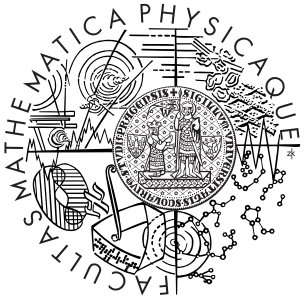
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Antonín Komora

Aplikace EM algoritmu

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce:
Ing. Marek Omelka, Ph. D

Studijní program: Matematika
Obor: Obecná matematika

2011

Rád bych poděkoval vedoucímu práce **Ing. Marku Omelkovi, Ph.D** za cenné rady, ochotu pomoci a projevenou trpělivost při tvorbě této práce.

Dále bych rád poděkoval Bc. Petru Kácovskému za nenahraditelnou pomoc při práci s programem TEX a všem, kteří mi pomáhali při vychytávání nedostatků v pozdějších etapách, obzvláště pak Bc. Kateřině Štádlerové.

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

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

V Praze dne Antonín Komora

Název práce: Aplikace EM algoritmu
Autor: Antonín Komora
Katedra: Katedra Pravděpodobnosti a Matematické Statistiky MFF UK
Vedoucí bakalářské práce: Ing. Marek Omelka, Ph.D.

Abstrakt: EM algoritmus je velmi cenným nástrojem pro výpočty statistických problémů, kde nám nejsou k dispozici všechna data. Jedná se o iterační algoritmus, který v prvním kroku hledá odhady chybějících hodnot na základě podoby parametru z předchozí iterace a zadaných dat. Činí tak přes podmíněné střední hodnoty. V další fázi metodami maximální věrohodnosti hledá odhad parametru maximalizující logaritmickou věrohodnostní funkci, který předá do další iterace. Tento postup je opakován až do bodu, kdy jsou přírůstky funkce mezi iteracemi tak malé, že se ukončení postupu na výsledku závažněji neprojeví. Důležitou charakteristikou je monotónní konvergence za značně obecných podmínek, ale ta na druhou stranu nepatří mezi nejrychlejší, a proto je mnohokrát zapotřebí velkého množtví iterací.

Klíčová slova: EM algoritmus, problém neúplných dat, metoda maximální věrohodnosti

Title: Application of EM-algorithm
Author: Antonín Komora
Department: Department of Probability and Mathematical Statistics MFF UK
Supervisor: Ing. Marek Omelka, Ph.D.

Abstract: EM algorithm is a very valuable tool in solving statistical problems, where the data presented is incomplete. It is an iterative algorithm, which in its first step estimates the missing data based on the parameter estimate from the last iteration and the given data and it does so by using the conditional expectation. In the second step it uses the maximum likelihood estimation to find the value that maximizes the logarithmic likelihood function and passes it along to the next iteration. This is repeated until the point, where the value increment of the logarithmic likelihood function is small enough to stop the algorithm without significant errors. A very important characteristic of this algorithm is its monotone convergence and that it does so under fairly general conditions. However the convergence itself is not very fast, and therefore at times requires a great number of iterations.

Keywords: EM algorithm, incomplete data problem, maximum likelihood method

Obsah

Kapitola 1

Úvod

Tento algoritmus byl oficiálně formulován roku 1977 v článku *Maximum Likelihood from Incomplete Data via the EM algorithm* autorů A. P. Dempstera, N. M. Lairda a D. B. Rubina a tím byl položen základ pro mnoho dalších prací, neboť se tento algoritmus velmi rychle stal značně oblíbeným statistickým nástrojem. Mezi autory, kteří se algoritmem zabývali a na jejichž díla je v této práci rovněž odkazováno, jsou například C. F. Jeff Wu (*On the convergence properties of the EM algorithm* v časopise *The Annals of Statistics*, 1983) nebo G. J. McLachlan a T. Krishnan (*The EM algorithm and Extensions*, 1997).

EM algoritmus (zkratka pro anglické Expectation-Maximization algorithm) je iteračním algoritmem, který je jedním ze způsobů výpočtů statistických problémů pomocí metod maximální věrohodnosti v případech, kdy neúplnost či absence jednodušší struktury dat buď neumožňuje anebo výrazně komplikuje užití těchto metod přímo. Tyto problémy jsou souborně nazývány jako problémy neúplných dat (orig. incomplete--data problems) a zahrnují nejen situace, kdy data skutečně chybí, ale i mnoho případů, kdy není neúplnost dat na první pohled patrná.

Základní myšlenkou je k problému neúplných dat nalézt vhodný problém dat úplných, který je možné efektivněji řešit jednoduššími prostředky, a propojit tyto dva v jeden celek. Ve své podstatě se jedná o přepis problému tak, aby vyhovoval způsobu řešení pro úplná data, zjištění vztahu mezi věrohodnostními funkcemi obou a využití jednoduššího postupu pro další výpočet. Existují však také problémy, kde přestože máme k dispozici data úplná (resp. dostatečná), je výhodnější na ně aplikovat EM algoritmus, neboť přidáním dodatečných proměnných, které často bývají s proměnnými v příkladu přímo spojeny, umožníme podstatně jednodušší či rychleji konvergující výpočet odhadu parametrů věrohodnostní funkce.

Samotný algoritmus pracuje ve dvou krocích - prvním je **E** (Expectation), kdy je dosazením parametrů za chybějící data vytvořen odhad věrohodnostní funkce pomocí podmíněné střední hodnoty, a druhým je **M** (Maximization), kde je metodou maximální věrohodnosti získaná funkce maximalizována. Získaný odhad parametru je poté využit v další iteraci k opětovnému odhadnutí parametrů a jejich maxima-

lizaci. Celý postup je opakován až do bodu, kdy je přírůstek věrohodnostní funkce již tak malý, že se zastavení algoritmu na výsledku nijak závažně neprojeví.

Zavedením potřebných proměnných, funkcí a postupů a následně také i popisem algoritmu se zabývá Kapitola 2. Zde si na Ilustračním příkladu 1 definujeme vše potřebné, a posléze na Příkladu 2 si ukážeme způsob práce v případě, kdy data skutečně chybí a odhadovaný parametr je ve složitějším tvaru (myšleno vícerozměrný a se složitější hustotou pro neúplná data), ale stále se jedná o známé a značně využívané rozdělení.

