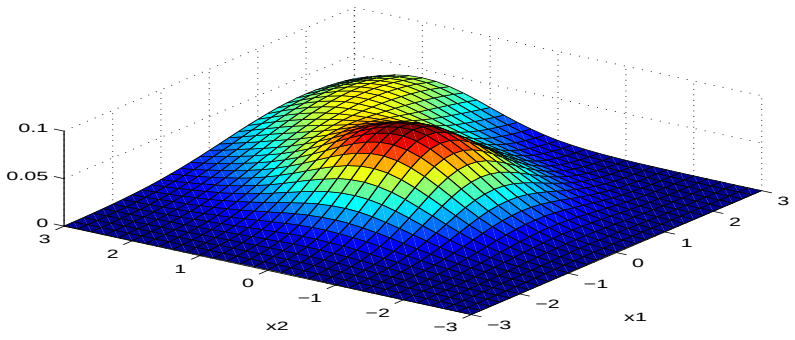


Obrázek 7: Dvousložková smíšená distribuce dvourozměrného nekorelovaného náhodného vektoru s dominantní složkou

Vidíme, že se, stejně jako v jednorozměrném případě, pouze snížil „kopec“ představující druhou složku. Podobnou změnu vyvolá změna rozptylu některé ze složek náhodného vektoru. Například změna parametru σ_{11} z první distribuce z hodnoty 1 na $2,5$ způsobí zploštění a protažení (ve směru osy $x1$) části grafu připadající první složce smíšené distribuce (obrázek 8, kód analogicky jako A4).

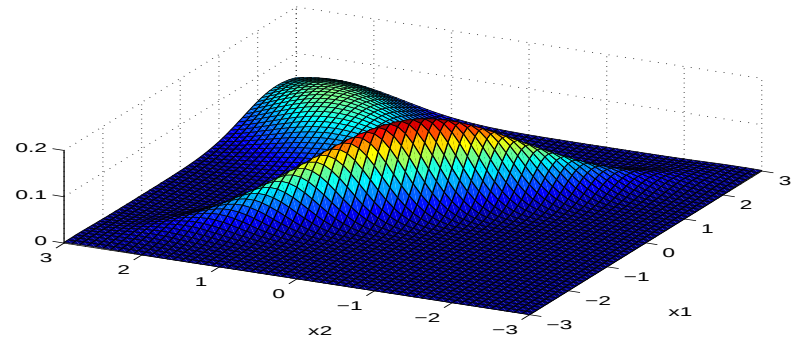
Zajímavější jsou změny grafů, když měníme varianční matice složek. Změny středních hodnot ovlivní umístění „kopce“ v souřadnicové soustavě. Vykresleme graf smíšené distribuce s parametry (viz obrázek 9, kód analogicky jako A4):

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \boldsymbol{\mu}_2 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma_1 = \begin{pmatrix} 1 & 0,5 \\ 0,5 & 1 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 1 & -0,9 \\ -0,9 & 1 \end{pmatrix}, \pi = 0,5.$$



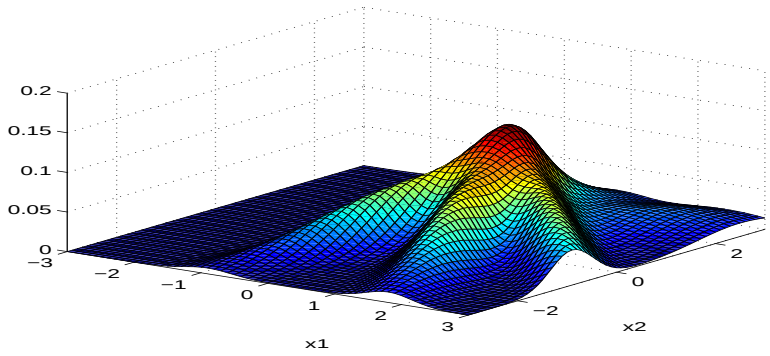
Obrázek 8: Dvousložková smíšená distribuce dvourozměrného nekorelovaného náhodného vektoru s různými hodnotami rozptylů

Můžeme pozorovat, že oba „kopce“ se protáhly. Směr a intenzitu protáhnutí určují varianční matice respektive korelace mezi proměnnými. První složka má kladnou korelaci $0,5$, což se projeví v protažení ve směru osy prvního a třetího kvadrantu. Druhá složka má zápornou a vysokou korelaci, což způsobí, že „kopec“ je více protáhlý než u první složky a protažení ve směru osy druhého a čtvrtého kvadrantu. Připomínám, že při nulové korelaci jsou „kopce“ pravidelně kulaté. Na závěr si ukažme ještě příklad pětisložkové smíšené distribuce (viz obrázek 10,



Obrázek 9: Dvousložková smíšená distribuce dvourozměrného náhodného vektoru s kladnou i zápornou korelací

kód analogicky jako A4).

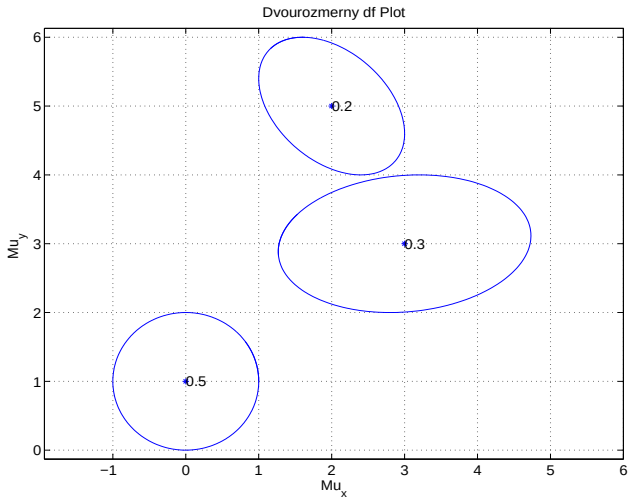


Obrázek 10: Pětisložková smíšená distribuce dvourozměrného náhodného vektoru

Stejně jako u smíšených distribucí z jednorozměrného rozdělení můžeme ke znázornění vícerozměrné distribuce použít *dF* plot. V dvourozměrném případě bude *dF* plot vypadat podobně. Zobrazíme proto na obrázku 11 (kód analogicky jako A2) *dF* plot pro model:

$$\pi_1 = 0,5, \quad \pi_2 = 0,3, \quad \pi_3 = 0,2, \quad \boldsymbol{\mu}_1 = (0 \quad 1)^T, \quad \boldsymbol{\mu}_2 = (3 \quad 3)^T, \quad \boldsymbol{\mu}_3 = (2 \quad 5)^T$$

$$\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \Sigma_2 = \begin{pmatrix} 3 & 0,2 \\ 0,2 & 1 \end{pmatrix}, \quad \Sigma_3 = \begin{pmatrix} 1 & -0,4 \\ -0,4 & 1 \end{pmatrix},$$



Obrázek 11: *dF* plot dvourozměrné smíšené distribuce

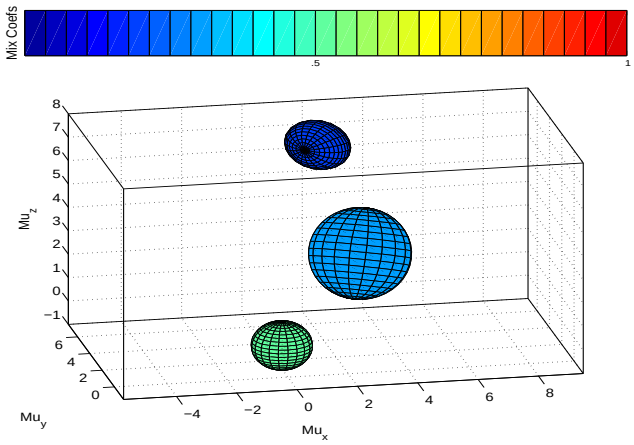
Střední hodnoty jsou reprezentovány umístěním v souřadnicové soustavě. Varianční struktura v modelu je popsána elipsami se středy určenými středními hodnotami. Pro úplnost zkusme ještě vykreslit *dF* plot pro třírozměrnou třísložkovou smíšenou distribuci (viz obrázek 12, kód analogicky jako A2) danou modelem:

$$\pi_1 = 0,5, \quad \pi_2 = 0,3, \quad \pi_3 = 0,2,$$

$$\boldsymbol{\mu}_1 = (0 \quad 1 \quad 0)^T, \quad \boldsymbol{\mu}_2 = (3 \quad 3 \quad 3)^T, \quad \boldsymbol{\mu}_3 = (2 \quad 5 \quad 7)^T$$

$$\Sigma_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \Sigma_2 = \begin{pmatrix} 3 & 0,2 & 0 \\ 0,2 & 1 & 1 \\ 0 & 1 & 3 \end{pmatrix}, \quad \Sigma_3 = \begin{pmatrix} 1 & -0,4 & 0 \\ -0,4 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Interpretace obrázku je analogická dvourozměrnému případu. Střední hodnoty



Obrázek 12: *dF* plot dvourozměrné smíšené distribuce

určují polohu elipsoidů, které jsou determinovány varianční strukturou. Barva znázorňuje váhu složek.

Ukázali jsme si, že smíšené distribuce normálních rozdělení náhodných vektorů se vyznačují velkou proměnlivostí. Jejich hustoty se mohou značně měnit už při malých změnách vah, středních hodnot či rozptylů (resp. variančních matic). Důležité je také určení počtu složek distribuce. Nyní se podívejme na to, jak lze určit parametry v modelech smíšených distribucí, máme-li k dispozici naměřená data.

2 Metody určování parametrů

Definice 2.1 Nechť $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ je náhodný výběr rozsahu n příslušný statistickému znaku Y. Výběrovou funkci

$$\overline{Y_n} = \frac{1}{n} \sum_{j=1}^{n} Y_j \tag{13}$$

nazýváme *výběrový průměr* a funkci

$$S_n^2 = \frac{1}{n-1} \sum_{j=1}^{n} (Y_j - \overline{Y_n})^2 \tag{14}$$

nazýváme *empirický rozptyl.*

Definice 2.2 Nechť $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$ je náhodný výběr rozsahu n z p-rozměrného normálního rozdělení. Výběrovou funkci

$$\overline{\boldsymbol{Y}}_n = \frac{1}{n} \sum_{j=1}^{n} \boldsymbol{Y}_j \tag{15}$$

nazýváme *výběrový průměr.* Čtvercovou matici rozměru p

$$W = \sum_{j=1}^{n} \left(\boldsymbol{Y}_j - \overline{\boldsymbol{Y}}_n\right) \left(\boldsymbol{Y}_j - \overline{\boldsymbol{Y}}_n\right)^T \tag{16}$$

nazýváme *Wishartova matice s* $n-1$ *stupni volnosti* a matici

$$S = \frac{1}{n-1} W \tag{17}$$

nazýváme *empirická varianční (kovarianční) matice.*

Poznámka 2.1 V některých případech se místo S používá $S' = \frac{1}{n}W$. Matice S a S' resp. $|S|$ a $|S'|$ se používají jako analogie jednorozměrného rozptylu.

Definice 2.3 Nechť $f(\boldsymbol{y}, \boldsymbol{\theta})$ je hustota obecného rozdělení s parametry uspořádanými ve vektoru $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$, pak *Fisherova informační matice pro parametr* $\boldsymbol{\theta}$ je dána vztahem

$$F(\boldsymbol{\theta}) = E_f \frac{\partial}{\partial \boldsymbol{\theta}} \ln f(\boldsymbol{y}, \boldsymbol{\theta}) \left[\frac{\partial}{\partial \boldsymbol{\theta}} \ln f(\boldsymbol{y}, \boldsymbol{\theta})\right]^T. \tag{18}$$

Mějme $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ náhodný výběr rozsahu n, kde $\boldsymbol{Y}_j$ je p-rozměrný náhodný vektor s hustotou $f(\boldsymbol{y}_j)$ na $\mathbb{R}^p$. V praxi $\boldsymbol{Y}_j$ obsahuje náhodné veličiny odpovídající p měřením vykonaných v j-tém opakování určitého jevu. Nechť $\boldsymbol{Y} = (\boldsymbol{Y}_1^T, \ldots, \boldsymbol{Y}_n^T)^T$, kde T znamená vektorovou transpozici. Vidíme, že $\boldsymbol{Y}$ reprezentuje celý vzorek, tj. n-tici bodů v $\mathbb{R}^p$. Realizaci náhodného vektoru budeme označovat malým písmenem. Např. $\boldsymbol{y} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_n^T)^T$ označuje realizaci náhodného výběru, kde $\boldsymbol{y}_j$ je pozorovaná hodnota nahodného vektoru $\boldsymbol{Y}_j$. Dále předpokládejme, že hustota $f(\boldsymbol{y}_j)$ je smíšenou distribucí c složek, tj.

$$f(\boldsymbol{y}_j) = \sum_{i=1}^{c} \pi_i f_i(\boldsymbol{y}_j; \boldsymbol{\theta}_i). \tag{19}$$

Nejprve se zaměříme na náhodný výběr, kde Y_j je jednorozměrná náhodná veličina, která je normálně rozdělená. Dostáváme pak smíšenou distribuci

$$f(y) = \sum_{i=1}^{c} \pi_i \phi(y; \mu_i, \sigma_i^2), \tag{20}$$

kde $\phi(y; \mu_i, \sigma_i^2)$ označuje hustotu normálního rozdělení se střední hodnotou μ_i a rozptylem σ_i^2. V tomto případě neznáme $c-1$ vah jednotlivých složek, c středních hodnot a c rozptylů. Celkem tedy $3c-1$ parametrů. V modelech smíšených distribucí je největší početní břemeno spojeno právě s odhadováním parametrů.

2.1 Metoda maximální věrohodnosti

Metoda maximální věrohodnosti (viz [5]) je jedna z metod, jak určit odhady neznámých parametrů daného rozdělení pravděpodobností.

Nechť $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$ je náhodný výběr z rozdělení spojitého typu s hustotou $f(y; \theta)$, kde $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k) \in \Theta$ je vektor neznámých paramatrů. Hustota náhodného výběru $\boldsymbol{Y}$ je

$$f(\boldsymbol{y}; \boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i; \boldsymbol{\theta}), \tag{21}$$

protože složky náhodného výběru jsou nezávislé náhodné veličiny.

Hustotu náhodného výběru $\boldsymbol{Y}$ uvažovanou při zvoleném $\boldsymbol{Y} = \boldsymbol{y}$ jako funkci parametru $\boldsymbol{\theta}$, nazýváme *funkcí věrohodnosti* a značíme ji symbolem $L(\boldsymbol{\theta})$, tedy

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i; \boldsymbol{\theta}). \tag{22}$$

Definice 2.4 Vektor statistik (výběrových funkcí) $\widehat{\boldsymbol{\theta}}(\boldsymbol{Y}) = (\hat{\theta}_1(\boldsymbol{Y}), \ldots, \hat{\theta}_k(\boldsymbol{Y}))$, který pro $\boldsymbol{Y} = \boldsymbol{y}$ splňuje vztah

$$L(\widehat{\boldsymbol{\theta}}(\boldsymbol{y})) \geq L(\boldsymbol{\theta}), \quad \forall \boldsymbol{\theta} \in \Theta \tag{23}$$

nazýváme *maximálně věrohodným odhadem* vektoru parametrů $\boldsymbol{\theta}$.

Definice maximálně věrohodného odhadu byla založena na následující úvaze: Nechť máme dva různé body $\boldsymbol{\theta}_1$ a $\boldsymbol{\theta}_2$ z parametrického prostoru Θ. Je-li $f(\boldsymbol{y}, \boldsymbol{\theta}_1)$ o mnoho menší než $f(\boldsymbol{y}, \boldsymbol{\theta}_2)$, znamená to, že výsledek pozorování $\boldsymbol{Y} = \boldsymbol{y}$ má při hodnotě parametru $\boldsymbol{\theta} = \boldsymbol{\theta}_1$ o mnoho menší pravděpodobnost než při hodnotě $\boldsymbol{\theta} = \boldsymbol{\theta}_2$. Jsme tudíž ochotni považovat za správnou hodnotu parametru $\boldsymbol{\theta}$ spíše $\boldsymbol{\theta}_2$ než $\boldsymbol{\theta}_1$. Proto je maximálně věrohodným odhadem parametru $\boldsymbol{\theta}$ takový vektor statistik $\widehat{\boldsymbol{\theta}}(\boldsymbol{Y}) = (\hat{\theta}_1(\boldsymbol{Y}), \ldots, \hat{\theta}_k(\boldsymbol{Y}))$, který maximalizuje pro daný výběr věrohodnostní funkci $L(\boldsymbol{\theta})$ na množině Θ možných hodnot parametru $\boldsymbol{\theta}$.

V praxi se ukázalo, že je vhodné místo maximalizace funkce $L(\boldsymbol{\theta})$ maximalizovat její logaritmus $\ln L(\boldsymbol{\theta})$, který se nazývá *logaritmická funkce věrohodnosti*.

Důležitou podmínkou pro dobré statistické vlastnosti maximálně věrohodného odhadu je tzv. regularita (podrobněji viz v [4]). Požaduje se zde zejména nezávislost množiny $\{\boldsymbol{y} : f(\boldsymbol{y}; \boldsymbol{\theta}) > 0\}$ na parametru $\boldsymbol{\theta}$ a existence spojitých parciálních derivací funkce $L(\boldsymbol{\theta})$ podle všech složek $\boldsymbol{\theta}$ při každém $\boldsymbol{y}$. Máme-li podmínky regularity splněny, pak je maximálně věrohodný odhad $\boldsymbol{\theta}$ dán řešením soustavy rovnic

$$\frac{\partial L(\boldsymbol{\theta})}{\partial \theta_j} = 0, \quad j = 1, \ldots, k, \quad \text{resp.} \quad \frac{\partial \ln L(\boldsymbol{\theta})}{\partial \theta_j} = 0, \quad j = 1, \ldots, k. \tag{24}$$

Tyto rovnice se nazývají *věrohodnostní rovnice*. Jestliže bychom chtěli odhadnout funkci $\tau(\boldsymbol{\theta})$ vektoru paramentrů θ, můžeme klást při použití metody maximální

věrohodnosti

$$\widehat{\tau(\boldsymbol{\theta})} = \tau(\hat{\theta}_1, \ldots, \hat{\theta}_k). \tag{25}$$

Například pro odhad podílu rozptylů $\frac{\sigma_1^2}{\sigma_2^2}$ dvou složek smíšené distribuce lze použít získané hodnoty σ_1^2 a σ_2^2 a jednoduše je vydělit.

Nyní si ukážeme, jak bychom metodou maximální věrohodnosti postupovali při určování parametrů smíšené distribuce složené ze dvou normálních rozdělení.

Příklad 2.1 Pokusme se pomocí metody maximální věrohodnosti najít odhady parametrů smíšené distribuce ve tvaru

$$f(x) = \pi_1 \phi(x; \mu_1, \sigma_1^2) + (1 - \pi_1)\phi(x; \mu_2, \sigma_2^2), \tag{26}$$

kde $\phi(x; \mu, \sigma^2)$ je hustota normálního rozdělení a $\pi_1 \in (0, 1)$.

Na první pohled vidíme, že věrohodnostní funkce bude mít 5 neznámých parametrů a můžeme ji zapsat následovně

$$L(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \pi_1) = \prod_{i=1}^{n} \left(\pi_1 \phi(x_i; \mu_1, \sigma_1^2) + (1 - \pi_1)\phi(x_i; \mu_2, \sigma_2^2) \right). \tag{27}$$

Při zápisu s použitím explicitního vyjádření hustoty ϕ dostaneme tento tvar

$$L(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \pi_1) = \prod_{i=1}^{n} \left(\pi_1 \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left[-\frac{(x_i - \mu_1)^2}{2\sigma_1^2} \right] + \right.$$

$$\left. + (1 - \pi_1) \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left[-\frac{(x_i - \mu_2)^2}{2\sigma_2^2} \right] \right) =$$

$$= \frac{1}{(\sqrt{2\pi})^n} \prod_{i=1}^{n} \left(\pi_1 \frac{1}{\sigma_1} \exp\left[-\frac{(x_i - \mu_1)^2}{2\sigma_1^2} \right] + (1 - \pi_1) \frac{1}{\sigma_2} \exp\left[-\frac{(x_i - \mu_2)^2}{2\sigma_2^2} \right] \right) \tag{28}$$

Nyní celou rovnici zlogaritmujeme. Dostáváme tedy

$$\ln\left(L(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \pi_1) \right) = -\frac{n}{2} \ln(2\pi) +$$

$$+ \sum_{i=1}^{n} \ln \left(\frac{\pi_1}{\sigma_1} \exp\left[-\frac{(x_i - \mu_1)^2}{2\sigma_1^2} \right] + \frac{1 - \pi_1}{\sigma_2} \exp\left[-\frac{(x_i - \mu_2)^2}{2\sigma_2^2} \right] \right) \tag{29}$$

Spočítáme-li parciální derivace podle všech pěti parametrů a položíme je rovny nule, získáme soustavu věrohodnostních rovnic, jejímž řešením bude maximálně věrohodný odhad vektoru parametrů $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \pi_1)$. Pro zjednodušení zápisu si označme argument logaritmu v rovnici (29) symbolem $\odot$ a místo označení $L(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \pi_1)$ budeme používat pouze L. Pak mají parciální derivace následující podobu

$$\frac{\partial \ln L}{\partial \mu_1} = \sum_{i=1}^{n} (\odot)^{-1} \frac{\pi_1 (x_i - \mu_1)}{\sigma_1^3} \exp\left[-\frac{(x_i - \mu_1)^2}{2\sigma_1^2} \right] \tag{30}$$

$$\frac{\partial \ln L}{\partial \mu_2} = \sum_{i=1}^{n} (\odot)^{-1} \frac{(1 - \pi_1)(x_i - \mu_2)}{\sigma_2^3} \exp\left[-\frac{(x_i - \mu_2)^2}{2\sigma_2^2} \right] \tag{31}$$

$$\frac{\partial \ln L}{\partial \sigma_1} = \frac{\pi_1}{\sigma_1^2} \sum_{i=1}^{n} (\odot)^{-1} \exp\left[-\frac{(x_i - \mu_1)^2}{2\sigma_1^2} \right] \left(\frac{1}{\sigma_1^2} (x_i - \mu_1)^2 - 1 \right) \tag{32}$$

$$\frac{\partial \ln L}{\partial \sigma_2} = \frac{1 - \pi_1}{\sigma_2^2} \sum_{i=1}^{n} (\odot)^{-1} \exp\left[-\frac{(x_i - \mu_2)^2}{2\sigma_2^2} \right] \left(\frac{1}{\sigma_2^2} (x_i - \mu_2)^2 - 1 \right) \tag{33}$$

$$\frac{\partial \ln L}{\partial \pi_1} = -\frac{n}{2\pi_1} + \sum_{i=1}^{n} (\odot)^{-1} \left(\frac{1}{\sigma_1} \exp\left[-\frac{(x_i - \mu_1)^2}{2\sigma_1^2} \right] - \frac{1}{\sigma_2} \exp\left[-\frac{(x_i - \mu_2)^2}{2\sigma_2^2} \right] \right). \tag{34}$$

Položíme-li výše uvedené parciální derivace rovny nule, nalezneme maximálně věrohodné odhady parametrů $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$ a π_1. V tomto případě to ovšem není zcela jednoduché vzhledem k velkému počtu složek a počtu dat.

2.2 Maximálně věrohodné odhady μ a Σ

Ukázali jsme si postup jak určit maximálně věrohodné odhady pro náhodnou veličinu (smíšenou distribuci z normálního rozdělení). Teď si předvedeme postup v případě náhodného vektoru z normálního rozdělení (viz [1]).

Příklad 2.2 Mějme $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ náhodný výběr z p-rozměrného normálního rozdělení s parametry $\boldsymbol{\mu}$ a Σ. Abychom mohli určit maximálně věrohodné odhady parametrů $\boldsymbol{\mu}$ a Σ, musíme opět sestavit věrohodnostní funkci, kterou budeme posléze maximalizovat při daných pevných hodnotách $\boldsymbol{y}$. Věrohodnostní funkce bude ve tvaru

$$L(\boldsymbol{\mu}, \Sigma) = \prod_{i=1}^{n} (2\pi)^{-\frac{p}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} (\boldsymbol{y}_i - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}) \right\} \tag{35}$$

Teď je nutno věrohodnostní funkci vhodně upravit.

$$L(\boldsymbol{\mu}, \Sigma) = (2\pi)^{-\frac{np}{2}} |\Sigma|^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}) \right\} =$$

$$= (2\pi)^{-\frac{np}{2}} |\Sigma|^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2} \mathrm{Tr} \left[\Sigma^{-1} \sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{\mu}) (\boldsymbol{y}_i - \boldsymbol{\mu})^T \right] \right\}.$$

Jako v předchozím využijeme výhod počítání s přirozeným logaritmem, tudíž

$$\ln L(\boldsymbol{\mu}, \Sigma) = -\frac{np}{2} \ln 2\pi - \frac{n}{2} \ln |\Sigma| - \frac{1}{2} \sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}) =$$

$$= \frac{np}{2} \ln 2\pi - \frac{n}{2} \ln |\Sigma| - \frac{1}{2} \mathrm{Tr} \left[\Sigma^{-1} \sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{\mu}) (\boldsymbol{y}_i - \boldsymbol{\mu})^T \right].$$

Nyní hledáme maximum $\ln L(\boldsymbol{\mu}, \Sigma)$. Určíme parciální derivace podle $\boldsymbol{\mu}$ a Σ.

$$\frac{\partial}{\partial \boldsymbol{\mu}} \ln L(\boldsymbol{\mu}, \Sigma) = -\frac{1}{2} \sum_{i=1}^{n} 2\Sigma^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}) \stackrel{!}{=} 0. \quad / \Sigma$$

Dostáváme výsledný odhad vektoru středních hodnot. Odhadem je výběrový průměr (vztah (15)).

$$\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{y}_i = \overline{\boldsymbol{y}}. \tag{36}$$

Obdobně budeme postupovat pro parametr Σ.

$$\frac{\partial}{\partial \Sigma} \ln L(\boldsymbol{\mu}, \Sigma) = \frac{n}{2} \frac{1}{|\Sigma^{-1}|} \left| \Sigma^{-1} \right| - \frac{1}{2} \sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{\mu}) (\boldsymbol{y}_i - \boldsymbol{\mu})^T \stackrel{!}{=} 0.$$

Výsledný odhad varianční matice je ve tvaru

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{y}_i - \overline{\boldsymbol{y}})(\boldsymbol{y}_i - \overline{\boldsymbol{y}})^T = \frac{1}{n}W = S'. \tag{37}$$

Teď bychom měli ještě ukázat, že se opravdu jedná o maximum. Důkaz je uveden v [1]. Je vhodné podotknout, že jsme určili maximálně věrohodné odhady v případě p-rozměrného normálního rozdělení. Kdybychom chtěli určit odhady pro několikasložkovou smíšenou distribuci z p-rozměrného normálního rozdělení, byli by výpočet i věrohodnostní rovnice velmi komplikované. Proto byl vyvinut postup, který si předvedeme v následující části.

2.3 EM (Expectation-Maximization) algoritmus

2.3.1 Algoritmus pro jednorozměrná data

V praxi velmi používanou metodou odhadování parametrů smíšených distribucí je tzv. *EM algoritmus* (viz [6, 7]). Je to iterační algoritmus k nalezení maximálně věrohodných odhadů. Tato metoda je užitečná hlavně tehdy, když jednodušší metody nelze efektivně použít. Stala se standardním nástrojem statistiků a je používána v mnoha aplikacích.

Potřebujeme odhadnout parametry $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_{c-1}, \mu_1, \ldots, \mu_c, \sigma_1^2, \ldots, \sigma_c^2)$. Proto maximalizujeme logaritmickou věrohodnostní funkci danou vztahem

$$L(\boldsymbol{\theta}|y_1, \ldots, y_n) = \sum_{i=1}^{n} \ln\left[\sum_{k=1}^{c} \pi_k \phi(y_i; \mu_k, \sigma_k^2)\right]. \tag{38}$$

Předpokládáme, že složky distribuce existují v pevných proporcích a jsou dány π_k. Proto má smysl počítat pravděpodobnost, že hodnota y_i patří do jedné ze složek distribuce. Jelikož pravděpodobonost příslušnosti hodnoty y_i do některé ze složek je neznámá, potřebujeme použít například EM algoritmus k maximalizaci rovnice (38). Pravděpodobnost, že hodnota pozorování y_i patří do k-té složky distribuce (aposteriorní pravděpodobnost), můžeme zapsat následovně

$$\hat{\tau}_{ik}(y_i) = \frac{\hat{\pi}_k \phi(y_i; \hat{\mu}_k, \hat{\sigma}_k^2)}{\hat{f}(y_i; \hat{\pi}_k, \hat{\mu}_k, \hat{\sigma}_k^2)}, \quad k = 1, \ldots, c;\ i = 1, \ldots, n, \tag{39}$$

kde

$$\hat{f}(y_i; \hat{\pi}_k, \hat{\mu}_k, \hat{\sigma}_k^2) = \sum_{k=1}^{c} \hat{\pi}_k \phi(y_i; \hat{\mu}_k, \hat{\sigma}_k^2). \tag{40}$$

Vidíme, že aposteriorní pravděpodobnost je odhadována na základě jiných odhadů. K získání maxima funkce (38) musíme najít první parciální derivace podle všech parametrů a položit je rovny nule. Tím dostaneme věrohodnostní rovnice. Jejich řešením jsou

$$\hat{\pi}_k = \frac{1}{n} \sum_{i=1}^{n} \hat{\tau}_{ik}, \tag{41}$$

$$\hat{\mu}_k = \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{\tau}_{ik} y_i}{\hat{\pi}_k} \tag{42}$$

$$\hat{\sigma}_k^2 = \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{\tau}_{ik} (y_i - \hat{\mu}_k)^2}{\hat{\pi}_k}. \tag{43}$$

EM algoritmus je dvoukrokový proces skládající se z E-kroku a M-kroku. Tyto kroky se opakují dokud odhadované hodnoty parametrů nekonvergují. Podívejme se blíže co jednotlivé kroky obnášejí.

E-krok: Spočítáme aposteriorní pravděpodobnost, že i-té pozorování náleží do k - té složky, která je dána aktuálními parametry. Použijeme rovnici (39).

M-krok: Aktualizujeme odhady parametrů použitím aposteriorní pravděpodobnosti a rovnic (41) až (43).

Celý EM algoritmus pro odhad parametrů smíšené distribuce z normálních rozdělení můžeme shrnout do následujících 5 kroků.

EM procedura

1. Určete počet složek distribuce (c).

2. Určete počáteční odhad parametrů složek smíšené distribuce (tj. určit odhady vah složek, středních hodnot a rozptylů).

3. Pro každý bod y_i určete aposteriorní pravděpodobnost použitím rovnice (39) a aktuálních hodnot parametrů. Provádíme E-krok.

4. Aktualizujte váhy složek, střední hodnoty a rozptyly použitím rovnic (41) až (43). Provádíme M-krok.

5. Opakujte kroky 3 a 4 dokud odhady konvergují.

Ve kroku 5 prakticky pokračujeme iteracemi dokud se změna hodnot odhadů ve dvou po sobě jdoucích iteracích neblíží určené toleranci. Je dobré si uvědomit, že při použití EM algoritmu používáme celý vzorek, což zejména pro velké vzorky klade větší nároky na výpočty.

Jak lze z dat určit počáteční hodnoty parametrů π_i, μ_i a σ_i? Asi nejjednodušším způsobem je setřídit data vzestupně podle naměřených hodnot. Potom v závislosti na předpokládaném počtu složek (c) soubor dat rozdělíme na poloviny, třetiny atd. Prvotním odhadem vah bude $\pi_i = \frac{1}{c}$. Střední hodnoty a rozptyly odhadneme výběrovým průměrem (13) a empirickým rozptylem (14) spočítanými pro každou c-tinu souboru.

2.3.2 Algoritmus pro vícerozměrná data

V případě vícerozměrných dat se v odhadovaných parametrech změní rozptyly na varianční matice a pochopitelně se také změní rozměr vektoru středních hodnot. Odhadujeme tedy $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_{c-1}, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_c, \Sigma_1, \ldots, \Sigma_c)$. Budeme maximalizovat logaritmickou věrohodnostní funkci

$$L(\boldsymbol{\theta}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) = \sum_{i=1}^{n} \ln \left[\sum_{k=1}^{c} \pi_k \phi(\boldsymbol{y}_i; \boldsymbol{\mu}_k, \Sigma_k) \right]. \tag{44}$$

Řešení věrohodnostních rovnic bude analogické. Dostaneme jej ve tvaru

$$\hat{\pi}_k = \frac{1}{n} \sum_{i=1}^{n} \hat{\tau}_{ik}, \tag{45}$$

$$\hat{\boldsymbol{\mu}}_k = \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{\tau}_{ik} \boldsymbol{y}_i}{\hat{\pi}_k} \tag{46}$$

$$\hat{\Sigma}_k = \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{\tau}_{ik}(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_k)(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_k)^T}{\hat{\pi}_k}. \tag{47}$$

EM procedura bude mít stejný průběh jako v předchozí části. Větší problém vyvstane, když budeme chtít nějakým způsobem určit prvotní odhady všech parametrů. Postup jak určit prvotní odhady středních hodnot a variančních matic v případě vícerozměrných dat je uveden dále v kapitole 4. Jakmile získáme prvotní odhady všech potřebných parametrů, můžeme použít EM algoritmus. V další kapitole si ukážeme, jakým způsobem lze metodu maximální věrohodnosti resp. EM proceduru použít při řešení praktické úlohy.

3 Příklad jednorozměrný

3.1 Motivace

Budeme aplikovat problematiku smíšených distribucí na studii, která byla zadána Laboratoří experimentální medicíny při Dětské klinice LF UP a FN v Olomouci. Studie zkoumá význam proteinů mnohočetné lékové rezistence u léčby astma bronchiale a nespecifických střevních zánětů. Bližší vysvětlení studie poskytl MUDr. Petr Džubák (viz následující odstavce).