V Kapitole 3 rozebereme vlastnosti algoritmu, uvedeme základní tvrzení a nejdůležitější z nich předvedeme pomocí příkladů. Nejdříve ukážeme, že posloupnost hodnot logaritmické věrohodnostní funkce pro jednotlivé odhady parametrů, tedy $\{L(\boldsymbol{\theta}^{(k)})\}$, je monotónní, a následně ukážeme samotnou konvergenci této posloupnosti k nějaké hodnotě $L^* = L(\boldsymbol{\theta}^*)$, kde je rovněž důležité pojednat o tom, za jakých podmínek $\boldsymbol{\theta}^{(k)}$ konverguje k oné hodnotě $\boldsymbol{\theta}^*$, pro niž platí $L^* = L(\boldsymbol{\theta}^*)$.

V závěru shrneme základní charakteristiky tohoto algoritmu a zdůrazníme hlavní problémy. Rovněž zde budou ve zkrácené podobě zmíněna dvě jeho rozšíření (ECM a ECME), ktera byla formulována s cílem řešit některé z problémů algoritmu zmíněné v tomto oddíle.

Kapitola 2

Formulace algoritmu

2.1 Ilustrační příklad 1

Nejznámějším příkladem pro odhalení způsobu, jakým EM algoritmus postupuje, je ten, který uvedli Dempster a kol. [1977, str. 2-3] a který hlouběji rozvedli McLachlan a Krishnan [1997, str. 9-16]. Originální zadání je založeno na datech, která uvedl Rao [1965, str. 368-369], avšak pro potřeby této práce příklad vypočítáme pro trochu odlišné zadání založené na populaci zvířecího druhu *Panthera tigris*, tedy tygra[1].

V dnešní době se předpokládá, že posledních 5480 žijících tygrů je rozděleno mezi následujících pět poddruhů: bengálský (1700), indočínský (2800), sumaterský (450), sibiřský (470) a čínský (60). Aby zadání bylo v souladu s příkladem, který uvedli Dempster a kol. [1977, str. 2-3], tak budeme předpokládat, že rozdělení je multinomické v pěti kategoriích:

$$\boldsymbol{y} = (y_1, y_2, y_3, y_4, y_5)^T = (1700, 2800, 450, 470, 60)^T,$$

s odpovídajícími pravděpodobnostmi pro nějaké $\theta \in (0, 1)$:

$$\left(\frac{3}{16}\theta, \frac{1}{2} + \frac{1}{4}\theta, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{1}{16}\theta\right)^T. \tag{2.1}$$

V tomto příkladu budeme θ chápat jako parametr pro výpočet přesné podoby pravděpodobností, které odpovídají vstupním datům.

Označíme-li $C(\boldsymbol{y}) = \frac{(\sum_{i=1}^{5} y_i)!}{\prod_{i=1}^{5} y_i!}$, pak pro pravděpodobnost, že se multinomicky rozdělený náhodný vektor $\boldsymbol{Y} = (Y_1, Y_2, Y_3, Y_4, Y_5)$ rovná vektoru $\boldsymbol{y}$, platí:

$$\begin{aligned} P(Y_1 = y_1, \ldots, Y_5 = y_5) = C(\boldsymbol{y}) \ \times \ \left(\frac{3}{16}\theta\right)^{y_1} \times \\ \times \left(\frac{1}{2} + \frac{1}{4}\theta\right)^{y_2} \times \ \left(\frac{1}{4}(1-\theta)\right)^{y_3} \times \ \left(\frac{1}{4}(1-\theta)\right)^{y_4} \times \ \left(\frac{1}{16}\theta\right)^{y_5}. \end{aligned} \tag{2.2}$$

[1] číselné údaje o populaci odhadnuty dle informací z http://en.wikipedia.org/wiki/Tiger

Jelikož se jedná o diskrétní rozdělení, tak je tato pravděpodobnost zároveň i věrohodnostní funkcí. Proto na tuto pravděpodobnost budeme nadále odkazovat jako na funkci $g(\boldsymbol{y} \mid \theta)$. Jelikož je hlavním prostředkem pro výpočet logaritmická věrohodnostní funkce, kterou předpokládáme ve tvaru $L(\boldsymbol{y} \mid \theta) = \log g(\boldsymbol{y} \mid \theta)$, je v tomto případě označení funkce g čistě jen symbolickou záležitostí.

2.1.1 Přímý výpočet metodou maximální věrohodnosti

Než si předvedeme, jak EM algoritmus pracuje, nalezneme nejdříve maximálně věrohodný odhad parametru θ z funkce (2.2) pro

$$\boldsymbol{y} = (y_1, y_2, y_3, y_4, y_5)^T = (1700, 2800, 450, 470, 60)^T \ .$$

Získanou hodnotu využijeme v Tabulce 2.1 v závěru příkladu pro předvedení, jak algoritmus postupuje.

Nejprve upravíme (2.2) na tvar:

$$C(\boldsymbol{y}) \left(\frac{3}{16}\right)^{y_1} \theta^{y_1+y_5} \left(\frac{1}{4}\right)^{y_2} (2+\theta)^{y_2} \left(\frac{1}{4}\right)^{y_3+y_4} (1-\theta)^{y_3+y_4} \left(\frac{1}{16}\right)^{y_5} .$$

Dále označíme $H(\boldsymbol{y}) = \left(\frac{3}{16}\right)^{y_1} \left(\frac{1}{4}\right)^{y_2+y_3+y_4} \left(\frac{1}{16}\right)^{y_5}$. Funkce $C(\boldsymbol{y})$ a $H(\boldsymbol{y})$ nezávisí na parametru θ, tedy se do dalšího výpočtu nepromítnou, a proto je není třeba rozepisovat. Jako další krok sestavíme logaritmickou věrohodnostní funkci:

$$L(\boldsymbol{y} \mid \theta) = \log C(\boldsymbol{y}) H(\boldsymbol{y}) + (y_1 + y_5) \log \theta + (y_3 + y_4) \log(1-\theta) + y_2 \log(2+\theta). \quad (2.3)$$

Hledání maximálně věrohodného odhadu θ je ve své podstatě snahou o nalezení maxima logaritmické věrohodnostní funkce. Zde můžeme využít derivaci podle θ:

$$\frac{\partial}{\partial \theta} L(\boldsymbol{y} \mid \theta) = \frac{y_1 + y_5}{\theta} - \frac{y_3 + y_4}{1-\theta} + \frac{y_2}{2+\theta} \ ;$$

položíme-li ji rovnu nule a doplníme-li číselné údaje, získáme tím rovnici:

$$\frac{1760}{\theta} - \frac{920}{1-\theta} + \frac{2800}{2+\theta} = 0 \ . \quad (2.4)$$

Vynásobíme-li tuto rovnici jejími jmenovateli, získáme rovnici kvadratickou. Ta má dva kořeny, z nichž jeden je záporný a nemusí nás tedy zajímat, druhý kořen je roven $0,7317828585$. Jestli se skutečně jedná o maximálně věrohodný odhad θ, zjistíme z druhé derivace:

$$\frac{\partial^2}{(\partial \theta)^2} L(\boldsymbol{y} \mid \theta) = -\frac{y_1 + y_5}{\theta^2} - \frac{y_3 + y_4}{(1-\theta)^2} - \frac{y_2}{(2+\theta)^2} \ .$$

Tato derivace je pro dané $\boldsymbol{y}$ a libovolné $0 < \theta < 1$ záporná. Tedy se jedná o maximum a hodnota maximálně věrohodného odhadu je $\theta^* = 0,7317828585$.