Jedním ze zkoumaných proteinů je P-glykoprotein (P-gp). Je to transmembránová pumpa, která transportuje řadu lipofilních látek včetně glukokortikosteroidů. Proto se soudí, že zvýšené množství P-glykoproteinu a dalších transportních proteinů je důvodem rezistence na léčbu u nespecifických střevních zánětů, astma bronchiale a samozřejmě u nádorových onemocnění. Současně bylo popsáno množství genetických variací genu pro tento protein, přičemž některé z nich měly vztah ke stabilitě a funkci proteinu. Data, která v minulých analýzách hodnotila pouhé množství P-gp byla značně nehomogenní a mohla být ovlivněna vlastní metodikou stanovení. Proto, abychom postihli i tyto odchylky, plánujeme provést mnohem komplexnější analýzu. Jako první bude vyšetřeno množství proteinů lékové rezistence P-gp na lymfocytech periferní krve a to za použití imunofluorescenčních technik průtokové cytometrie. Množství proteinů bude srovnáváno s klinickým stavem pacientů, genetickým profilem a expresí genů[1] na úrovni mRNA.

Výsledky, které jsme získali v pilotních experimentech, ukazují, že množství P-glykoproteinu se na lymfocytech periferní krve liší a stratifikuje je do několika skupin s nízkým, středním a vysokým množstvím P-gp. Přitom každý pacient je odlišný, jak množstvím P-gp, tak množstvím skupin. Software, který běžně používáme pro hodnocení flow cytometrických dat neumožňuje mnohorozměrnou analýzu křivek za použití matematického modelování. Proto je pro nás zajímavé

[1]Exprese genu (také genová exprese) je proces, kterým je v genu uložená informace převedena v reálně existující buněčnou strukturu nebo funkci. Tomu reálně odpovídá několikakroková syntéza proteinu, který tomuto genu (tedy sekvenci jeho DNA) odpovídá (je jím kódován) a kterým (a nebo skrze něj) je později daná funkce realizována (viz [3]).

vytvoření matematického nástroje, který bude umět proložit těmito histogramy křivky a případně odlišit jednotlivé skupiny a to i na základě jiných parametrů, jako je velikost a granularita. Výstupem by měly být obecně používané statistické veličiny, popisující jednotlivé vzorky (medián, průměr, variační koeficient a další). V ideálním případě by bylo vhodné propojení této analýzy do komplexu mnohorozměrných analýz hodnotících vztahy mezi klinickým stavem pacienta, jeho antropometrickými parametry a parametry molekulární biologie a genetiky. Tato studie si klade za cíl zjistit vztah mezi odpovědí na terapii a proteiny mnohočetné lékové rezistence, přičemž výstupem by v dlouhodobém horizontu mohla být optimalizace a individualizace cílené terapie výše zmíněných pacientů.

3.2 Postup

Naším úkolem bude zjistit počet skupin a rozdělení P-gp do těchto skupin pro každého pacienta.

Příklad budeme řešit pomocí EM algoritmu. K dispozici máme vzorky od 8 pacientů. Každý vzorek obsahuje několik tisíc měření, které budou tvořit náš náhodný výběr. U každého pacienta nejprve provedeme úvahu o počtu složek distribuce. Ten lze zjistit například vytvoříme-li histogram z naměřených dat. Podle tvaru histogramu usoudíme na počet složek. Naměřené hodnoty pak vzestupně setřídíme a rozdělíme je na tolik částí, kolik je počet složek. Pro každou část takto rozděleného souboru spočítáme výběrový průměr a výběrový rozptyl. Tímto získáme včechny prvotní odhady vah (1/počet složek), středních hodnot a rozptylů složek smíšené distribuce. Pak můžeme použít EM algoritmus.

Při počítání v Matlabu použijeme proceduru *csfinmix*, která nám pro zadané vstupní parametry spočítá odhady parametrů $\pi_1, \ldots, \pi_c; \mu_1, \ldots, \mu_c$ a $\sigma_1^2, \ldots, \sigma_c^2$.

Procedura má následující syntaxi:

[wts,mus,vars] = csfinmix(DATA,$[\mu_1, \ldots, \mu_c]$,$[\sigma_1^2, \ldots, \sigma_c^2]$,$[\pi_1, \ldots, \pi_c]$,Iterace,Tol),

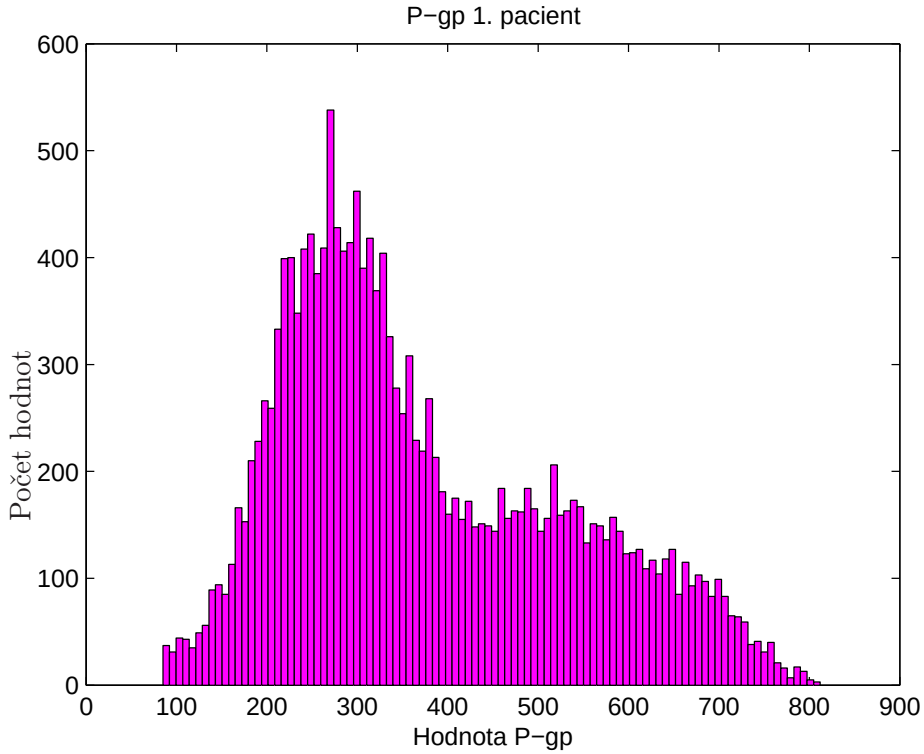
kde [wts,mus,vars] jsou získané odhadované hodnoty parametrů v pořadí: váhy

složek, střední hodnoty a rozptyly. Parametr DATA je vektor naměřených hodnot, pak následují vektory prvotních odhadů středních hodnot, rozptylů a vah složek. Parametr Iterace nám určuje maximální počet iterací, které bude algortimus počítat, a parametr Tol nám dává omezení na výpočet. Zadáme si jím hraniční hodnotu rozdílu počítaných parametrů smíšené distribuce ve dvou po sobě jdoucích iteracích. Bude-li rozdíl všech hodnot parametrů distribuce ve dvou iteracích jdoucích za sebou menší než tato hranice, výpočet se zastaví. Parametry Iterace a Tol zadáváme omezení na výpočet.

3.3 Výpočet

3.3.1 1. pacient

Nejdříve si vytvořme histogram, ze kterého pak provedeme odhad počtů složek ve smíšené distribuci.



Obrázek 13: Histogram četností hodnot P-gp v intervalech délky 10

Z histogramu vidíme, že hodnoty P-gp u 1. pacienta můžeme vysvětlit dvou, případně třísložkovou smíšenou distribucí. Zkusme tedy spočítat parametry těchto distribucí. Nejprve uvažujme případ dvousložkové smíšené distribuce. Z dat spočítáme prvotní odhady parametrů. Dostaneme tedy $\pi_1 = 0,5; \pi_2 = 0,5; \mu_1 = 248,2608; \mu_2 = 500,2781; \sigma_1^2 = 2863,8489$ a $\sigma_2^2 = 14135,4044$. Na základě těchto hodnot spočítáme pomocí EM algoritmu přesnější odhady parametrů této distribuce. Ukažme si nyní blíže, jak budou vypadat jednotlivé iterace a jaké budou změny hodnot parametrů (viz tabulka 2). Zdrojový kód k histogramu nalezneme v příloze A5.

Poznámka: Počáteční hodnoty rozptylů σ_1^2 a σ_2^2 jsou počítány pomocí funkce VAR v programu MS Excel. Ta ovšem používá ve jmenovateli pro výpočet výběrového rozptylu n namísto $n-1$. Tento rozdíl se však vzhledem k velkému počtu dat projeví minimálně a iterační postup jej prakticky vymaže.

Vidíme, že hodnoty parametrů se ustalují při toleranci 0,0001 až po dvousté iteraci (viz tabulka 2). V našem případě ovšem tolerance 0,0001 představuje velmi přísný požadavek, neboť zadaná data jsou celočíselného charakteru. Spočítejme tedy stejným způsobem odhady parametrů, uvažujeme-li třísložkovou smíšenou distribuci a získané hodnoty porovnejme. Pro zajímavost si ještě spočítejme parametry, uvažujeme-li obyčejné normální rozdělení. Získané výsledky jsou v následující tabulce 3.

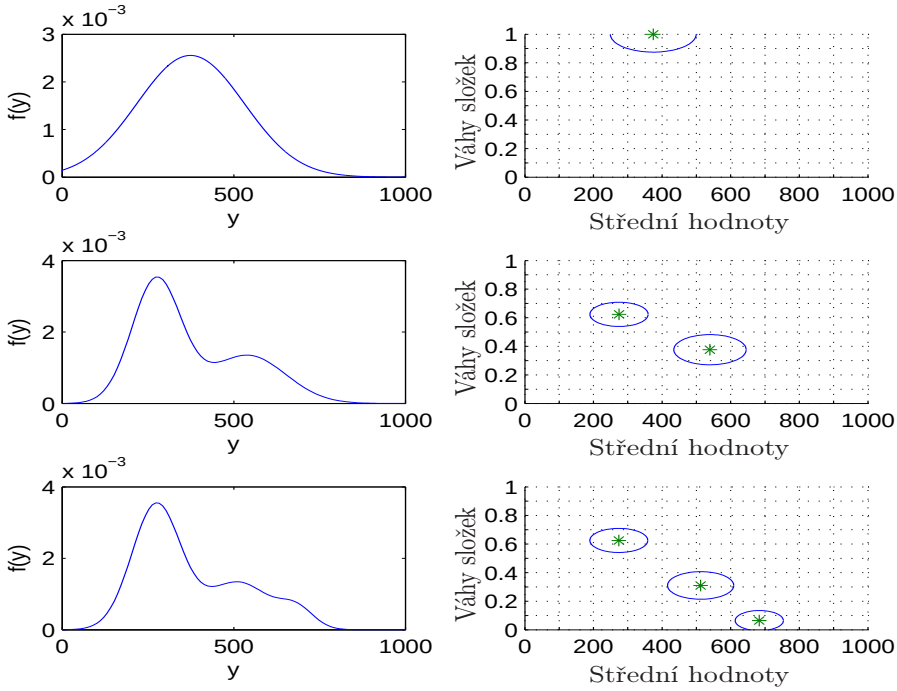
Tabulka 2: Průběh iterací

iterace	π_1	π_2	μ_1	μ_2	σ_1^2	σ_2^2
0	0,5	0,5	248,2608	500,2781	2863	14135
1	0,5033	0,4967	254,9605	495,1409	3548	16450
2	0,5100	0,4900	258,9748	494,2660	3871	17487
3	0,5163	0,4837	261,3258	494,8277	4056	17919
4	0,5219	0,4781	262,7587	496,0202	4174	18037
5	0,5271	0,4729	263,6886	497,5100	4255	17988
6	0,5318	0,4682	264,3412	499,1346	4315	17849
7	0,5363	0,4637	264,8401	500,8081	4362	17665
8	0,5405	0,4595	265,2531	502,4841	4400	17457
9	0,5445	0,4555	265,6178	504,1370	4434	17240
10	0,5483	0,4517	265,9547	505,7528	4465	17021
11	0,5520	0,4480	266,2750	507,3240	4493	16804
12	0,5555	0,4445	266,5847	508,8466	4520	16592
13	0,5589	0,4411	266,8869	510,3187	4545	16386
14	0,5621	0,4379	267,1830	511,7396	4570	16186
15	0,5652	0,4348	267,4736	513,1092	4594	15994
16	0,5682	0,4318	267,7587	514,4276	4617	15808
17	0,5711	0,4289	268,0384	515,6956	4640	15630
18	0,5738	0,4262	268,3123	516,9135	4661	15459
19	0,5764	0,4236	268,5800	518,0824	4683	15296
20	0,5789	0,4211	268,8414	519,2029	4703	15139
21	0,5813	0,4187	269,0960	520,2761	4723	14989
22	0,5836	0,4164	269,3438	521,3029	4743	14846
23	0,5857	0,4143	269,5843	522,2844	4762	14710
24	0,5878	0,4122	269,8174	523,2217	4780	14580
25	0,5898	0,4102	270,0431	524,1161	4798	14456
...	...	...	...	...	...	...
...	...	...	...	...	...	...
100	0,6234	0,3766	274,4408	539,4936	5158	12394
101	0,6234	0,3766	274,4435	539,5022	5159	12393
...	...	...	...	...	...	...
...	...	...	...	...	...	...
205	0,6237	0,3763	274,4845	539,6302	5162	12377
206	0,6237	0,3763	274,4845	539,6303	5162	12377

Tabulka 3: Odhady parametrů smíšených distribucí 1. pacienta

Parametry			
c	π	μ	σ^2
1	1	374,2695	24377,82
2	(0,6237; 0,3763)	(274,4845; 539,6303)	(5162; 12377)
3	(0,6253; 0,3103; 0,0644)	(273,911; 512,405; 683,371)	(5064; 8661; 2428)

Pomocí výše uvedených dat si sestrojme grafy a dF ploty příslušných hustot (viz obrázek 14, kód analogicky podle A1, A2) a porovnejme je. Při vizualizaci dF plotů v Matlabu byly při výpočtu směrodatné odchylky stokrát zvětšeny pro zachování názornosti.

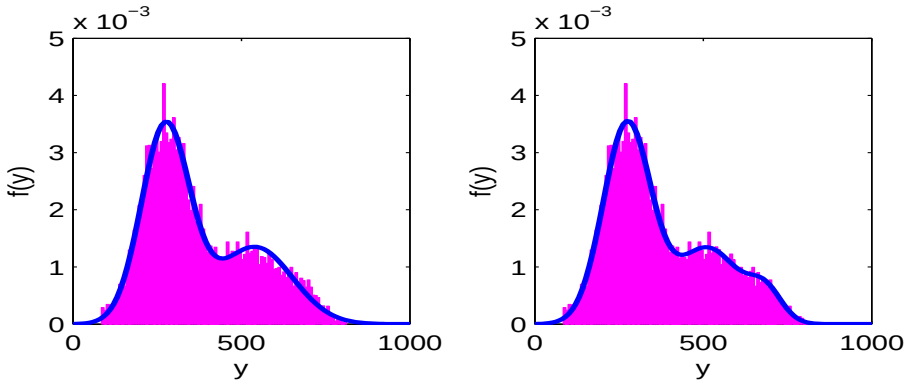


Obrázek 14: Srovnání grafů hustot a dF plotů pro P-gp 1. pacienta pro jedno, dvou a třísložkovou smíšenou distribuci. Směrodatné odchylky u dF plotů jsou 100 krát zvětšeny

Z tabulky 3 a obrázku 14 vidíme, že při přechodu ze dvousložkové na třísložkovou smíšenou distribuci se 1. složka prakticky nezměnila, zatímco 2. složka se rozdělila ve dvě.

Otázkou teď je, která z distribucí nejlépe odpovídá skutečnosti. Z jistotou lze říci, že to určitě nebude obyčejné normální rozdělení. Zkusme tedy ještě srovnat grafy hustot s histogramem (viz obrázek 15, kód příloha A6) a podívejme se na výsledek.

Vidíme, že spíše bychom byli pro použití distribuce se třemi složkami, která



Obrázek 15: Srovnání histogramu s hustotami dvou a třísložkové smíšené distribuce pro P-gp 1. pacienta

zejména pro vyšší hodnoty P-gp lépe odpovídá histogramu. Pro přesnější rozhodnutí je možno využít shlukové analýzy a vhodných statistických testů (viz [8]), což ovšem překračuje rámec této práce.