2.1.2 Výpočet pomocí EM algoritmu

Nyní ilustrujme, jakým způsobem EM algoritmus pracuje. Řekněme, že se populace tygrů indočínských (y_2) skládá ze dvou poddruhů (x_2 a x_3). Předpokládejme, že o nich v tuto chvíli víme jen to, že jich dohromady je 2800 a že pravděpodobnost je rozdělena na $\frac{1}{2}$ pro prvek x_2 a $\frac{1}{4}\theta$ pro x_3. Jinými slovy dochází k rozdělení prvku y_2 do dvou, čímž získáme vektor úplných dat $\boldsymbol{x} = (x_1, x_2, x_3, x_4, x_5, x_6)^T$, kde
$y_1 = x_1, y_2 = x_2 + x_3, y_3 = x_4, y_4 = x_5, y_5 = x_6$ s odpovídajícími pravděpodobnostmi:

$$\left(\frac{3}{16}\theta, \frac{1}{2}, \frac{1}{4}\theta, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{1}{16}\theta\right)^T.$$

Označíme-li $C'(\boldsymbol{x}) = \frac{(\sum_{i=1}^{6} x_i)!}{\prod_{i=1}^{6} x_i!}$, pak věrohodnostní funkce odpovídající vektoru úplných dat $\boldsymbol{x}$ je ve tvaru:

$$f(\boldsymbol{x} \mid \theta) = C'(\boldsymbol{x}) \left(\frac{3}{16}\theta\right)^{x_1} \left(\frac{1}{2}\right)^{x_2} \left(\frac{1}{4}\theta\right)^{x_3} \left(\frac{1}{4}(1-\theta)\right)^{x_4} \left(\frac{1}{4}(1-\theta)\right)^{x_5} \left(\frac{1}{16}\theta\right)^{x_6}$$
$$f(\boldsymbol{x} \mid \theta) = C'(\boldsymbol{x}) \left(\frac{3}{16}\right)^{x_1} \left(\frac{1}{2}\right)^{x_2} \left(\frac{1}{4}\right)^{x_3+x_4+x_5} \left(\frac{1}{16}\right)^{x_6} \theta^{x_1+x_3+x_6} (1-\theta)^{x_4+x_5}. \tag{2.5}$$

Abychom předešli nejasnostem při odkazování na krok E a práci se středními hodnotami, jimž rovněž náleží písmeno E, tak budeme po zbytek této práce střední hodnotu označovat jako $\mathbb{E}$. Symbol $\mathbb{E}_a$ snačí střední hodnotu, která je podmíněná nějakou pevnou hodnotou a

V kroku E algoritmu hledáme $Q(\theta \mid \theta^{(k)}) = \mathbb{E}_{\theta^{(k)}}[L(\boldsymbol{x} \mid \theta) \mid \boldsymbol{y}]$ pro vektor úplných dat $\boldsymbol{x}$ a předpokládáme vypozorovaná data $\boldsymbol{y}$ a odhad $\theta^{(k)}$. Tato funkce nám značně zjednodušuje výpočty věrohodnostní funkce, neboť ty části $\boldsymbol{x}$, které jsou shodné s některou z částí $\boldsymbol{y}$, jsou ve své podstatě nahrazeny, a tedy jediné, u čeho je třeba zjišťovat podmíněnou střední hodnotu, jsou neznámé (chybějící) části $\boldsymbol{x}$. V dalším postupu se stačí omezit na ty části věrohodnostní funkce, které obsahují $\theta^{(k)}$, a tedy[2]:

$$L_c(\boldsymbol{x} \mid \theta^{(k)}) = (x_1 + x_3 + x_6)\ \log \theta^{(k)} + (x_4 + x_5)\ \log(1 - \theta^{(k)}). \tag{2.6}$$

Pro další postup využijeme skutečnosti, že L_c je lineární funkcí našich jediných dvou neznámých x_2 a x_3, tedy tento krok spočívá pouze v jejich nahrazení odpovídající podmíněnou střední hodnotou. Jinými slovy předpokládáme-li náhodné veličiny X_2 a X_3 odpovídající prvkům x_2 a x_3, pak z vlastností multinomického rozdělení (a jeho marginálních částí) jsou obě binomicky rozdělené s $n = y_2$ a jejich pravděpodobnosti odpovídají $\frac{\frac{1}{2}}{\frac{1}{2}+\frac{1}{4}\theta^{(k)}}$ a $\frac{\frac{1}{4}\theta^{(k)}}{\frac{1}{2}+\frac{1}{4}\theta^{(k)}}$. Proto z definice binomického rozdělení platí:

[2] index c značí věrohodnostní funkci pro úplná data

$$x_2^{(k)} = \mathbb{E}_{\theta^{(k)}}\left[X_2 \mid y_2\right] = y_2 \ \frac{\frac{1}{2}}{\frac{1}{2} + \frac{1}{4}\theta^{(k)}}$$

$$x_3^{(k)} = \mathbb{E}_{\theta^{(k)}}\left[X_3 \mid y_2\right] = y_2 \ \frac{\frac{1}{4}\theta^{(k)}}{\frac{1}{2} + \frac{1}{4}\theta^{(k)}} \ . \tag{2.7}$$

V kroku M algoritmus vezme upravený vektor úplných dat

$$\boldsymbol{x}^* = (1700, x_2^{(k)}, x_3^{(k)}, 450, 470, 60)^T$$

a metodou maximální věrohodnosti hledá odhad parametru $\theta^{(k)}$, který označí jako $\theta^{(k+1)}$. Postup je stejný jako ten uvedený výše a jelikož krok M navazuje na krok E, tak nyní stačí vzít pouze logaritmickou věrohodnost L_c (viz (2.6)) a dosadit do ní náš zkoumaný vektor úplných dat.

$$L_c(\boldsymbol{x}^* \mid \theta) = (x_3^{(k)} + 1760) \ \log\theta + 920 \ \log(1-\theta)$$

$$\frac{\partial}{\partial\theta} L_c(\boldsymbol{x}^* \mid \theta) = \frac{x_3^{(k)} + 1760}{\theta} - \frac{920}{1-\theta}$$

$$\frac{x_3^{(k)} + 1760}{\theta} - \frac{920}{1-\theta} = 0 \tag{2.8}$$

$$\theta^{(k+1)} = \frac{x_3^{(k)} + 1760}{x_3^{(k)} + 2680} = \frac{4560\ \theta^{(k)} + 3520}{5480\ \theta^{(k)} + 5360} \ . \tag{2.9}$$

2.1.3 Závěrečné shrnutí

Z Tabulky 2.1 plyne, že z počáteční hodnoty $\theta^0 = 0,5$ algoritmus postupuje velmi rychle a při výpočtu pro 10 desetinných míst je přírůstek již po sedmém kroku prakticky zanedbatelný. V tabulce je rovněž využita hodnota maximálně věrohodného odhadu získaná v oddíle 2.1.1., tedy $\theta^* = 0,7317828585$.