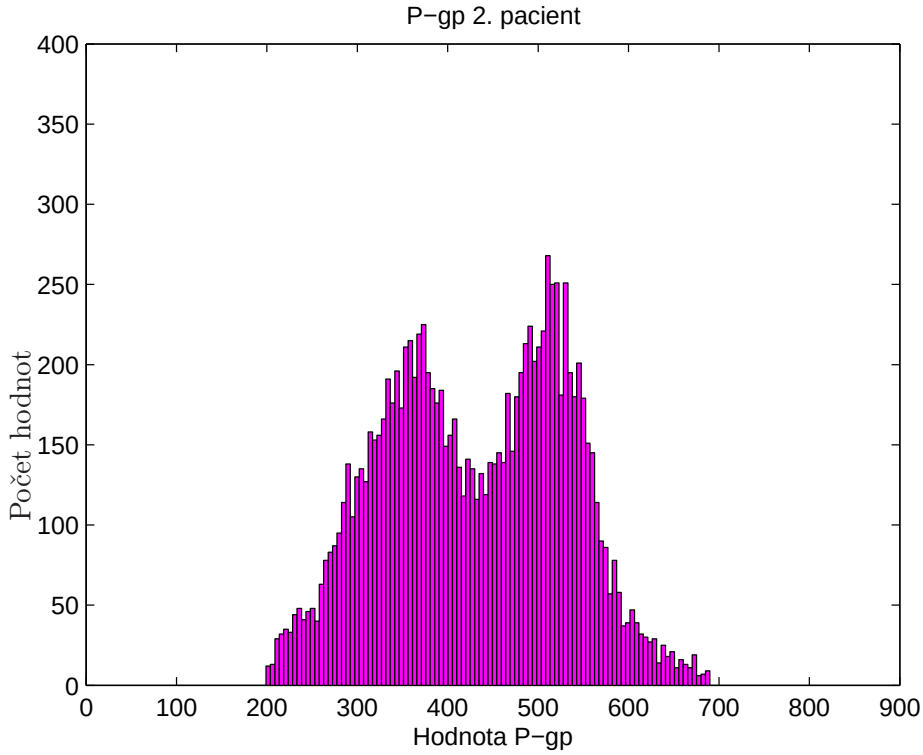
3.3.2 2. pacient

U 2. pacienta můžeme na základě histogramu četností P-gp (viz obrázek 16, kód analogicky podle A5) předpokládat 2 složky s přibližně stejnou vahou. Spočítejme opět střední hodnoty, rozptyly a váhy složek, uvažujeme-li 2 složky a porovnejme je s hodnotami, uvažujeme-li 3 složky. Spočtené hodnoty jsou uvedeny v tabulce 4.

Tabulka 4: Odhady parametrů smíšených distribucí 2. pacienta

Parametry			
c	π	μ	σ^2
2	(0,4919; 0,5081)	(346,275; 521,134)	(3430; 3188)
3	(0,5533; 0,026; 0,4208)	(356,119; 634,438; 515,83)	(4091; 802; 1743)

Vidíme, že v případě třísložkové smíšené distribuce má 2. složka pouze 2,6% váhu. Proto bychom ji mohli i vypustit. Porovnáme-li střední hodnoty složek jednotlivých distribucí, tak se příliš neliší. Podívejme se proto jak vypadá graf, složíme-li dohromady histogramy četností a grafy hustot dvou a třísložkovkých distribucí P-gp 2. pacienta (viz obrázek 17, kód analogicky podle A6).

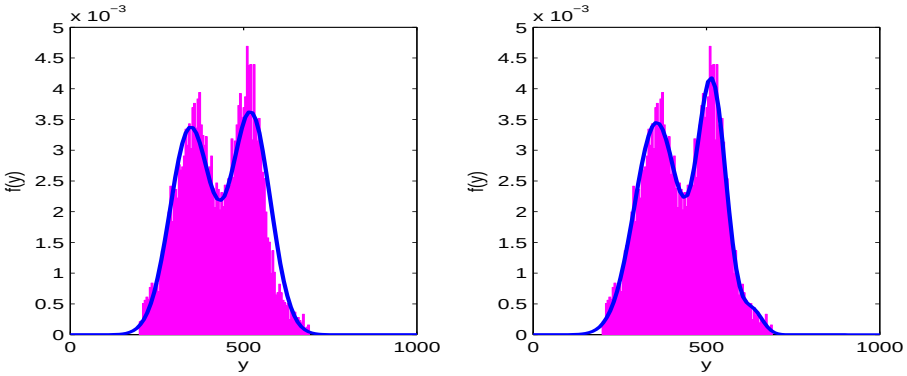


Obrázek 16: Histogram četností hodnot P-gp v intervalech délky 10

Na první pohled se grafy příliš neliší, ale je patrné, že přítomnost 3. složky lépe koresponduje s histogramem pro vyšší hodnoty P-gp. Všimněme si, že zatímco u dvousložkové distribuce jsou si váhy složek téměř rovny, což vizuálně poznáme podle velikosti „kopců“, tak u třísložkové distribuce má větší váhu složka se střední hodnotou 515,8. I pro nižší hodnoty P-gp se jeví model třísložkové smíšené distribuce nepatrně lépe.

Matematicky bychom se asi opět přiklonili k variantě se třemi složkami, ale záleží na ošetřujícím lékaři, který model použije, zvláště pak, vezmeme-li v uváhu nízkou váhu 2. složky v třísložkové smíšené distribuci, kvůli které by z lékařského hlediska mohla být tato složka vypuštěna.

Podívejme se ještě na průběh P-gp u dalšího pacienta. Vyberme si pacienta, jehož histogram četností P-gp se poměrně značně odlišuje od ostatních.



Obrázek 17: Srovnání histogramu s hustotami dvou a třísložkové smíšené distribuce pro P-gp 2. pacienta

Poznámka: Ve studii se jedná o 4. pacienta. V našem příkladu to však bude 3. a poslední pacient.

3.3.3 3. pacient

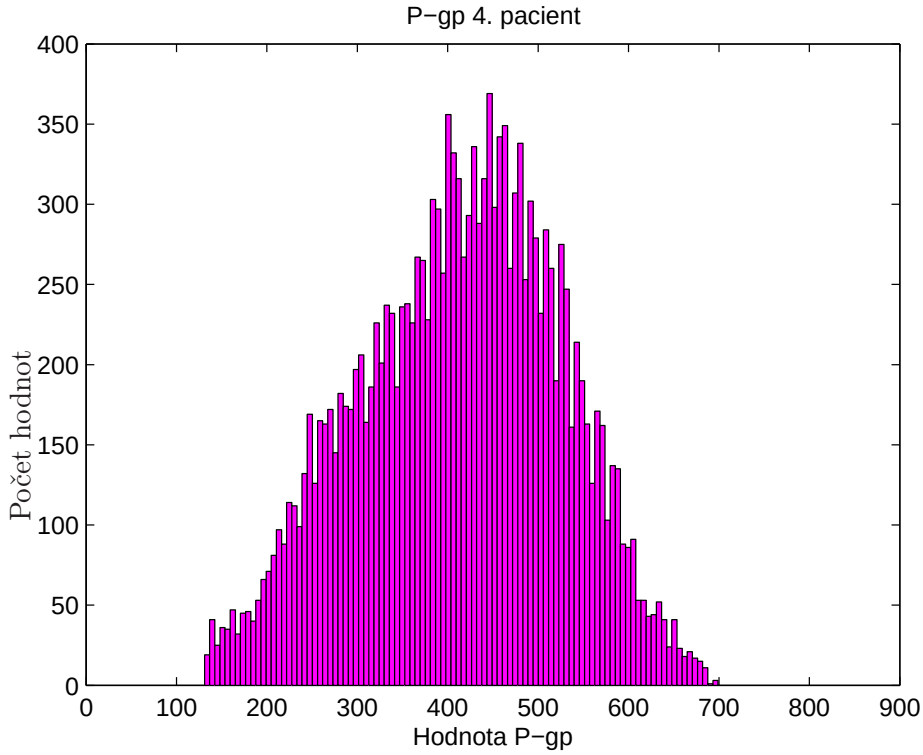
Podívejme se opět na histogram četností P-gp (obrázek 18, kód analogicky podle A5) a prověďme úvahu o počtu složek. Na první pohled lze očekávat „obyčejné“ normální rozdělení, případně lze použít distribuci se dvěma složkami. Hodnoty získané výpočty nám ukazuje tabulka 5.

Tabulka 5: Odhady parametrů smíšených distribucí 3. pacienta

Parametry			
c	π	μ	σ^2
1	1	412,9849	12436
2	(0,2209; 0,7791)	(272,92; 452,701)	(3995; 7688)

Srovnejme nyní grafy hustot získaných na základě parametrů z tabulky 5 s histogramem četností (viz obrázek 19, kód analogicky podle A6).

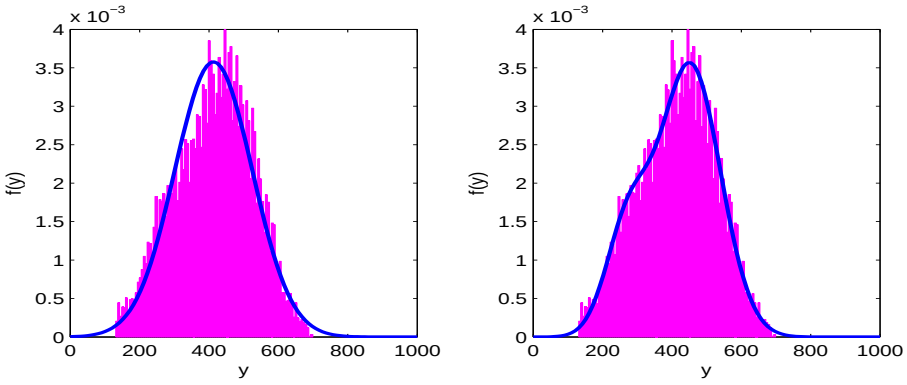
Vidíme, že i v tomto případě lépe opovídá dvousložková smíšená distribuce, nicméně i normální rozdělení nepopisuje P-gp u 3. pacienta s příliš velkými odchylkami. Jako i v předchozích případech se ukázalo, že hustoty smíšených distribucí s větším počtem složek lépe odpovídají histogramu četností P-gp.



Obrázek 18: Histogram četností hodnot P-gp v intervalech délky 10

3.4 Shrnutí

Ukázali jsme si, jak vypadají smíšené distribuce pro P-gp u jednotlivých pacintů. Je zřejmé, že pro každého pacienta mají jiný průběh. Proto je dobré léčbu odvíjet v závislosti na získaných tvarech hustot P-gp pro každého pacienta, protože při všeobecné standardizaci by léčba nemusela být efektivní. Z 8 pacientů pouze u 1 vykazovaly hodnoty P-gp možnost použití normálního rozdělení (pacient 3). U zbývajících 7 se chovalo jako dvou nebo třísložková smíšená distribuce. Je zajímavé, že u 4 pacientů měla větší váhu složka popisující nižší hodnoty P-gp (pacient 1) a u 3 měla naopak větší váhu složka popisující vyšší hodnoty P-gp (pacient 2). Lze předpokládat, že čím více složek v modelu použijeme, tím lépe odpovídá histogramu četností P-gp. Což se ukázalo u všech 3 pacientů. Vzhledem k poměrně velké variabilitě P-gp u pacientů je přínosné dále tuto problematiku



Obrázek 19: Srovnání histogramu s hustotami normálního rozdělení a dvousložkové distribuce pro P-gp 4. pacienta

zkoumat, abychom docílili co největší efektivity a specializace léčby u každého pacienta. Nyní přejděme k další části, ve které se budeme věnovat stejné problematice, ale situace se zkomplikuje tím, že budeme uvažovat dvourozměrné smíšené distribuce.

4 Příklad dvourozměrný

Ve skutečnosti se studie z předchozího příkladu nezabývala pouze jedním, ale hned třemi proteiny jimiž jsou P-gp, MRP a LRP. Proto budeme analyzovat stejná data, ale pouze ve dvojrozměrném případě. Při větším počtu složek náhodného vektoru roste totiž počet odhadovaných parametrů a to komplikuje a zpomaluje výpočty. Je totiž nutno odhadnout váhy složek $c-1$ parametrů, příslušné středních hodnoty cp parametrů a taky prvky variačních maticí $cp(p+1)/2$ parametrů, kde p je rozměr vektoru. To dává pro dvousložkové a dvourozměrné smíšené distribuce 11 parametrů. Kdybychom chtěli analyzovat všechny 3 proteiny, tak při dvousložkové distribucí dostáváme celkem 19 parametrů a navíc bychom nebyli schopni výsledné distribuce graficky znázornit.

4.1 Postup

Opět použijeme funkci csfinmix, tentokrát ale s jinými vstupními parametry.
[wts,mus,vars] = csfinmix(DATA,$[\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_c]$,$[\Sigma_1,\ldots,\Sigma_c]$,$[\pi_1,\ldots,\pi_c]$,Iterace,Tol)

V úloze bychom mohli obecně odhadovat parametry $\pi_1,\ldots,\pi_c,\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_c$ a $\Sigma_1,\ldots,\Sigma_c$. My se ale omezíme pouze na smíšené distribuce ve tvaru (11). Jelikož výše zmíněné parametry jsou vstupy funkce csfinmix, je nutné stanovit počáteční odhady těchto parametrů. Nyní si ukážeme jak.

Máme dáno opakované měření P-gp a MRP v matici DATA s rozměry $n \times 2$. V případě jedné proměnné stačilo soubor dat setřídit dle velikosti, rozdělit na části dle uvažovaného počtu složek a na základě tohoto rozdělení určit počáteční střední hodnoty a rozptyly. Zde to již nebude tak jednoduché. Je nutno dvojrozměrná data nějak transformovat, abychom provedli dělení souboru na tolik částí kolik uvažujeme složek. Provedeme proto projekci dat, jež se všechna nachází v prvním kvadrantu, na přímky dané směrovými vektory f_i. Projekci provedeme následujícím způsobem:

$$h_i = DATA * f_i,\ \ i = 1,2,3,4,5,6. \tag{48}$$

Vektory f_i zvolíme následovně:

$$f_1 = \begin{pmatrix} \cos 0 \\ \sin 0 \end{pmatrix}, f_2 = \begin{pmatrix} \cos 22.5 \\ \sin 22.5 \end{pmatrix}, f_3 = \begin{pmatrix} \cos 45 \\ \sin 45 \end{pmatrix},$$

$$f_4 = \begin{pmatrix} \cos 67.5 \\ \sin 67.5 \end{pmatrix}, f_5 = \begin{pmatrix} \cos 90 \\ \sin 90 \end{pmatrix}, f_6 = \begin{pmatrix} \cos 135 \\ \sin 135 \end{pmatrix}.$$

Touto operací získáme 6 vektorů délky n. Z nich vytvoříme histogramy (viz obrázek 20). Identifikaci 2 složek smíšené distribuce provedeme orientačně a to podle vzdálenosti dvou největších „kopců“ v histogramech. Vybereme 2 histogramy, kde jsou kopce nejvzdálenější, a na základě těchto dvou vypočteme souřadnice $\hat{\boldsymbol{\mu}}_1$ a $\hat{\boldsymbol{\mu}}_2$. Platí totiž, že

$$f_1^T \boldsymbol{\mu}_1 = a_1, \quad f_1^T \boldsymbol{\mu}_2 = a_2, \quad f_2^T \boldsymbol{\mu}_1 = b_1, \quad f_2^T \boldsymbol{\mu}_2 = b_2. \tag{49}$$

Parametry a_1, a_2, b_1 a b_2 jsou hodnoty, které odpovídají x-ovým souřadnicím vrcholů kopců, přičemž a_1 a a_2 získáme z prvního histogramu a b_1 a b_2 z druhého. Máme tedy 4 lineární rovnice se 4 neznámými $\mu_{11}, \mu_{12}, \mu_{21}$, a μ_{22}. Vyřešíme je a získáme tak prvotní odhady středních hodnot pro proceduru *csfinmix*. Teď už zbývá jen určit odhady prvků variačních matic. Váhy složek vezmeme v poměru 1:1. Pro odhad variačních matic rozdělíme soubor dat na dvě části a to pomocí přímky, která bude kolmá na spojnici bodů $\hat{\boldsymbol{\mu}}_1$ a $\hat{\boldsymbol{\mu}}_2$. Z každé části pak vypočteme odhad variační matice podle vztahu

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{y}_i - \overline{\boldsymbol{y}})(\boldsymbol{y}_i - \overline{\boldsymbol{y}})^T. \tag{50}$$

Celý postup můžeme shrnout do několika kroků.

1. Určit směrové vektory

$$f_i = \begin{pmatrix} \cos \alpha \\ \sin \alpha \end{pmatrix}, \quad \alpha \in \langle 0, 180 \rangle, \quad i = 1, 2, \ldots$$

   Volit i alespoň 5.

2. Vytvořit histogramy četností pro všechny projekce získané různou volbou směrových vektorů.

3. Vybrat ty 2 histogramy, ve kterých jsou „kopce“ nejvzdálenější.

4. Výpočet $\hat{\boldsymbol{\mu}}_1$ a $\hat{\boldsymbol{\mu}}_2$ pomocí rovnic (49).

5. Vést kolmici na spojnici bodů $\hat{\boldsymbol{\mu}}_1$ a $\hat{\boldsymbol{\mu}}_2$ v jejím středu, která rozdělí soubor dat na 2 části.

6. Z každé části vypočítat empirickou varianční matici $\hat{\Sigma}$ dle vztahu (50).

7. Pokračovat iteračním algoritmem pomocí funkce *csfinmix*.

4.2 Výpočet

4.2.1 1. pacient

Poznámka 4.1 Číslování pacientů se neshoduje s číslováním v jednorozměrném případě.

Na vstupu máme dán vektor DATA, ve kterém je zaznamenáno opakované měření P-gp a MRP. Celkem máme 17600 naměřených hodnot. Zvolíme vektory f_1 až f_6, jak bylo ukázáno. A vytvoříme histogramy (viz obrázek 20, kód v příloze A7). Podle kroku 3 vybereme první a čtvrtý histogram. Podle nich pak zvolíme čísla a_1, a_2, b_1 a b_2. Zvolme tedy $a_1 = 280, a_2 = 530, b_1 = 500$ a $b_2 = 1000$. Dosazením do rovnic (49) dostáváme prvotní odhad $\hat{\boldsymbol{\mu}}_1$ a $\hat{\boldsymbol{\mu}}_2$.

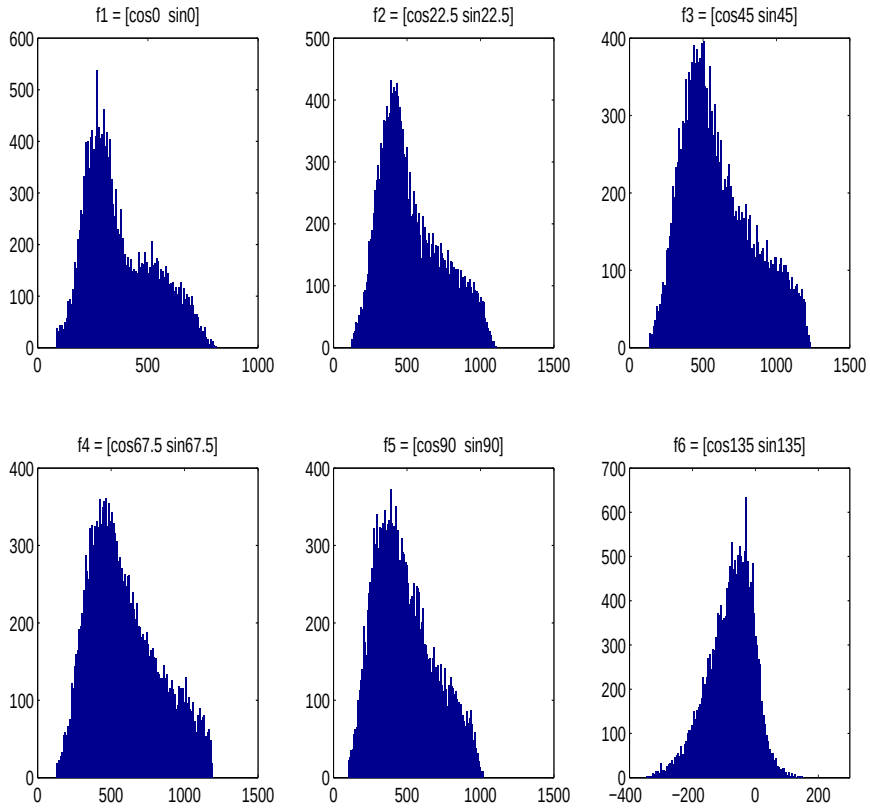
$$\hat{\boldsymbol{\mu}}_1 = \begin{pmatrix} 280 \\ 427,8 \end{pmatrix}, \hat{\boldsymbol{\mu}}_2 = \begin{pmatrix} 530 \\ 797,7 \end{pmatrix}.$$

Nyní určeme dělící přímku. Výpočtem dostáváme její analytický tvar:

$$x + 1,4796y - 1359,39 = 0. \tag{51}$$

Přímka dělí soubor 17600 měření v poměru 12922:4678 (data ležící pod:nad přímkou). Teď lze spočítat empirické varianční matice $\hat{\Sigma}_1$ a $\hat{\Sigma}_2$.

$$\hat{\Sigma}_1 = \begin{pmatrix} 8459 & 7289 \\ 7289 & 15038 \end{pmatrix}, \hat{\Sigma}_2 = \begin{pmatrix} 9194 & 5560 \\ 5560 & 12532 \end{pmatrix}.$$



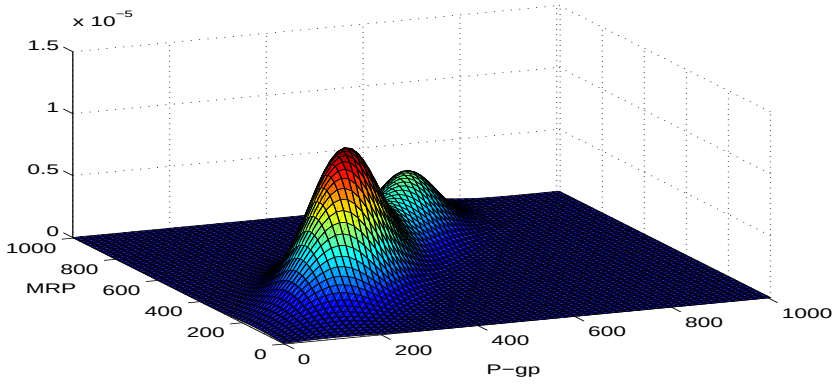
Obrázek 20: Histogramy četností po projekci na různé směry

Máme tedy všechny potřebné parametry pro použití funkce *csfinmix*. Po provedení 50 iterací dostáváme výsledné odhady parametrů smíšené distribuce:

$$\hat{\pi} = 0,7087, \quad \hat{\boldsymbol{\mu}}_1 = \begin{pmatrix} 294,0 \\ 387,2 \end{pmatrix}, \quad \hat{\boldsymbol{\mu}}_2 = \begin{pmatrix} 569,7 \\ 729,3 \end{pmatrix},$$

$$\hat{\Sigma}_1 = \begin{pmatrix} 7802 & 7137 \\ 7137 & 15880 \end{pmatrix}, \quad \hat{\Sigma}_2 = \begin{pmatrix} 10005 & 8494 \\ 8494 & 17133 \end{pmatrix}.$$

Výslednou smíšenou distribuci lze vykreslit (viz obrázek 21, kód v příloze A8). Vypočítejme ještě korelační koeficienty mezi jednotlivými proměnnými v obou složkách distribuce. Zjistíme tak jestli a jaká je závislost mezi hodnotami namě-



Obrázek 21: Graf odhadnuté smíšené distribuce pro 1. pacienta

řených proteinů. Výsledné hodnoty jsou

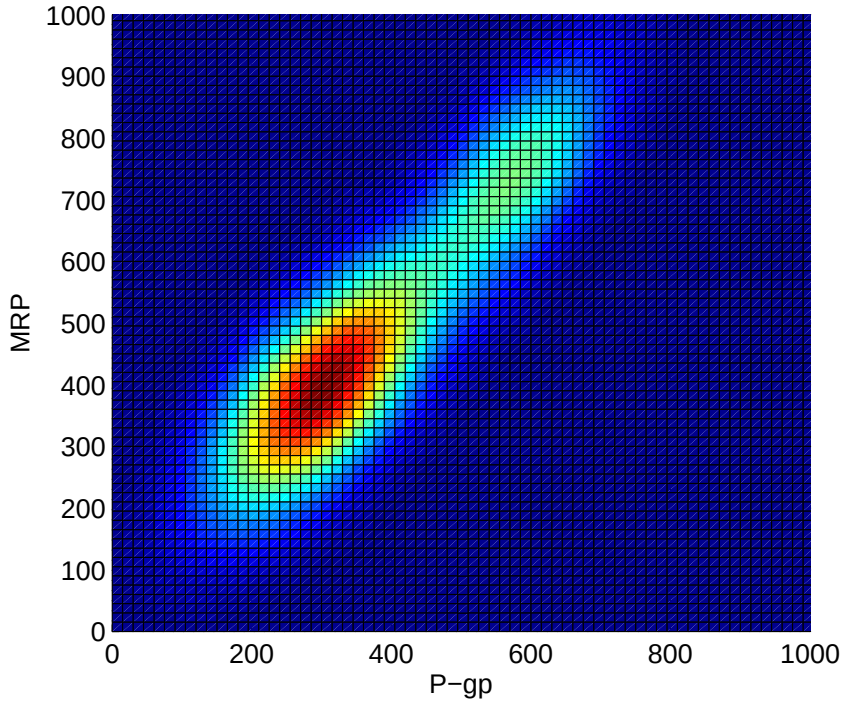
$$\rho^1_{(P-gp,MRP)} = 0,6412, \quad \rho^2_{(P-gp,MRP)} = 0,6488.$$

Mezi proteiny je tedy poměrně vysoká kladná závislost, které si všimneme na obrázku 22. Tento obrázek je horním pohledem na graf odhadnuté smíšené distribuce z obrázku 21. Kladná závislost se projeví protáhlým tvarem vrstevnic smíšené distribuce ve směru osy prvního a třetího kvadrantu. Pro zajímavost si ještě vykresleme distribuční funkci odpovídající odhadnuté smíšené distribuce pro prvního pacienta (viz obrázek 23, kód v příloze A9).

Dosud jsme se zabývali pouze odhadem parametrů smíšených distribucí náhodných vektorů. Mohla a měla by nás ale také zajímat přesnost těchto odhadů. K tomu využijeme *Fisherovu informační matici*.

Nyní si ukážeme jak spočítat empirickou Fisherovu informační matici máme-li dán předpis smíšené distribuce a data. V našem případě bude Fisherova informační matice mít rozměr 11×11, protože vycházíme z předpisu smíšené distribuce (11), který má 11 parametrů. Kvůli přehlednosti provedeme malou úpravu tohoto vztahu. Místo parametru π použijeme parametr c. Předpis smíšené distribuce má tvar

$$f(\boldsymbol{y}) = c(2\pi)^{-\frac{2}{2}} |\Sigma_1|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu}_1)^T \Sigma_1^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_1)\right\} +$$



Obrázek 22: Korelace proteinů P-gp a MRP

$$+(1-c)(2\pi)^{-\frac{2}{2}}\left|\Sigma_2\right|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu}_2)^T\Sigma_2^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_2)\right\}. \tag{52}$$

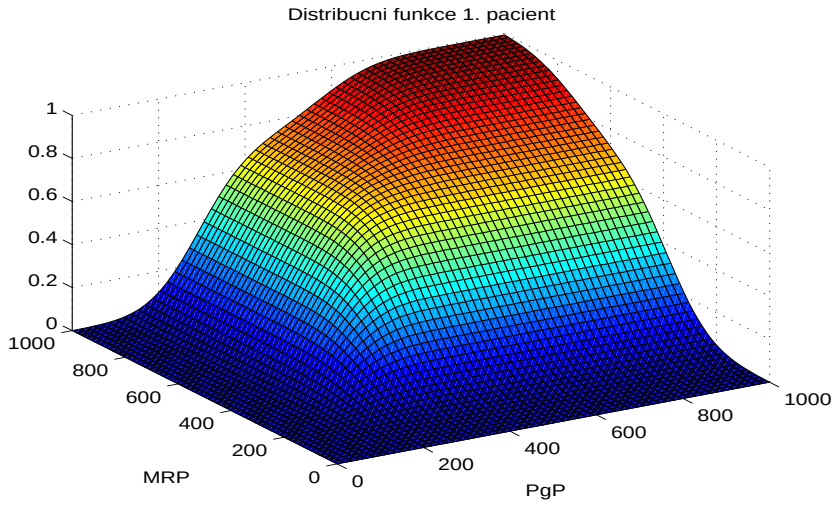
My budeme navíc používat $\ln(f(\boldsymbol{y}))$. Budeme muset spočítat parciální derivace $f(\boldsymbol{y})$ podle všech 11 parametrů.

$$\frac{\partial f(\boldsymbol{y})}{\partial c} = (2\pi)^{-\frac{2}{2}}\left|\Sigma_1\right|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu}_1)^T\Sigma_1^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_1)\right\} -$$

$$-(2\pi)^{-\frac{2}{2}}\left|\Sigma_2\right|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu}_2)^T\Sigma_2^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_2)\right\}.$$

$$\frac{\partial f(\boldsymbol{y})}{\partial \boldsymbol{\mu}_1} = \underbrace{c(2\pi)^{-\frac{2}{2}}\left|\Sigma_1\right|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu}_1)^T\Sigma_1^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_1)\right\}}_{k} *$$

$$*\frac{\partial}{\partial \boldsymbol{\mu}_1}\left[-\frac{1}{2}\left(\boldsymbol{y}^T\Sigma_1^{-1}\boldsymbol{y} - 2\boldsymbol{\mu}_1^T\Sigma_1^{-1}\boldsymbol{y} + \boldsymbol{\mu}_1^T\Sigma_1^{-1}\boldsymbol{\mu}_1\right)\right] =$$



Obrázek 23: Distribuční funkce odpovídající odhadlé smíšené distribuci prvního pacienta

$$= k\left[-\frac{1}{2}\left(-2\Sigma_1^{-1}\boldsymbol{y} + 2\Sigma_1^{-1}\boldsymbol{\mu}_1\right)\right].$$

$$\frac{\partial f(\boldsymbol{y})}{\partial \Sigma_1^{-1}} = \frac{\partial}{\partial \Sigma_1^{-1}} \underbrace{c(2\pi)^{-\frac{2}{2}}}_{h}\left[\left|\Sigma_1^{-1}\right|\right]^{\frac{1}{2}} \exp\left\{-\frac{1}{2}Tr\left[(\boldsymbol{y}-\boldsymbol{\mu}_1)(\boldsymbol{y}-\boldsymbol{\mu}_1)^T\Sigma_1^{-1}\right]\right\} =$$

$$= h\frac{1}{2}\left[\left|\Sigma_1^{-1}\right|\right]^{\frac{1}{2}}\left|\Sigma_1^{-1}\right|\Sigma_1 \ \exp\left\{-\frac{1}{2}Tr\left[(\boldsymbol{y}-\boldsymbol{\mu}_1)(\boldsymbol{y}-\boldsymbol{\mu}_1)^T\Sigma_1^{-1}\right]\right\} +$$

$$+h\left[\left|\Sigma_1^{-1}\right|\right]^{\frac{1}{2}} \exp\left\{-\frac{1}{2}Tr\left[(\boldsymbol{y}-\boldsymbol{\mu}_1)(\boldsymbol{y}-\boldsymbol{\mu}_1)^T\Sigma_1^{-1}\right]\right\}\left[-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu}_1)(\boldsymbol{y}-\boldsymbol{\mu}_1)^T\right].$$

Parciální derivace podle $\boldsymbol{\mu}_2$ a Σ_2 spočítáme analogicky tak, že jen prohodíme indexy.

Poznámka: Parciální derivací funkce $f(\boldsymbol{y})$ podle $\boldsymbol{\mu}_1$ je vektor parciálních derivací této funkce podle μ_{11} a μ_{12}. Podobně parciální derivací funkce $f(\boldsymbol{y})$ podle Σ_1 je matice složená z parciálních derivací této funkce podle jednotlivých prvků Σ_1.

$$\frac{\partial f(\boldsymbol{y})}{\partial \boldsymbol{\mu}_1} = \begin{pmatrix} \frac{\partial f(\boldsymbol{y})}{\partial \mu_{11}} \\ \frac{\partial f(\boldsymbol{y})}{\partial \mu_{12}} \end{pmatrix}, \frac{\partial f(\boldsymbol{y})}{\partial \Sigma_1} = \begin{pmatrix} \frac{\partial f(\boldsymbol{y})}{\partial \sigma_{11}} & \frac{\partial f(\boldsymbol{y})}{\partial \sigma_{12}} \\ \frac{\partial f(\boldsymbol{y})}{\partial \sigma_{21}} & \frac{\partial f(\boldsymbol{y})}{\partial \sigma_{22}} \end{pmatrix}.$$

Nyní již máme analyticky vyjádřeny všechny potřebné parciální derivace k výpočtu Fisherovy informační matice. Tu vypočítáme tak, že použijeme všechny

naměřené hodnoty a to tak, že sestavíme sloupcový vektor všech parciálních derivací, do kterých dosadíme odhadnuté parametry. Každý prvek v tomto vektoru vydělíme ještě hodnotou $f(\boldsymbol{y})$. Nyní vezmeme první měření a dosadíme jej do tohoto vektoru. Získáme tak číselný vektor. Pak vezmeme druhé měření a opět dosadíme. Tímto způsobem získáme matici $\mathbf{Y}$ rozměru $11 \times n$. Teď ještě vypočteme vektor $\overline{\mathbf{Y}}$, kde $\overline{\mathbf{Y}} = \frac{1}{n}\mathbf{Y}\mathbf{1}_n$. Odhad Fisherovy informační matice pak zjistíme dle vztahu (53).

$$\hat{\mathbf{F}} = \frac{1}{n}(\mathbf{Y} - \mathbf{1}_n \otimes \overline{\mathbf{Y}})(\mathbf{Y} - \mathbf{1}_n \otimes \overline{\mathbf{Y}})^T \tag{53}$$

Označíme-li $\hat{\boldsymbol{\beta}} = (\hat{c}, \hat{\mu}_{11}, \hat{\mu}_{12}, \hat{\mu}_{21}, \hat{\mu}_{22}, \hat{\sigma}_{11}, \hat{\sigma}_{12}, \hat{\sigma}_{22}, \hat{\sigma}'_{11}, \hat{\sigma}'_{12}, \hat{\sigma}'_{22})^T$ vektor odhadovaných parametrů, tak přesnost odhadů zjistíme z diagonály varianční matice tohoto vektoru, na které jsou rozptyly odhadů jednotlivých parametrů. Varianční matici získáme pomocí vypočtené Fisherovy informační matice podle vztahu 54.