Tabulka 2.1: Průběh EM algoritmu pro multinomický případ

iterace	$\theta^{(k)}$	$\theta^* - \theta^{(k)}$	$\frac{\theta^* - \theta^{(k+1)}}{\theta^* - \theta^{(k)}}$
0	0,5000000000	0,2317828585	0,0678802389
1	0,7160493827	0,0157334758	0,0592237031
2	0,7308510638	0,0009317947	0,0587107515
3	0,7317281522	0,0000547064	0,0586806348
4	0,7317796483	0,0000032102	0,0586788675
5	0,7317826702	0,0000001884	0,0586787632
6	0,7317828475	0,0000000111	0,0586787548
7	0,7317828579	0,0000000006	0,0586786168

- Třetí sloupec udává odchylku od hodnoty θ^* (spočtené v oddíle 2.1.1).
- Čtvrtý sloupec obsahuje poměr odchylek ve dvou po sobě jdoucích iteracích (tento poměr se velmi rychle ustálí a je takřka konstantní).

EM algoritmus je využíván hlavně proto, že značně zjednodušuje hledání maximálně věrohodného odhadu. Kdyby byl totiž parametr složitěji zakomponován do rozložení pravděpodobností vstupních dat (2.1) nebo by byla věrohodnost (2.2) ve složitějším tvaru, tak by se při přímém výpočtu mohlo stát, že bychom hledali kořeny složitých nelineárních rovnic vysokých stupňů (v tomto případě až pátého). Aplikací algoritmu se však v kroku E dostáváme k odhadu jen neznámých (chybějících) dat a za zbývající dosadíme vstupy ze zadání, čímž v kroku M docházíme k podstatně jednodušší rovnici. V tomto příkladu to sice není příliš patrné, ale i tak jsme se z rovnice (2.4) dostali k jednodušší (2.8).

2.2 Zavedení základních prvků

V této části zavedeme základní prvky EM algoritmu a učiníme tak odkazem na předchozí ilustrační příklad. Mějme dva výběrové prostory $\mathcal{X}$ a $\mathcal{Y}$ a zobrazení $\varphi : \mathcal{X} \longrightarrow \mathcal{Y}$ definované předpisem $\varphi(\boldsymbol{x}) = \boldsymbol{y}$ pro $\boldsymbol{x} \in \mathcal{X}$, v němž jednomu prvku z $\mathcal{Y}$ může odpovídat mnoho prvků z $\mathcal{X}$, ale každému prvku $\boldsymbol{x}$ odpovídá jen jedno $\boldsymbol{y}$. V našem příkladu:

$$\varphi(\boldsymbol{x}) = \varphi(x_1, x_2, x_3, x_4, x_5, x_6) = (y_1, y_2, y_3, y_4, y_5)^T, \tag{2.10}$$

kde $y_1 = x_1$, $y_2 = x_2 + x_3$, $y_3 = x_4$, $y_4 = x_5$ a $y_5 = x_6$.

Vektor $\boldsymbol{y} = \varphi(\boldsymbol{x})$ nazveme vektorem neúplných dat, $\boldsymbol{x}$ vektorem dat úplných. Důležitou charakteristikou je to, že s vektorem $\boldsymbol{x}$ není možné pracovat přímo, ale pouze skrze $\boldsymbol{y}$. O $\boldsymbol{x}$ víme jen to, že leží v nějaké podmnožině $\mathcal{X}$ definované jako $\mathcal{X}(\boldsymbol{y}) = \{\hat{\boldsymbol{x}} \mid \varphi(\hat{\boldsymbol{x}}) = \boldsymbol{y}\}$, kde $\boldsymbol{y}$ jsou vypozorovaná data.

Na prostoru $\mathcal{X}$ předpokládáme rodinu rozdělení závislých na vektoru parametrů $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)^T$ z parametrického prostoru Ω s hustotami ve tvaru $f(\boldsymbol{x} \mid \boldsymbol{\theta})$. (V našem příkladu se jedná o jednorozměrný parametr z prostoru $\Omega = (0, 1)$ a hustota odpovídá věrohodnosti multinomického rozdělení - viz (2.5)). Z této rodiny lze odvodit rodinu rozdělení na $\mathcal{Y}$ s hustotami $g(\boldsymbol{y} \mid \boldsymbol{\theta})$ - viz (2.2). Platí:

$$g(\boldsymbol{y} \mid \boldsymbol{\theta}) = \int_{\mathcal{X}(\boldsymbol{y})} f(\boldsymbol{x} \mid \boldsymbol{\theta}) \mathrm{d}\nu(\boldsymbol{x}),$$

kde pro míru ν platí: $\nu(\boldsymbol{x}) = \bigotimes_{i=1}^{6} \nu_i(x_i)$, kde ν_i jsou čítací míry takové, že $\nu_2 \times \nu_3$ je definováno ve smyslu $x_2 + x_3 = y_2$ a zbývající ν_i jsou také definována ve smyslu (2.10). Hustota $f(\boldsymbol{x} \mid \boldsymbol{\theta})$ odpovídá úplným datům a $g(\boldsymbol{y} \mid \boldsymbol{\theta})$ datům neúplným.