$$var(\hat{\boldsymbol{\beta}}) \cong \hat{\mathbf{F}}^{-1}. \tag{54}$$

$$\mathbf{Y} = \begin{pmatrix}
\frac{\partial f(\boldsymbol{y}_1)}{\partial c}\frac{1}{f(\boldsymbol{y}_1)}, & \frac{\partial f(\boldsymbol{y}_2)}{\partial c}\frac{1}{f(\boldsymbol{y}_2)}, & \cdots, & \frac{\partial f(\boldsymbol{y}_n)}{\partial c}\frac{1}{f(\boldsymbol{y}_n)} \\
\frac{\partial f(\boldsymbol{y}_1)}{\partial \mu_{11}}\frac{1}{f(\boldsymbol{y}_1)}, & \frac{\partial f(\boldsymbol{y}_2)}{\partial \mu_{11}}\frac{1}{f(\boldsymbol{y}_2)}, & \cdots, & \frac{\partial f(\boldsymbol{y}_n)}{\partial \mu_{11}}\frac{1}{f(\boldsymbol{y}_n)} \\
\frac{\partial f(\boldsymbol{y}_1)}{\partial \mu_{12}}\frac{1}{f(\boldsymbol{y}_1)}, & \frac{\partial f(\boldsymbol{y}_2)}{\partial \mu_{12}}\frac{1}{f(\boldsymbol{y}_2)}, & \cdots, & \frac{\partial f(\boldsymbol{y}_n)}{\partial \mu_{12}}\frac{1}{f(\boldsymbol{y}_n)} \\
\frac{\partial f(\boldsymbol{y}_1)}{\partial \mu_{21}}\frac{1}{f(\boldsymbol{y}_1)}, & \frac{\partial f(\boldsymbol{y}_2)}{\partial \mu_{21}}\frac{1}{f(\boldsymbol{y}_2)}, & \cdots, & \frac{\partial f(\boldsymbol{y}_n)}{\partial \mu_{21}}\frac{1}{f(\boldsymbol{y}_n)} \\
\frac{\partial f(\boldsymbol{y}_1)}{\partial \mu_{22}}\frac{1}{f(\boldsymbol{y}_1)}, & \frac{\partial f(\boldsymbol{y}_2)}{\partial \mu_{22}}\frac{1}{f(\boldsymbol{y}_2)}, & \cdots, & \frac{\partial f(\boldsymbol{y}_n)}{\partial \mu_{22}}\frac{1}{f(\boldsymbol{y}_n)} \\
\frac{\partial f(\boldsymbol{y}_1)}{\partial \sigma_{11}}\frac{1}{f(\boldsymbol{y}_1)}, & \frac{\partial f(\boldsymbol{y}_2)}{\partial \sigma_{11}}\frac{1}{f(\boldsymbol{y}_2)}, & \cdots, & \frac{\partial f(\boldsymbol{y}_n)}{\partial \sigma_{11}}\frac{1}{f(\boldsymbol{y}_n)} \\
\frac{\partial f(\boldsymbol{y}_1)}{\partial \sigma_{12}}\frac{1}{f(\boldsymbol{y}_1)}, & \frac{\partial f(\boldsymbol{y}_2)}{\partial \sigma_{12}}\frac{1}{f(\boldsymbol{y}_2)}, & \cdots, & \frac{\partial f(\boldsymbol{y}_n)}{\partial \sigma_{12}}\frac{1}{f(\boldsymbol{y}_n)} \\
\frac{\partial f(\boldsymbol{y}_1)}{\partial \sigma_{22}}\frac{1}{f(\boldsymbol{y}_1)}, & \frac{\partial f(\boldsymbol{y}_2)}{\partial \sigma_{22}}\frac{1}{f(\boldsymbol{y}_2)}, & \cdots, & \frac{\partial f(\boldsymbol{y}_n)}{\partial \sigma_{22}}\frac{1}{f(\boldsymbol{y}_n)} \\
\frac{\partial f(\boldsymbol{y}_1)}{\partial \sigma'_{11}}\frac{1}{f(\boldsymbol{y}_1)}, & \frac{\partial f(\boldsymbol{y}_2)}{\partial \sigma'_{11}}\frac{1}{f(\boldsymbol{y}_2)}, & \cdots, & \frac{\partial f(\boldsymbol{y}_n)}{\partial \sigma'_{11}}\frac{1}{f(\boldsymbol{y}_n)} \\
\frac{\partial f(\boldsymbol{y}_1)}{\partial \sigma'_{12}}\frac{1}{f(\boldsymbol{y}_1)}, & \frac{\partial f(\boldsymbol{y}_2)}{\partial \sigma'_{12}}\frac{1}{f(\boldsymbol{y}_2)}, & \cdots, & \frac{\partial f(\boldsymbol{y}_n)}{\partial \sigma'_{12}}\frac{1}{f(\boldsymbol{y}_n)} \\
\frac{\partial f(\boldsymbol{y}_1)}{\partial \sigma'_{22}}\frac{1}{f(\boldsymbol{y}_1)}, & \frac{\partial f(\boldsymbol{y}_2)}{\partial \sigma'_{22}}\frac{1}{f(\boldsymbol{y}_2)}, & \cdots, & \frac{\partial f(\boldsymbol{y}_n)}{\partial \sigma'_{22}}\frac{1}{f(\boldsymbol{y}_n)}
\end{pmatrix}.$$

Ukažme si tedy, jak vypadá diagonála $var(\hat{\boldsymbol{\beta}})$, na které jsou rozptyly odhadů jednotlivých parametrů. Po výpočtu dostáváme hodnoty

$$\text{Diag}(var(\hat{\boldsymbol{\beta}})) = (0,86; 34420; 62654; 184980; 231420;$$

$$2,42 * 10^{-7}; 2,49 * 10^{-7}; 4,54 * 10^{-8}; 3,49 * 10^{-7}; 4,65 * 10^{-7}; 9,34 * 10^{-8}).$$

Přesnost odhadů je dána směrodatnými odchylkami, které v tomto případě jsou

$$(0,93; 185; 250; 430; 481; 0,0005; 0,0005; 0,0002; 0,0006; 0,0007; 0,0003).$$

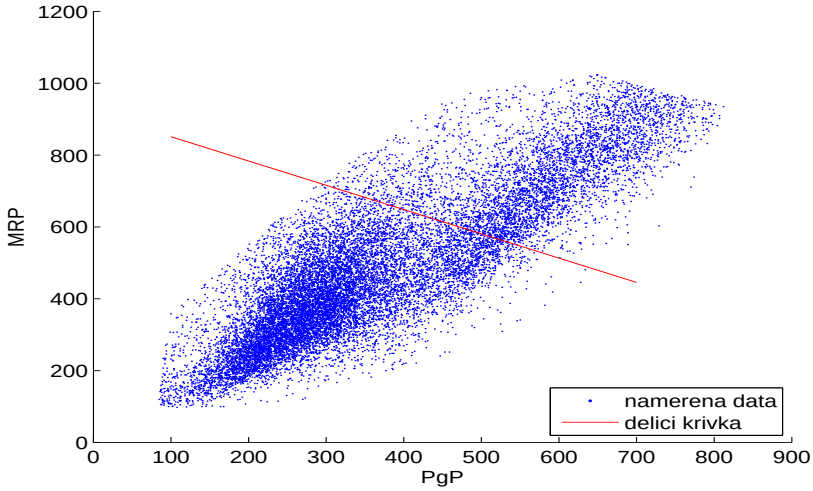
Všimněme si, že odhady váhy první složky a středních hodnot mají vysoké směrodatné odchylky vzhledem ke svým hodnotám. Především směrodatná odchylka odhadu váhy první složky je dokonce vyšší $(0,93)$ než samotný odhad parametru $(0,7087)$, který může nabývat hodnot pouze na intervalu $(0,1)$. Zatímco odhady prvků varianční matice jsou velmi přesné. Může to být způsobeno například špatnou volbou dělící křivky pro výpočet empirických variančních matic. Předpoklad normality by vzhledem k vysokému počtu a charakteru dat neměl být porušen. Srovnejme pro přehlednost ještě odhady a jejich směrodatné odchylky do tabulky (viz tabulka 6).

Tabulka 6

Parametry					
	c	μ_{11}	μ_{12}	μ_{21}	μ_{22}
odhad	0,7087	294,0	387,2	569,7	729,3
směrodatná odchylka	0,93	185	250	430	481
směrodatná odchylka/odhad (%)	131%	63%	65%	75%	66%

	σ_{11}	σ_{12}	σ_{22}	σ'_{11}	σ'_{12}	σ'_{22}
odhad	7802	7137	15880	10005	8494	17133
směrodatná odchylka	0,0005	0,0005	0,0002	0,0006	0,0007	0,0003

Podívejme se ještě jednou na volbu dělící křivky při výpočtu empirických variančních matic. Určili jsme, že křivka má tvar $x + 1,4796y - 1359,39 = 0$. Zkusme ji zakreslit do grafu společně s daty (viz obrázek 24, kód v příloze A10). Vidíme, že dělící křivka je určena poměrně nevhodně, protože zasahuje i do druhého „shluku“, který vymezuje druhou složku distribuce. Proto zkusme určit dělící křivku lépe. Vyzkoušením několika tvarů přímek jsme odhadli, že dělící křivka by mohla mít tvar $4,8x + y - 2550 = 0$ (viz obrázek 25, kód analogicky jako A10). Pro tuto novou volbu vypočítejme odhady parametrů smíšené distribuce a inverzi Fisherovy informační matice a srovnejme je s původními hodnotami.



Obrázek 24: Volba dělící křivky

Původní odhady:

$$\hat{\pi} = 0,7087, \quad \hat{\boldsymbol{\mu}}_1 = \begin{pmatrix} 294,0 \\ 387,2 \end{pmatrix}, \quad \hat{\boldsymbol{\mu}}_2 = \begin{pmatrix} 569,7 \\ 729,3 \end{pmatrix},$$

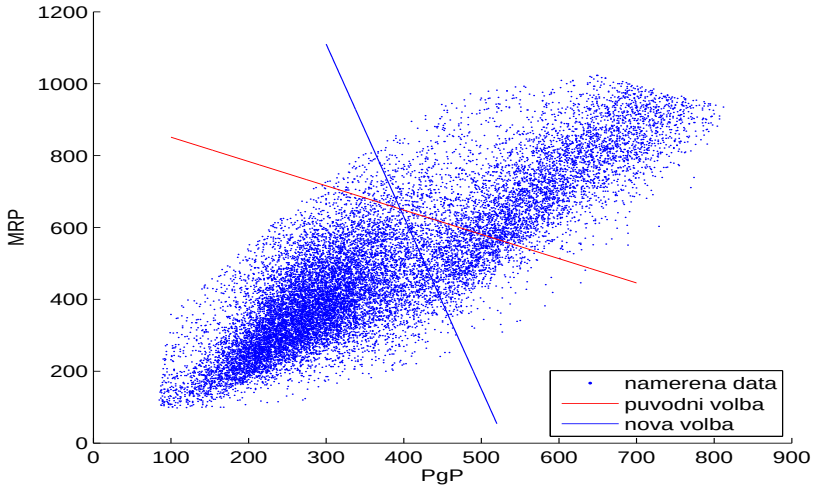
$$\hat{\Sigma}_1 = \begin{pmatrix} 7802 & 7137 \\ 7137 & 15880 \end{pmatrix}, \quad \hat{\Sigma}_2 = \begin{pmatrix} 10005 & 8494 \\ 8494 & 17133 \end{pmatrix}.$$

Nové odhady:

$$\hat{\pi} = 0,6553, \quad \hat{\boldsymbol{\mu}}_1 = \begin{pmatrix} 278,9 \\ 381,5 \end{pmatrix}, \quad \hat{\boldsymbol{\mu}}_2 = \begin{pmatrix} 557,1 \\ 688,7 \end{pmatrix},$$

$$\hat{\Sigma}_1 = \begin{pmatrix} 5433 & 6273 \\ 6273 & 16944 \end{pmatrix}, \quad \hat{\Sigma}_2 = \begin{pmatrix} 10158 & 11897 \\ 11897 & 25043 \end{pmatrix}.$$

Vidíme, že odhady středních hodnot se nepatrně posunuly, i důležitost složek v distribuci se moc nezměnila. Hodnoty v první varianční matici se snížily, tudíž máme menší rozptýlenost hodnot kolem střední hodnoty, ale ve druhé varianční matici se hodnoty naopak zvýšili. Celkově lze říci, že nová volba dělící křivky příliš odhady nezlepšila, i když, jak uvidíme dále, nějaké pozitivní změny proběhly. Nakonec se ještě podívejme na inverzi Fisherovy informační matice, respektive



Obrázek 25: Nová volba dělící křivky

na její diagonálu. Připomínám, že nás znepokojovali vysoké směrodatné odchylky odhadů středních hodnot a váhy první složky distribuce. Prvky z diagonály jsou uspořádány ve vektoru

$$(0,78; 142; 217; 339; 409; 0,0008; 0,0006; 0,0002; 0,0007; 0,0007; 0,0002).$$

Srovnejme opět nové odhady parametrů s příslušnými směrodatnými odchylkami v tabulce 7.

Tabulka 7

Parametry					
	c	μ_{11}	μ_{12}	μ_{21}	μ_{22}
odhad	0,6553	278,9	381,5	557,1	688,7
směrodatná odchylka	0,78	142	217	339	409
směrodatná odchylka/odhad (%)	119%	51%	57%	61%	59%

	σ_{11}	σ_{12}	σ_{22}	σ'_{11}	σ'_{12}	σ'_{22}
odhad	7802	7137	15880	10005	8494	17133
směrodatná odchylka	0,0008	0,0006	0,0002	0,0007	0,0007	0,0002

Je vidět, že přesnější volba dělící čáry pomohla ke zpřesnění odhadů. To orientačně měříme podle procentuálního podílu hodnoty směrodatné odchylky a hodnoty odhadu. Při původní volbě dělící čáry jsou tyto podíly přibližně o $10 - 20$

procentních bodů vyšší než při zpřesněné volbě dělící čáry. Sledovat podíly u hodnot odhadů prvků variančních matic je v tomto případě bezpředmětné. Na základě tohoto zjištění by se dalo usuzovat, že čím přesnější volba dělící čáry bude, tím přesnější odhady dostaneme.

4.2.2 2. pacient

Zkusme ještě stejným postupem odhadnout parametry pro jiného pacienta, abychom se přesvědčili o jedinečnosti hodnot P-gP a MRP u každého pacienta. Připomeňme si raději, že průběh hodnot P-gp a MRP modelujeme dvousložkovou dvourozměrnou smíšenou distribucí z normálního rozdělení ve tvaru:

$$f(\boldsymbol{y}) = \pi\phi(\boldsymbol{y}; \boldsymbol{\mu}_1, \Sigma_1) + (1-\pi)\phi(\boldsymbol{y}; \boldsymbol{\mu}_2, \Sigma_2), \tag{55}$$

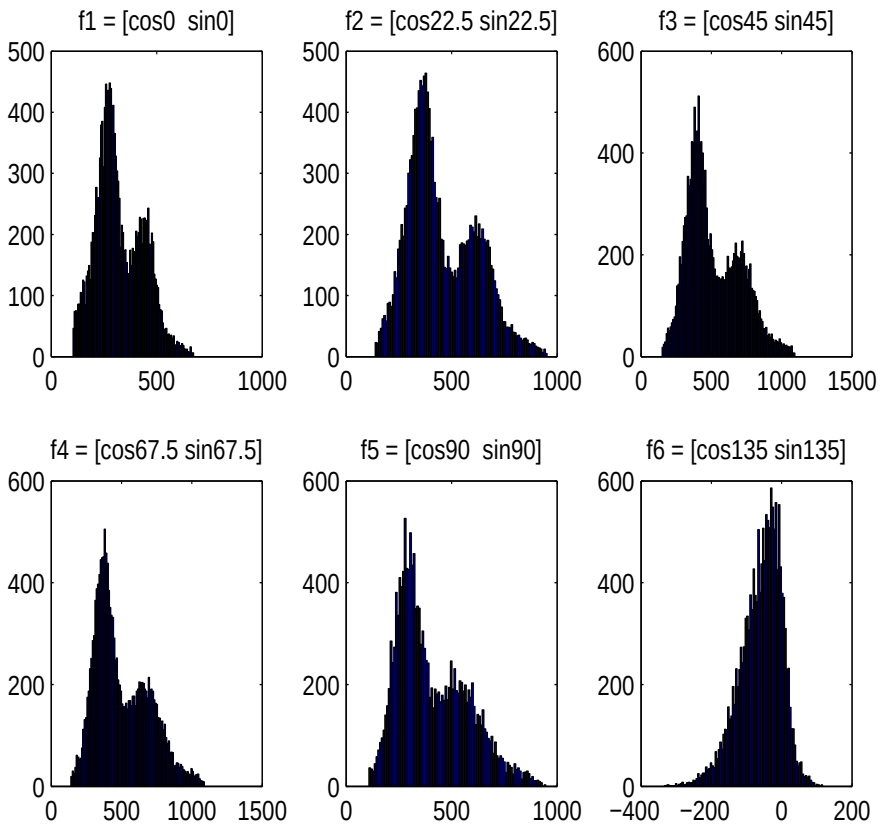
Stejně jako v předchozí části musíme nejprve určit prvotní odhady všech 11 parametrů, abychom mohli použít proceduru *csfinmix*. Vektory f_i volíme stejně. Poté provedeme projekce dat a získáme vektory h_i (podle vztahu (48)). Pro každý vektor h_i vytvoříme histogram četností, abychom mohli určit hodnoty a_1, a_2, b_1, b_2, pomocí kterých určíme prvotní odhady $\hat{\boldsymbol{\mu}}_1$ a $\hat{\boldsymbol{\mu}}_2$ (viz 49). Podívejme se na to, jak histogramy hodnot z vektorů h_i vypadají (viz obrázek 26, kód analogicky jako A7).

Teď bychom měli vybrat ty dva histogramy, jejichž „kopce“ jsou nejvzdálenější. V tomto případě to není tak jednoznačné. Vyberme tedy histogram h_2 a pro lehčí výpočty h_5. Hodnoty a_1, a_2, b_1, b_2 volíme následovně: $a_1 = 390, a_2 = 630, b_1 = 300, b_2 = 520$. Vypočteme prvotní odhady jsou

$$\hat{\boldsymbol{\mu}}_1 = \begin{pmatrix} 297,7 \\ 300 \end{pmatrix}, \hat{\boldsymbol{\mu}}_2 = \begin{pmatrix} 466,3 \\ 520 \end{pmatrix}.$$

Teď nám chybí určit jen prvotní odhady variančních matic $\hat{\Sigma}_1$ a $\hat{\Sigma}_2$. K tomu potřebujeme data rozdělit. Z předchozího příkladu víme, že bychom měli dělící čáru určit pokud možno co nejpřesněji. Na základě vykreslení všech dat jsme určili její analytický tvar následovně:

$$3x + y - 1500 = 0. \tag{56}$$

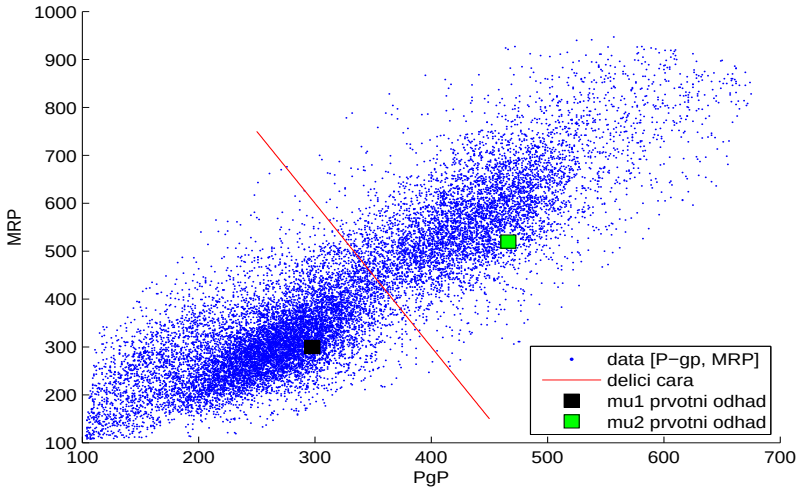


Obrázek 26: Histogramy četností po projekci na různé směry

Podívejme se na celou situaci na následujícím obrázku 27 (kód v příloze A11).

Vidíme, že jsme se přibližným postupem (tj. volbami a_1, a_2, b_1, b_2) netrefili s hodnotami parametrů $\boldsymbol{\mu}_1$ a $\boldsymbol{\mu}_2$ zcela ideálně. Nicméně, odhady ponechejme a podívejme se, jak si s touto nepřesností poradí iterační metoda *csfinmix*. Někdo by mohl namítnout, že by bylo lepší určit prvotní odhady $\hat{\boldsymbol{\mu}}_1$ a $\hat{\boldsymbol{\mu}}_2$ přímo z obrázku 27. Bylo by to možné, ale pouze v případě dvourozměrné smíšené distribuce. Při větším rozměru už nejsme schopni efektivně data znázornit, abychom mohli provést přímý odhad. Proto se i zde držíme výše zmíněného postupu.

Vypočtěme ještě odhady parametrů Σ_1 a Σ_2 a použijme metodu *csfinmix*



Obrázek 27: Volba dělící čáry u 2. pacienta

k určení výsledných odhadů parametrů pro 2. pacienta. Dostáváme hodnoty

$$\hat{\Sigma}_1 = \begin{pmatrix} 3303 & 2679 \\ 2679 & 6201 \end{pmatrix}, \hat{\Sigma}_2 = \begin{pmatrix} 4087 & 4680 \\ 4680 & 12383 \end{pmatrix}.$$

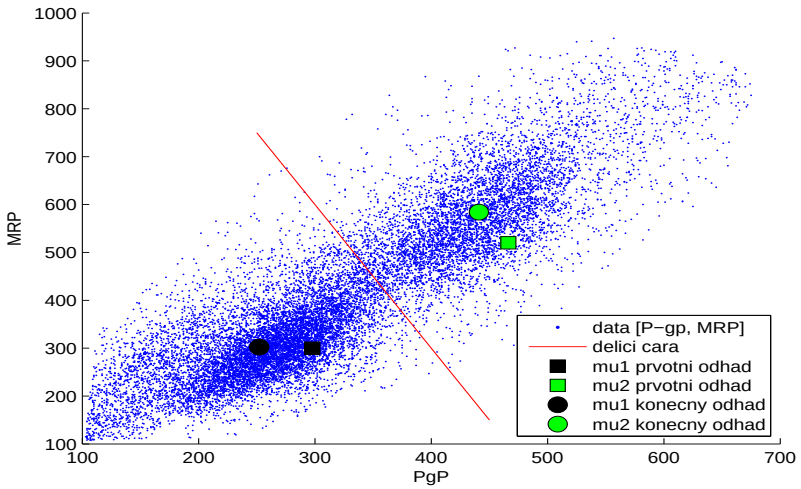
Po provedení 50 iterací dostáváme výsledné odhady parametrů smíšené distribuce:

$$\hat{\pi} = 0,6251, \quad \hat{\boldsymbol{\mu}}_1 = \begin{pmatrix} 252,2 \\ 302,2 \end{pmatrix}, \quad \hat{\boldsymbol{\mu}}_2 = \begin{pmatrix} 441,0 \\ 584,1 \end{pmatrix},$$

$$\hat{\Sigma}_1 = \begin{pmatrix} 3556 & 2848 \\ 2848 & 6116 \end{pmatrix}, \quad \hat{\Sigma}_2 = \begin{pmatrix} 4955 & 5751 \\ 5751 & 13951 \end{pmatrix}.$$

Poznámka 4.2 Vzhledem k tomu, že operujeme se vstupními daty, které mají cca 15000×2 hodnot, trvají výpočty poměrně dlouho. Požadované přesnosti bylo dosaženo při použití 50 iterací a výpočet odhadů trval 244 sekund. Kdybychom použili 60 iterací, doba výpočtu by se zvýšila na 320 sekund, při 100 iteracích na 556 sekund. Tak bychom mohli pokračovat dále. Připomeňme si, že v případě jednorozměrné smíšené distribuce jsme použili 206 iterací a výpočet trval jen několik sekund.

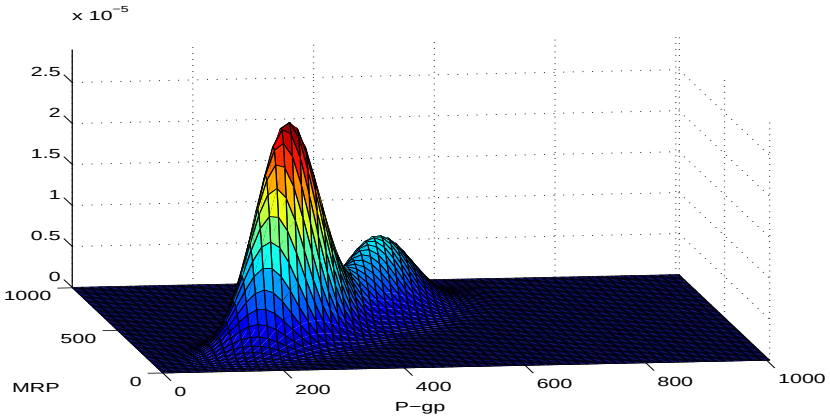
Nejprve se podívejme zdali a jak se změnila poloha prvotních odhadů $\hat{\boldsymbol{\mu}}_1$ a $\hat{\boldsymbol{\mu}}_2$. Změna je znázorněna na obrázku 28 (kód analogicky jako A11). Pozorujeme,



Obrázek 28: Posun hodnot odhadů $\hat{\boldsymbol{\mu}}_1$ a $\hat{\boldsymbol{\mu}}_2$

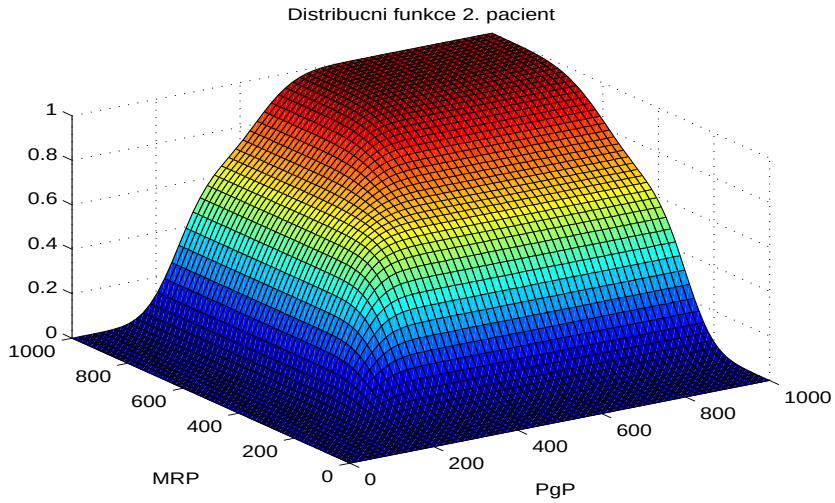
že výsledné odhady parametrů $\boldsymbol{\mu}_1$ a $\boldsymbol{\mu}_2$ odpovídají datům mnohem lépe než ty prvotní. Můžeme konstatovat, že metoda *csfinmix* si dokáže poradit i s neúplně přesnými prvotními odhady parametrů.

Dále se podívejme na graf odhadnuté smíšené distribuce a na její distribuční funkci (obrázky 29,30, kód analogicky jako A8,A9). Všimněme si velikosti „kopce“



Obrázek 29: Graf odhadnuté smíšené distribuce pro 2. pacienta

první složky, která je větší než u prvního pacienta, i přesto, že váha první složky druhého pacienta je menší než váha první složky pacienta prvního. Příčinu této skutečnosti můžeme pozorovat na obrázku 31 (kód analogicky jako A8). U druhého pacienta jsou složky smíšené distribuce položeny blíže k sobě než u prvního. Více se vzájemně prolínají, a to způsobuje, výše popsaný jev. Graf distribuční funkce smíšené distribuce druhého pacienta vypadá následovně:

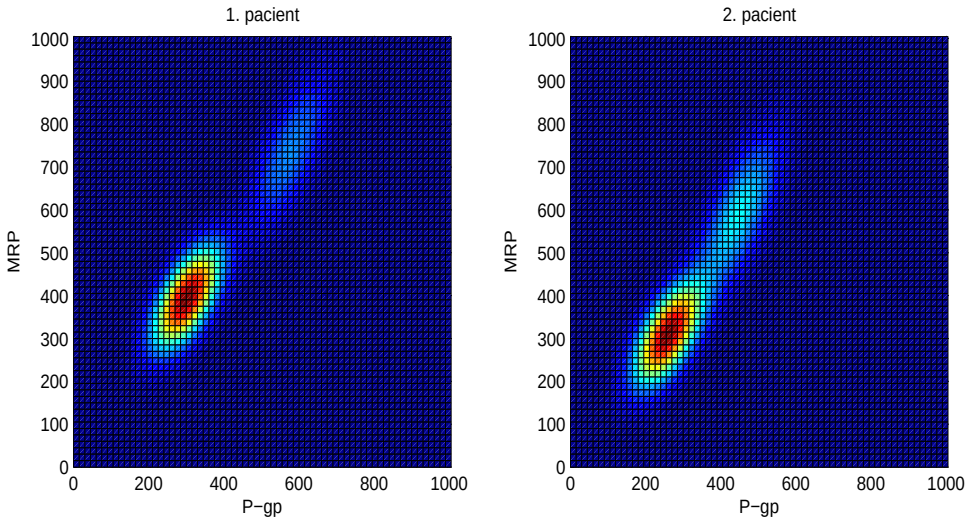


Obrázek 30: Distribuční funkce odpovídající odhadlé smíšené distribuci druhého pacienta

Doplňme ještě velikosti korelačních koeficientů mezi P-gp a MRP v obou složkách distribuce a projekci smíšené distribuce z ptačí perspektivy (viz obrázek 31), kterou srovnáme s prvním pacientem. Na obrázku uvidíme především polohu jednotlivých složek a jejich protažení, které je určeno korelačními koeficienty. Korelační koeficienty mají přibližně stejné hodnoty, tudíž je zachována vzájemná závislost mezi jednotlivými proteiny.

$$\rho^1_{(P-gp,MRP)} = 0,6107, \quad \rho^2_{(P-gp,MRP)} = 0,6917.$$

Na závěr výpočtů se ještě zaměříme na přesnost odhadů. Opět použijeme empirickou Fisherovu informační matici a zjistíme, jaké jsou hodnoty na diago-



Obrázek 31: Srovnání odhadnutých smíšených distribucí obou pacientů z ptačí perspektivy

nále $var(\hat{\boldsymbol{\beta}})$, na které jsou rozptyly odhadů jednotlivých parametrů. Po výpočtu dostáváme hodnoty

$$\mathrm{Diag}(var(\hat{\boldsymbol{\beta}})) = (0,5063; 13017; 19764; 40827; 123470;$$

$$9,36 * 10^{-7}; 1,05 * 10^{-6}; 3,13 * 10^{-7}; 5,92 * 10^{-7}; 6,87 * 10^{-7}; 9,86 * 10^{-8}).$$

Přesnost odhadů určíme pomocí směrodatných odchylek, které v tomto případě jsou

$$(0,711; 114; 141; 202; 351; 0,0009; 0,0010; 0,0006; 0,0008; 0,0008; 0,0003).$$

Opět pozorujeme, že odhady prvků variančních matic jsou velmi přesné, zatímco odhady středních hodnot a váhy první složky přesné až tak nejsou. Pro lepší posouzení se na přesnost odhadů podívejme v tabulce 8.

Tabulka 8

Parametry					
	c	μ_{11}	μ_{12}	μ_{21}	μ_{22}
odhad	0,6251	252,2	302,2	441,0	584,1
směrodatná odchylka	0,711	114	141	202	351
směrodatná odchylka/odhad (%)	114%	45%	47%	46%	60%

	σ_{11}	σ_{12}	σ_{22}	σ'_{11}	σ'_{12}	σ'_{22}
odhad	3556	2848	6116	4955	5751	13951
směrodatná odchylka	0,0009	0,0010	0,0006	0,0008	0,0008	0,0003

Nyní zkusme shrnout všechny výsledky a poznatky, kterých jsme v případě dvourozměrných smíšených distribucí docílili.

4.3 Shrnutí

Ukázalo se, že metody odhadů parametrů smíšených distribucí v případě vícerozměrných dat jsou mnohem složitější a náročnější než v případě dat jednorozměrných. V našem příkladě jsme odhadli parametry dvourozměrných dvousložkových smíšených distribucí pro 2 pacienty. Kdybychom chtěli data modelovat pomocí vícesložkových (3,4,..) smíšených distribucí, vznikaly by další problémy typu: „Jak rozdělit soubor dat pro určení počátečních parametrů?“, „Jak identifikovat složky?“. Proto jsme se v této práci zaměřili pouze na dvousložkové smíšené distribuce, u kterých tyto problémy jsou poměrně snadno řešitelné.

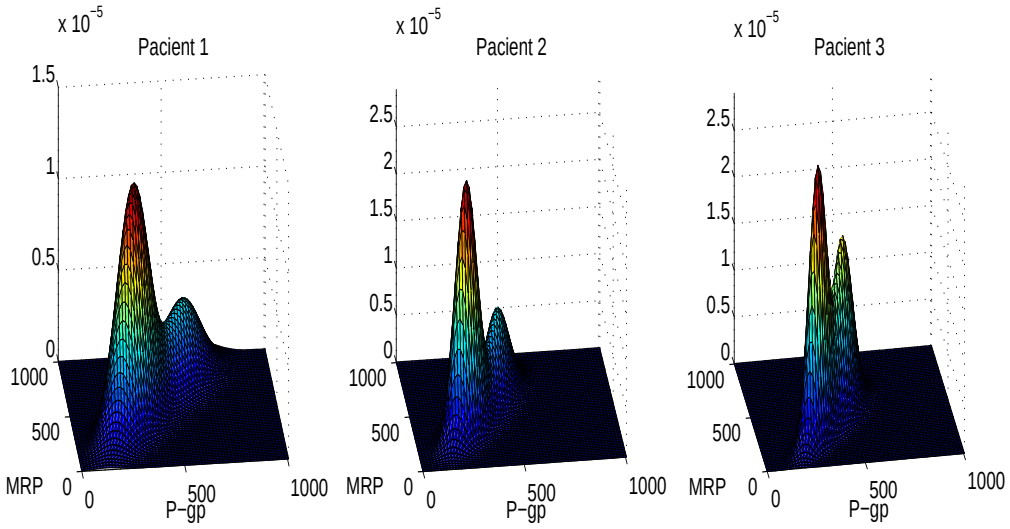
Opět se potvrdila individualita jednotlivých pacientů v případě naměřených hodnot P-gp a MRP, která se projevila především umístěním středních hodnot složek. Pro větší průkaznost ještě přidáme odhadlé střední hodnoty dalšího pacienta a výsledky srovnáme v tabulce 9. Především hodnoty $\boldsymbol{\mu}_2$ se mění poměrně hodně. I poměr vah složek v modelech se liší. U prvního pacienta má první složka váhu 70,91%, u druhého 62,51% a u třetího pouze 49,17%.