EM algoritmus při zadaném $\boldsymbol{y}$ hledá hodnotu $\boldsymbol{\theta}$ maximalizující funkci $g(\boldsymbol{y} \mid \boldsymbol{\theta})$, ale činí tak přes odpovídající $f(\boldsymbol{x} \mid \boldsymbol{\theta})$. Samotné kroky algoritmu jsou prováděny přes věrohodnostní funkci, tedy nechť $L(\boldsymbol{\theta}) = \log g(\boldsymbol{y} \mid \boldsymbol{\theta})$ je logaritmická věrohodnostní funkce vektoru parametrů $\boldsymbol{\theta}$ založená na vektoru pozorování $\boldsymbol{y}$ (viz (2.3)) a $L_c(\boldsymbol{\theta}) = \log f(\boldsymbol{x} \mid \boldsymbol{\theta})$ budiž lodpovídající funkce pro vektor úplných dat $\boldsymbol{x}$. Jak již bylo uvedeno v úvodu, EM algoritmus v kroku E na základě zadaných neúplných dat a momentální podoby parametru odhadne chybějící části vektoru dat úplných, a v kroku M již pracujeme jen s věrohodnostní funkcí odpovídající tomuto upravenému vektoru $\boldsymbol{x}^*$, tedy $f(\boldsymbol{x}^* \mid \boldsymbol{\theta})$ (viz (2.6)).

Krok E: předpokládejme, že jsme z minulé iterace získali hodnotu parametru $\boldsymbol{\theta}^{(k)}$. Algoritmus znovu odhaduje hodnotu logaritmické věrohodnosti úplného datového souboru a postupuje přes funkci Q definovanou následovně:

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)}) = \mathbb{E}_{\boldsymbol{\theta}^{(k)}}[\log f(\boldsymbol{x} \mid \boldsymbol{\theta}) \mid \boldsymbol{y}]. \tag{2.11}$$

V našem příkladu se jedná o odhadnutí podmíněných středních hodnot pro neznámé x_2 a x_3, neboť zbylé členy jsou stejné jak pro vektor $\boldsymbol{x}$, tak pro $\boldsymbol{y}$.

Krok M: algoritmus maximalizuje funkci $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})$ z kroku E. Jinými slovy hledáme $\boldsymbol{\theta}^{(k+1)}$ takové, aby platilo:

$$Q(\boldsymbol{\theta}^{(k+1)} \mid \boldsymbol{\theta}^{(k)}) \;\geq\; Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)}) \qquad \forall \boldsymbol{\theta} \in \Omega \; .$$

Pro nás to znamenalo hledání maximálně věrohodného odhadu jednorozměrného parametru θ z věrohodnosti multinomického rozdělení $f(\boldsymbol{x}^* \mid \theta)$ (viz (2.5)), kde:

$$\boldsymbol{x}^* = (1700, x_2^{(k)}, x_3^{(k)}, 450, 470, 60)^T \;\text{ a }\; \theta = \theta^{(k)} \; .$$

Pro další iteraci již vezmeme místo $\boldsymbol{\theta}^{(k)}$ parametr $\boldsymbol{\theta}^{(k+1)}$ a opakujeme, dokud není přírůstek L (tedy $L(\boldsymbol{\theta}^{(i)}) - L(\boldsymbol{\theta}^{(i-1)})$ pro všechna i) již tak malý, že ho lze zanedbat. Tedy co se týče konvergence posloupnosti $\{L(\boldsymbol{\theta}^{(k)})\}$ - viz Kapitola 3.

2.3 Regulární exponenciální rodina

Tato rodina je velmi důležitá pro aplikaci EM algoritmu, neboť obsahuje většinu často využívaných a široce známých rozdělení (např. multinomické, normální, χ^2 apod.). Hustotu absolutně spojitých (resp. věrohodnost diskrétních) rozdělení z této rodiny je možné převést do obecného tvaru:

$$g(\boldsymbol{x} \mid \boldsymbol{\theta}) = \frac{b(\boldsymbol{x})}{a(\boldsymbol{\theta})} \exp\left[\boldsymbol{\theta}^T t(\boldsymbol{x})\right] \ , \tag{2.12}$$

kde $t(\boldsymbol{x})$ je suficientní statistika vektoru parametrů $\boldsymbol{\theta} \in \Omega$ (tedy podmíněné rozdělení vektoru $\boldsymbol{x}$ při daném $t(\boldsymbol{x})$ nezávisí na $\boldsymbol{\theta}$ - viz Anděl [2007, str. 124]), pro niž platí:

$$\mathbb{E}_{\boldsymbol{\theta}}\left[t(\boldsymbol{x})\right] = \frac{\partial}{\partial \boldsymbol{\theta}} \log a(\boldsymbol{\theta}) \ , \tag{2.13}$$

kde $a(\boldsymbol{\theta})$ a $b(\boldsymbol{x})$ jsou skalární funkce a Ω je konvexní parametrický prostor všech $\boldsymbol{\theta}$ takových, že platí:

$$\int_{\mathcal{X}} \frac{b(\boldsymbol{x})}{a(\boldsymbol{\theta})} \exp\left[\boldsymbol{\theta}^T t(\boldsymbol{x})\right] d\boldsymbol{x} < \infty.$$

Krok E: vynecháme členy, které neobsahují $\boldsymbol{\theta}$. Funkce Q je pro regulání exponenciální rodinu definována následovně (L_c zastupuje logaritmickou věrohodnostní funkci pro úplná data):

$$Q\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)}\right) = \mathbb{E}_{\boldsymbol{\theta}^{(k)}}\left[L_c(\boldsymbol{x} \mid \boldsymbol{\theta}) \mid \boldsymbol{y}\right] = \boldsymbol{\theta}^T \ \mathbb{E}_{\boldsymbol{\theta}^{(k)}}\left[t(\boldsymbol{x}) \mid \boldsymbol{y}\right] - \log \ a(\boldsymbol{\theta}). \tag{2.14}$$

Krok M: odhad $\boldsymbol{\theta}^{(k+1)}$ metodou maximální věrohodnosti získáme derivací (2.14) podle $\boldsymbol{\theta}$ a následně tento výraz položíme roven nule, tedy:

$$\mathbb{E}_{\boldsymbol{\theta}^{(k)}}\left[t(\boldsymbol{x}) \mid \boldsymbol{y}\right] - \frac{\partial}{\partial \boldsymbol{\theta}} \log \ a(\boldsymbol{\theta}) = 0 \ ,$$

označíme-li $t^{(k)} = \mathbb{E}_{\boldsymbol{\theta}^{(k)}}\left[t(\boldsymbol{x}) \mid \boldsymbol{y}\right]$, pak v souladu s definicí (2.13) získáme výsledný odhad jako řešení rovnice:

$$t^{(k)} = E_{\boldsymbol{\theta}}\left[t(\boldsymbol{x})\right] \ . \tag{2.15}$$

Důležitou vlastností regulární exponenciální rodiny je skutečnost, že má-li rovnice (2.15) řešení, je toto dáno jednoznačně (plyne z konkávity logaritmické věrohodnosti této rodiny). Avšak pokud tato rovnice řešitelná není, pak je třeba pro výpočet využít složitější postup, neboť $\boldsymbol{\theta}^{(k+1)}$ leží někde na hranici Ω.