Tabulka 9

Srovnání odhadnutých středních hodnot			
	pacient 1	pacient 2	pacient 3
μ_{11}	294	252	299
μ_{12}	387	302	246
μ_{21}	570	441	461
μ_{22}	729	584	474

Váhy složek a různé hodnoty prvků odhadnutých variančních matic způsobují to, že výsledné grafy smíšených distribucí se chovají analogicky, jak je to ukázáno na obrázcích 8 a 9 (nejvíce jsou ovlivněny velikosti „kopců“ složek). Je to patrné

na obrázku 32 (kód analogicky jako A8) , kde „kopce“ v případě třetího pacienta jsou různě vysoké, přestože váhy jsou téměř v poměru 1:1.



Obrázek 32: Srovnání odhadnutých smíšených distribucí všech pacientů

Naopak korelace v datech vykazovala poměrně stabilní charakter. U první složky jsme určili korelační koeficienty postupně 0,6412, 0, 6107 a 0,7101. U druhé 0,6488, 0,6917 a 0,7103. Dále se ukázalo, že postup, který jsme používali, dává velmi přesné odhady prvků variančních matic v modelu, ale odhady vah a středních hodnot mají přesnost o několik řádů menší. Důležitou roli při odhadování měla i volba dělící čáry. Pozitivně můžeme hodnotit to, že metoda *csfinmix* i přes nepříliš přesné prvotní odhady středních hodnot (viz pacient 2) poskytuje odhady, které odpovídají datům poměrně dobře.

Závěr

Seznámili jsme se s tím, co jsou to smíšené distribuce, jak mohou vypadat a jak lze odhadovat parametry smíšených distribucí. V první kapitole jsme si ukázali, že smíšené distribuce se vyznačují velkou proměnlivostí a přizpůsobitelností. I malé změny hodnot vah, středních hodnot či prvků variančních matic mohou mít velký vliv na celkový tvar smíšené distribuce. Právě flexibilita a rozmanitost dělá ze smíšených distribucí užitečný nástroj pro modelování různé škály dat. Zjistili jsme, že i přes velkou početní náročnost metody maximální věrohodnosti existuje efektivní řešení odhadování parametrů v podobě EM algoritmu. Ten umožňuje najít řešení věrohodnostních rovnic jednoduchým iterativním způsobem.

V našem příkladě se ukázalo, že každý z pacientů, na kterých byla studie prováděna, má unikátní průběh P-gp a MRP. A proto je cenné, že můžeme použít metodu, která tuto nehomogenitu dokáže modelovat. Nejprve jsme modelovali pouze průběh P-gp. V této části byl postup poměrně jednoduchý a výsledky věrohodně odpovídali datům. Zjistili jsme, že použití většího počtu složek v modelu zajistí lepší analytický popis dat. Poté jsme analyzovali dvourozměrný soubor dat (P-gp, MRP). Postup odhadování parametrů již nebyl tak jednoduchý. Museli jsme najít způsob jak data rozdělit na části, ze kterých pak bude možné určit počáteční odhady parametrů. V obou částech jsme použili proceduru *csfinmix*. Dospěli jsme k poměrně uspokojivým odhadům. Překvapivá byla rozdílnost přesnosti odhadů parametrů v modelu, kterou jsme určili pomocí inverze empirické Fisherovy informační matice.

Při zpracovávání tematu jsme nenarazili na vážnější problémy. Jedním z problémů, který by bylo vhodné zkoumat, je určení počtu složek smíšené distribuce tak, abychom jich nepoužívali zbytečně moc, ale aby také dobře popisovaly data. K tomu je potřeba použití vhodných statistických testů, což by mohlo být předmětem dalšího rozšíření této práce. Další potenciální problém se týká výpočetní techniky. Při zpracovávání dat ze studie jsme počítali s řádově tisíci údaji a to se projevilo i na rychlosti výpočtů v programu Matlab. Obvykle nám zpracování běžných úloh nezabere déle než několik sekund. Zde trval výpočet odhadů spolu

s výpočtem přesností odhadů přibližně 7 minut. Proto by nás měli zajímat technické nároky na výpočty v případech, kdybychom počítali s většími objemy dat či složitějšími vzorci. Jiný problém se týká přesnosti odhadů, která již byla výše zmíněna.

Téma smíšených distribucí zahrnuje širokou problematiku. V naší práci jsme se zaměřili pouze na spojité náhodné veličiny (vektory) z normálního rozdělení. Celá tematika by si jistě zasloužila větší pozornost. Věříme, že celá práce čtenáře obohatí a že mu poskytne nejen dostatečné množství informací o problematice, ale dokáže i inspirovat.

Literatura

[1] Anderson, T.W., *An Introduction to Multivariate Statistical Analysis* , Wiley, 2003

[2] Eisenberger, I., *Genesis of bimodal distributions*, Technometrics **6**, 357-363, 1964

[3] Exprese genu [online], dostupné z http://cs.wikipedia.org/wiki/Exprese_genu [citováno 12. 3. 2008]

[4] Kubáček, L., *Základy teórie odhadu*, VEDA, Bratislava, 1983

[5] Kunderová, P., *Základy pravděpodobnosti a matematické statistiky,* Univerzita Palackého v Olomouci, Olomouc, 2004

[6] Martinez, W.L., Martinez, A.R., *Computional statistics handbook with Matlab,* Chapman and Hall, Boca Raton, 2002

[7] Martinez, W.L., Martinez, A.R., *Exploratory data analysis with Matlab,* Chapman and Hall, Boca Raton, London, New York, Washington D.C., 2005

[8] McLachlan, G., Peel, D., *Finite mixture models,* John and Sons, New York, 2000 T.W. Anderson, An Introduction to Multivariate Statistical Analysis, Wiley, 1984

Seznam příloh

1. Příloha A1: Kód v Matlabu pro vytvoření grafu třísložkové smíšené distribuce s danými parametry
2. Příloha A2: Kód v Matlabu pro vytvoření dF plotu modelu třísložkové smíšené distribuce
3. Příloha A3: Kód v Matlabu pro vytvoření grafu dvousložkových smíšených distribucí se stejnými vahami
4. Příloha A4: Kód v Matlabu pro vytvoření dvousložkové dvourozměrné smíšené distribuce
5. Příloha A5: Kód v Matlabu pro vytvoření histogramu četností P-gp 1. pacienta
6. Příloha A6: Kód v Matlabu pro vytvoření grafu srovnávajícího histogram 1. pacienta a spočtené dvou a třísložkové smíšené distribuce
7. Příloha A7: Kód v Matlabu pro vytvoření histogramů četností hodnot P-gp projektovaných na směrové vektory
8. Příloha A8: Kód v Matlabu pro vytvoření grafu odhadnuté smíšené distribuce 1. pacienta
9. Příloha A9: Kód v Matlabu pro vytvoření grafu odhadnuté distribuční funkce 1. pacienta
10. Příloha A10: Kód v Matlabu pro vytvoření grafu zobrazujícího data a dělící čáru
11. Příloha A11: Kód v Matlabu pro vytvoření grafu zobrazujícího data, dělící čáru a prvotní odhady parametrů v modelu

Přílohy

Příloha A1

Kód v Matlabu pro vytvoření grafu třísložkové smíšené distribuce s danými parametry

```
x=linspace(-6,5);
% create the model - normal components used
mix = [0.3 0.3 0.4]; % mixing coefficients
mus = [-3 0 2]; % terms means
vars = [1 1 0.5];
nterm= 3; % use Statistics toolbox function to evaluate normal pdf.
fhat = zeros(size(x));
for i= 1:nterm
   fhat = fhat+mix(i)*normpdf(x,mus(i),vars(i));
end
plot(x,fhat)
```

Příloha A2

Kód v Matlabu pro vytvoření dF plotu modelu třísložkové smíšené distribuce.

```
mu = [-3 0 2];
wts = [0.3 0.3 0.4];
covm = [1 1 0.5];
minx = -5;
maxx = 5;
csdfplot(mu,covm,wts,minx,maxx)
tick = (maxx-minx)/10;
set(gca,'Ytick',minx:tick:maxx)
set(gca,'Xtick',minx:tick:maxx)
set(gca,'YTickLabel','0|0.1|0.2|0.3|0.4|0.5|0.6|0.7|0.8|0.9|1')
xlabel('S'),ylabel('V')
```

Příloha A3

Kód v Matlabu pro vytvoření grafu dvousložkových smíšených distribucí se stejnými vahami

```
x=linspace(-3,7);
mix = [0.5 0.5]; mus = [0 1]; vars = [1 1]; nterm= 2;
fhat = zeros(size(x));
for i= 1:nterm
    fhat = fhat+mix(i)*normpdf(x,mus(i),vars(i));
end
subplot(2,2,1); plot(x,fhat); title('(a)'); axis([-3 7 0 0.4]);
xlabel('y'); ylabel('f(y)');

x=linspace(-3,7);
mix = [0.5 0.5]; mus = [0 2]; vars = [1 1]; nterm= 2;
fhat = zeros(size(x));
for i= 1:nterm
    fhat = fhat+mix(i)*normpdf(x,mus(i),vars(i));
end
subplot(2,2,2); plot(x,fhat); title('(b)'); axis([-3 7 0 0.4]);
xlabel('y'); ylabel('f(y)');

x=linspace(-3,7);
mix = [0.5 0.5]; mus = [0 3]; vars = [1 1]; nterm= 2;
fhat = zeros(size(x));
for i= 1:nterm
    fhat = fhat+mix(i)*normpdf(x,mus(i),vars(i));
end
subplot(2,2,3); plot(x,fhat); title('(c)'); axis([-3 7 0 0.4]);
xlabel('y'); ylabel('f(y)';
```

```
x=linspace(-3,7);
mix = [0.5 0.5]; mus = [0 4]; vars = [1 1]; nterm= 2;
fhat = zeros(size(x));
for i= 1:nterm
    fhat = fhat+mix(i)*normpdf(x,mus(i),vars(i));
end
subplot(2,2,4); plot(x,fhat); title('(d)'); axis([-3 7 0 0.4]);
xlabel('y'); ylabel('f(y)');
```

Příloha A4

Kód v Matlabu pro vytvoření dvousložkové dvourozměrné smíšené distribuce

```
mu = [2 2];
Sigma = [1 0 ;0 1];
x1 = -3:.2:3; x2 = -3:.2:3;
[X1,X2] = meshgrid(x1,x2);
F = mvnpdf([X1(:) X2(:)],mu,Sigma);
F = reshape(F,length(x2),length(x1));

mu = [0 0];
Sigma = [1 0 ;0 1];
x1 = -3:.2:3; x2 = -3:.2:3;
[X1,X2] = meshgrid(x1,x2);
F2 = mvnpdf([X1(:) X2(:)],mu,Sigma);
F2 = reshape(F2,length(x2),length(x1));

fhat = 0.5*F + 0.5*F2 ;
surf(x1,x2,fhat);
%caxis([min(F(:))-.5*range(F(:)),max(F(:))]);
axis([-3 3 -3 3 0 .1])
xlabel('x1'); ylabel('x2');
```

Příloha A5

Kód v Matlabu pro vytvoření histogramu četností P-gp 1. pacienta

```
load data1sloupec %data1sloupec obsahuje namerene hodnoty
hist(Sloupec1,100)
title('P-gp 1. pacient')
xlabel('Hodnota P-gp')
ylabel('Pocet hodnot')
```

Příloha A6

Kód v Matlabu pro vytvoření grafu srovnávajícího histogram 1. pacienta a spočtené dvou a třísložkové smíšené distribuce.

```
subplot(1,2,1)
load data1sloupec
[ww,xx]=ecdf(Sloupec1);ecdfhist(ww,xx,100);
hold on
x=linspace(0,1000);
mix = [0.6237 0.3763]; mus = [274.4845 539.6303];
vars = [71.84706 111.252];nterm= 2;
fhat = zeros(size(x));
for i= 1:nterm
    fhat = fhat+mix(i)*normpdf(x,mus(i),vars(i));
end
plot(x,fhat);xlabel('y');ylabel('f(y)')

subplot(1,2,2)
load data1sloupec
[ww,xx]=ecdf(Sloupec1);ecdfhist(ww,xx,100);
hold on
x=linspace(0,1000);
```

```
mix = [0.6253 0.3103 0.0644];mus = [273.9113 512.4047 683.3713];
vars = [71.166 93.06664 49.27677];
nterm= 3;
fhat = zeros(size(x));
for i= 1:nterm
    fhat = fhat+mix(i)*normpdf(x,mus(i),vars(i));
end
plot(x,fhat);xlabel('y');ylabel('f(y)')
```

Příloha A7

Kód v Matlabu pro vytvoření histogramů četností hodnot P-gp projektovaných na směrové vektory

```
load pacient1; %pacient1 je matice dat P-gp a MRP
f1 = [1;0];
f2 = [cos(pi/8);sin(pi/8)];
f3 = [1/sqrt(2);1/sqrt(2)];
f4 = [cos(3*pi/8);sin(3*pi/8)];
f5 = [0;1];
f6 = [1/sqrt(2);-1/sqrt(2)];

h1 = pacient1*f1;
h2 = pacient1*f2;
h3 = pacient1*f3;
h4 = pacient1*f4;
h5 = pacient1*f5;
h6 = pacient1*f6;

subplot(2,3,1)
hist(h1,100);
title('f1 = [cos0  sin0]');
```

```
subplot(2,3,2)
hist(h2,100);
title('f2 = [cos22.5 sin22.5]');
subplot(2,3,3)
hist(h3,100);
title('f3 = [cos45 sin45]');
subplot(2,3,4)
hist(h4,100);
title('f4 = [cos67.5 sin67.5]');
subplot(2,3,5)
hist(h5,100);
title('f5 = [cos90  sin90]');
subplot(2,3,6)
hist(h6,100);
title('f6 = [cos135 sin135]');
```

Příloha A8

Kód v Matlabu pro vytvoření grafu odhadnuté smíšené distribuce 1. pacienta

```
%1.slozka
mu = [294.19 387.34];
Sigma = [7820 7150;7150 15889]
x1 = 0:15:1005; x2 = 0:15:1005;
[X1,X2] = meshgrid(x1,x2);
F = mvnpdf([X1(:) X2(:)],mu,Sigma);
F = reshape(F,length(x2),length(x1));
%2.slozka
mu = [569.74 729.38];
Sigma = [9997 8475; 8475 17094]
x1 = 0:15:1005; x2 = 0:15:1005;
[X1,X2] = meshgrid(x1,x2);
```

```
F2 = mvnpdf([X1(:) X2(:)],mu,Sigma);
F2 = reshape(F2,length(x2),length(x1));

%vysledna smisena distribuce
fhat = 0.7091*F + 0.2909*F2;
surf(x1,x2,fhat);

axis([0 1005 0 1005 0 .000015])
xlabel('P-gp'); ylabel('MRP');
```

Příloha A9

Kód v Matlabu pro vytvoření grafu odhadnuté distribuční funkce 1. pacienta

```
figure
MU = [294.1928 387.3435;569.7417 729.3894];
SIGMA = cat(3,[7820 7150.9;7150.9 15889.6],[9997.4 8475.3;8475.3 17094]);
p = [0.7091 0.2909];
obj = gmdistribution(MU,SIGMA,p);

ezsurf(@(x,y)cdf(obj,[x y]),[0 1000],[0 1000])
xlabel('PgP'); ylabel('MRP');
title('Distribucni funkce 1. pacient');
```

Příloha A10

Kód v Matlabu pro vytvoření grafu zobrazujícího data a dělící čáru

```
figure
load pacient1;
scatter(pacient1(:,1),pacient1(:,2),10,'.')
xlabel('PgP'); ylabel('MRP');
hold on;
```

```
x = [100:0.1:700];
y=(-x+1359.39)/1.4796;
plot(x,y,'red')
```

Příloha A11

Kód v Matlabu pro vytvoření grafu zobrazujícího data, dělící čáru a prvotní odhady parametrů v modelu

```
figure
load pacient3;
scatter(pacient3(:,1),pacient3(:,2),10,'.')
xlabel('PgP'); ylabel('MRP');
hold on;
x = [250:0.1:450];
y=-3*x+1500;

plot(x,y,'red')
mu11=297.7;
mu12=300;
plot(mu11,mu12,'--rs','MarkerEdgeColor','k',...
'MarkerFaceColor','k','MarkerSize',10)
mu21=466.3;
mu22=520;
plot(mu21,mu22,'--rs','MarkerEdgeColor','k',...
'MarkerFaceColor','g','MarkerSize',10)
```