Při výpočtu následujícího příkladu rozložíme exponent místo na $\boldsymbol{\theta}^T t(\boldsymbol{x})$ na tvar $c(\boldsymbol{\theta})^T t'(\boldsymbol{x})$ (t' je takovou modifikací t, aby hodnota exponentu zůstala zachována), kde $c(\boldsymbol{\theta})$ je vhodně zvolená vektorová funkce. Učiníme tak vycházejíce ze Zehnaovy věty, kterou uvedl Anděl [2007, věta 7.87, str. 148]. Tato věta nám zajišťuje, že maximálně věrohodný odhad je nezávislý na parametrizaci. V příkladu se to projeví tak, že zvolíme-li vhodnou variantu parametrizace, pak nám tato umožní následovat snazší cestu výpočtu podmíněných středních hodnot v kroku E.

2.4 Ilustrační příklad 2

Pro ilustraci vlastností absolutně spojitých členů regulární exponenciální rodiny přepokládejme, že Tabulka 2.2 odpovídá náhodnému výběru z dvourozměrného normálního rozdělení o neznámých středních hodnotách i rozptylech, kde u dvou pozorování chybí první a u dvou druhá složka - pro přehlednost si je označíme α, β, γ a δ.

Tabulka 2.2: Data pro dvourozměrné normální rozdělení

X	8	11	16	18	6	4	α	β	20	25
Y	10	14	16	15	20	4	18	22	γ	δ

Jinými slovy máme: $\boldsymbol{W}_i = (X_i, Y_i) \sim \mathrm{N}\left((\mu_1, \mu_2)^T, \boldsymbol{V}\right)$, kde $\boldsymbol{V} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}$ je varianční matice, $X_i \sim \mathrm{N}(\mu_1, \sigma_{11})$, $Y_i \sim \mathrm{N}(\mu_2, \sigma_{22})$ a $\sigma_{12} = \mathrm{cov}(X_i, Y_i)$ pro $i = 1, \ldots, 10$.

Pro další výpočty je vhodné zdůraznit, že $\boldsymbol{w}_i = (x_i, y_i)^T$ jsou definována tak, že $\boldsymbol{w}_7 = (\alpha, y_7)^T$, $\boldsymbol{w}_8 = (\beta, y_8)^T$, $\boldsymbol{w}_9 = (x_9, \gamma)^T$ a $\boldsymbol{w}_{10} = (x_{10}, \delta)^T$. Vektor napozorovaných dat $\boldsymbol{y}$ je tedy ve tvaru $\boldsymbol{y} = (\boldsymbol{w}_1^T, \ldots, \boldsymbol{w}_6^T, \boldsymbol{m}^T)^T$, kde $\boldsymbol{m} = (y_7, y_8, x_9, x_{10})^T$ zahrnuje zadané hodnoty v neúplných pozorováních (první či druhá složka daného pozorování je neznámá). Nechť $\boldsymbol{v} = (\boldsymbol{w}_1^T, \ldots, \boldsymbol{w}_{10}^T)^T$ je odpovídající vektor úplných dat, v němž jsou chybějící data vyjádřena vektorem $\boldsymbol{z} = (\alpha, \beta, \gamma, \delta)^T$.

V tomto případě je parametr ve tvaru $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma_{11}, \sigma_{22}, \sigma_{12})^T$. Parametrický prostor Ω zde odpovídá $\mathbb{R} \times \mathbb{R} \times (0, \infty) \times (0, \infty) \times \mathbb{R}$ a hustota dvourozměrného normálního rozdělení (viz Anděl [2007, věta 4.10, str. 65]) vypadá následovně:

$$g(\boldsymbol{w}_i \mid \boldsymbol{\theta}) = \frac{1}{2\pi\sqrt{\det \boldsymbol{V}}} \exp\left\{ -\frac{1}{2}(\boldsymbol{w}_i - \boldsymbol{\mu})^T \boldsymbol{V}^{-1} (\boldsymbol{w}_i - \boldsymbol{\mu}) \right\} ,$$

což v rozloženém tvaru odpovídá:

$$g(\boldsymbol{w}_i \mid \boldsymbol{\theta}) = \frac{1}{2\pi\sqrt{\det \boldsymbol{V}}} \exp\left\{ -\frac{\sigma_{11}\sigma_{22}}{2 \det \boldsymbol{V}} \times \right.$$
$$\left. \times \left(\frac{(x_i - \mu_1)^2}{\sigma_{11}} - \frac{2\sigma_{12}}{\sigma_{11}\sigma_{22}}(x_i - \mu_1)(y_i - \mu_2) + \frac{(y_i - \mu_2)^2}{\sigma_{22}} \right) \right\} . \quad (2.16)$$

Nyní je třeba převést hustotu tohoto rozdělení do tvaru (2.12). Toho dosáhneme hlavně úpravou exponentu. Roznásobíme a sdružíme jej tak, abychom osamostatnili jednotlivé mocniny proměnných x_i a y_i. Tedy pomocí elementárních úprav dosáhneme:

$$\exp\left\{-\frac{\sigma_{11}\sigma_{22}}{2\det \boldsymbol{V}}\left[\frac{(x_i-\mu_1)^2}{\sigma_{11}} - \frac{2\sigma_{12}}{\sigma_{11}\sigma_{22}}(x_i-\mu_1)(y_i-\mu_2) + \frac{(y_i-\mu_2)^2}{\sigma_{22}}\right]\right\} =$$

$$\exp\left\{P\left(\sigma_{22}x_i^2 + \sigma_{11}y_i^2 + 2(\sigma_{12}\mu_2 - \sigma_{22}\mu_1)x_i + 2(\sigma_{12}\mu_1 - \sigma_{11}\mu_2)y_i - 2\sigma_{12}x_iy_i\right)\right\} \times$$

$$\times \exp\left\{-P(\sigma_{22}\mu_1^2 - 2\sigma_{12}\mu_1\mu_2 + \sigma_{11}\mu_2^2)\right\} = \exp\{M\}\ \exp\{-H\}\ ,$$

kde $P = \frac{-1}{2\det \boldsymbol{V}}$. Jinými slovy, dospěli jsme k hustotě ve tvaru:

$$g(\boldsymbol{w}_i \mid \boldsymbol{\theta}) = \frac{e^M}{2\pi\ \sqrt{\det \boldsymbol{V}}\ e^H}\ .$$

Zohledníme-li náhodný výběr, pak v souladu s definicí (2.12) a Zehnaovou větou (viz Anděl [2007, věta 7.87, str. 148]) máme nyní:

$$b(\boldsymbol{x}) = \left(\frac{1}{2\pi}\right)^n \qquad a(\boldsymbol{\theta}) = \left[\sqrt{\det \boldsymbol{V}}\ \exp\{H\}\right]^n$$

$$c(\boldsymbol{\theta}) = P\left(\sigma_{22}, \sigma_{11}, 2(\sigma_{12}\mu_2 - \sigma_{22}\mu_1), 2(\sigma_{12}\mu_1 - \sigma_{11}\mu_2), -2\sigma_{12}\right)^T$$

$$t'(\boldsymbol{w}) = \left(\sum_{i=1}^n x_i^2, \sum_{i=1}^n y_i^2, \sum_{i=1}^n x_i, \sum_{i=1}^n y_i, \sum_{i=1}^n x_iy_i\right)^T . \tag{2.17}$$

Prvky $t'(\boldsymbol{w})$ si pro přehlednost dalšího postupu označíme jako T_1, T_2, T_3, T_4 a T_5.

2.4.1 Výpočet pomocí EM algoritmu

Krok E: předpokládejme, že z předchozí iterace máme získán odhad $\boldsymbol{\theta}^{(k)}$. Dle (2.14) závisí nynější krok na výpočtu podmíněné střední hodnoty $\mathbb{E}_{\boldsymbol{\theta}^{(k)}}[t(\boldsymbol{x}) \mid \boldsymbol{v}]$. Připomeňme, že $\boldsymbol{v} = (\boldsymbol{w}_1^T, \ldots, \boldsymbol{w}_{10}^T)^T$ je vektor úplných dat (včetně vektoru neznámých pozorování $\boldsymbol{z} = (\alpha, \beta, \gamma, \delta)^T$) Vzhledem k linearitě střední hodnoty se výpočet rozpadne na odhad pěti podmíněných středních hodnot T_i, pro $i = 1, \ldots, 5$.

Zde využijeme několika základních vlastností:

(i) $\mathbb{E}_{\boldsymbol{\theta}^{(k)}}\left[\sum_{i=1}^{10} X_i \mid \boldsymbol{v}\right] = \sum_{i=1}^{10} \mathbb{E}_{\boldsymbol{\theta}^{(k)}}[X_i \mid \boldsymbol{v}]$

(ii) $\mathrm{var}_{\boldsymbol{\theta}^{(k)}}[X_i \mid \boldsymbol{v}] = \mathbb{E}_{\boldsymbol{\theta}^{(k)}}[X_i^2 \mid \boldsymbol{v}] - \left(\mathbb{E}_{\boldsymbol{\theta}^{(k)}}[X_i \mid \boldsymbol{v}]\right)^2$

(iii) $\mathrm{cov}_{\boldsymbol{\theta}^{(k)}}[X_i, Y_i \mid \boldsymbol{v}] = \mathbb{E}_{\boldsymbol{\theta}^{(k)}}[X_iY_i \mid \boldsymbol{v}] - \mathbb{E}_{\boldsymbol{\theta}^{(k)}}[X_i \mid \boldsymbol{v}]\ E_{\boldsymbol{\theta}^{(k)}}[Y_i \mid \boldsymbol{v}]$

(iv) jev $[X_i \mid Y_i = y]$ má rozdělení $\mathrm{N}\left[\mu_1 + \frac{\sigma_{12}}{\sigma_{22}}(y-\mu_2), \left(\sigma_{11} - \frac{\sigma_{12}^2}{\sigma_{22}}\right)\right]$, jev $[Y_i \mid X_i = x]$ je obdobný s rozdělením $\mathrm{N}\left[\mu_2 + \frac{\sigma_{12}}{\sigma_{11}}(x-\mu_1), \left(\sigma_{22} - \frac{\sigma_{12}^2}{\sigma_{11}}\right)\right]$ (plyne z vlastností dvourozměrného normálního rozdělení - viz Anděl [2007, věta 4.12, str. 67])

Ve světle výše uvedených vlastností a (2.17) počítáme:

$$\mathbb{E}_{\boldsymbol{\theta}^{(k)}}\left[T_1 \mid \boldsymbol{v}\right] = \sum_{i=1}^{10} \mathbb{E}_{\boldsymbol{\theta}^{(k)}}\left[x_i^2 \mid \boldsymbol{v}\right] = \mathbb{E}_{\boldsymbol{\theta}^{(k)}}\left[\alpha^2 \mid y_7\right] + \mathbb{E}_{\boldsymbol{\theta}^{(k)}}\left[\beta^2 \mid y_8\right] + \sum_{\substack{i=1,\ldots,10 \\ i\neq 7,8}}^{10} x_i^2$$

$$\mathbb{E}_{\boldsymbol{\theta}^{(k)}}\left[\alpha^2 \mid y_7\right] \;=\; \mathrm{var}_{\boldsymbol{\theta}^{(k)}}\left[\alpha \mid y_7\right] \;+\; \left(\mathbb{E}_{\boldsymbol{\theta}^{(k)}}\left[\alpha \mid y_7\right]\right)^2 \;= \\ = \;\left(\sigma_{11} - \frac{\sigma_{12}^2}{\sigma_{22}}\right) \;+\; \left(\mu_1 + \frac{\sigma_{12}}{\sigma_{22}}(y_7 - \mu_2)\right)^2 \;. \quad (2.18)$$

Obdobným postupem získáme také $\mathbb{E}_{\boldsymbol{\theta}^{(k)}}\left[\beta^2 \mid y_8\right]$, ale je nutné si uvědomit, že u $\mathbb{E}_{\boldsymbol{\theta}^{(k)}}\left[\gamma^2 \mid x_9\right]$ a $\mathbb{E}_{\boldsymbol{\theta}^{(k)}}\left[\delta^2 \mid x_{10}\right]$ se jedná o druhou variantu podmíněného rozdělení z vlastnosti (iv). Odhady T_3 a T_4 se sestávají pouze z podmíněných středních hodnot α, β, γ a δ (viz vlastnost (iv)), a tedy posledním, který je třeba zjistit, je odhad T_5.

$$\mathbb{E}_{\boldsymbol{\theta}^{(k)}}\left[T_5 \mid \boldsymbol{v}\right] = \sum_{i=1}^{10} \mathbb{E}_{\boldsymbol{\theta}^{(k)}}\left[x_i y_i \mid \boldsymbol{v}\right] = \sum_{i=7}^{10} \mathbb{E}_{\boldsymbol{\theta}^{(k)}}\left[x_i y_i \mid \boldsymbol{v}\right] + \sum_{i=1}^{6} x_i y_i = \\ = y_7 \mathbb{E}_{\boldsymbol{\theta}^{(k)}}\left[\alpha \mid y_7\right] + y_8 \mathbb{E}_{\boldsymbol{\theta}^{(k)}}\left[\beta \mid y_8\right] + x_9 \mathbb{E}_{\boldsymbol{\theta}^{(k)}}\left[\gamma \mid x_9\right] + x_{10} \mathbb{E}_{\boldsymbol{\theta}^{(k)}}\left[\delta \mid x_{10}\right] + 896 \;.$$

Nyní vše vezmeme a sestavíme rovnice podmíněných středních hodnot T_i pro $i = 1, \ldots, 5$, k čemuž je zapotřebí doplnit číselné vstupy (vektor $\boldsymbol{v}$) a jisté elementární přepočty, čímž získáme:

$$\mathbb{E}_{\boldsymbol{\theta}^{(k)}}\left[T_1 \mid \boldsymbol{v}\right] = 1842 + 2\left(\sigma_{11}^{(k)} - \frac{(\sigma_{12}^{(k)})^2}{\sigma_{22}^{(k)}}\right) + 2(\mu_1^{(k)})^2 + \\ + 2\frac{\mu_1^{(k)}\sigma_{12}^{(k)}}{\sigma_{22}^{(k)}}(40 - 2\mu_2^{(k)}) + \left(\frac{\sigma_{12}^{(k)}}{\sigma_{22}^{(k)}}\right)^2 \left(2(\mu_2^{(k)})^2 - 80\mu_2^{(k)} + 808\right)$$

$$\mathbb{E}_{\boldsymbol{\theta}^{(k)}}\left[T_2 \mid \boldsymbol{v}\right] = 2001 + 2\left(\sigma_{22}^{(k)} - \frac{(\sigma_{12}^{(k)})^2}{\sigma_{11}^{(k)}}\right) + 2(\mu_2^{(k)})^2 + \\ + 2\frac{\mu_2^{(k)}\sigma_{12}^{(k)}}{\sigma_{11}^{(k)}}(45 - 2\mu_1^{(k)}) + \left(\frac{\sigma_{12}^{(k)}}{\sigma_{11}^{(k)}}\right)^2 \left(2(\mu_1^{(k)})^2 - 90\mu_1^{(k)} + 1025\right)$$

$$\mathbb{E}_{\boldsymbol{\theta}^{(k)}}\left[T_3 \mid \boldsymbol{v}\right] = 108 + 2\mu_1^{(k)} + \frac{\sigma_{12}^{(k)}}{\sigma_{22}^{(k)}}(40 - 2\mu_2^{(k)})$$

$$\mathbb{E}_{\boldsymbol{\theta}^{(k)}}\left[T_4 \mid \boldsymbol{v}\right] = 109 + 2\mu_2^{(k)} + \frac{\sigma_{12}^{(k)}}{\sigma_{11}^{(k)}}(45 - 2\mu_1^{(k)})$$

$$\mathbb{E}_{\boldsymbol{\theta}^{(k)}}\left[T_5 \mid \boldsymbol{v}\right] = 896 + 40\mu_1^{(k)} + 45\mu_2^{(k)} + \frac{\sigma_{12}^{(k)}}{\sigma_{22}^{(k)}}(808 - 40\mu_2^{(k)}) + \frac{\sigma_{12}^{(k)}}{\sigma_{11}^{(k)}}(1025 - 45\mu_1^{(k)}) \;.$$

Krok M: nyní využijeme definici (2.15), kde $t^{(k)}$ odpovídá podmíněným středním hodnotám vypočteným v kroku E. Pro tento výpočet ještě potřebujeme $\mathbb{E}_{\boldsymbol{\theta}}[t'(\boldsymbol{w})]$ a pro výpočet užijeme vlastnosti (i) - (iii) z kroku E. Dále uplatníme to, že $\boldsymbol{W}$ je náhodný výběr a že vektor $\boldsymbol{\theta}$ neudává v tomto případě nic jiného než samotné parametry rozdělení.

$$\mathbb{E}_{\boldsymbol{\theta}}[T_1] = \mathbb{E}_{\boldsymbol{\theta}}\left[\sum_{i=1}^{10} x_i^2\right] = 10\mathbb{E}_{\boldsymbol{\theta}}\left[x_1^2\right] = 10\left(\sigma_{11} + \mu_1^2\right)$$

$$\mathbb{E}_{\boldsymbol{\theta}}[T_3] = \mathbb{E}_{\boldsymbol{\theta}}\left[\sum_{i=1}^{10} x_i\right] = 10\mathbb{E}_{\boldsymbol{\theta}}[x_1] = 10\mu_1$$

$$\mathbb{E}_{\boldsymbol{\theta}}[T_5] = \mathbb{E}_{\boldsymbol{\theta}}\left[\sum_{i=1}^{10} x_i y_i\right] = 10\mathbb{E}_{\boldsymbol{\theta}}[x_1 y_1] = 10\left(\sigma_{12} + \mu_1\mu_2\right)$$

Obdobným postupem získáme i $\mathbb{E}_{\boldsymbol{\theta}}[T_2] = 10\left(\sigma_{22} + \mu_2^2\right)$ a $\mathbb{E}_{\boldsymbol{\theta}}[T_4] = 10\mu_2$.

Podmíněné střední hodnoty $\mathbb{E}_{\boldsymbol{\theta}^{(k)}}[T_i \mid \boldsymbol{v}]$ si pro přehlednost označíme jako $T_i^{(k)}$ $(i = 1, \ldots, 5)$. Pak z (2.15) plyne, že odhad $\boldsymbol{\theta}^{(k+1)}$ je ve tvaru:

$$\mu_1^{(k+1)} = \frac{1}{10}T_3^{(k)} \qquad \mu_2^{(k+1)} = \frac{1}{10}T_4^{(k)}$$

$$\sigma_{11}^{(k+1)} = \frac{1}{10}T_1^{(k)} - \left(\mu_1^{(k+1)}\right)^2 = \frac{1}{10}T_1^{(k)} - \left(\frac{1}{10}T_3^{(k)}\right)^2$$

$$\sigma_{22}^{(k+1)} = \frac{1}{10}T_2^{(k)} - \left(\mu_2^{(k+1)}\right)^2 = \frac{1}{10}T_2^{(k)} - \left(\frac{1}{10}T_4^{(k)}\right)^2$$

$$\sigma_{12}^{(k+1)} = \frac{1}{10}T_5^{(k)} - \mu_1^{(k+1)}\mu_2^{(k+1)} = \frac{1}{10}T_5^{(k)} - \frac{1}{100}T_3^{(k)}T_4^{(k)} \ .$$

Tabulka 2.3: Průběh EM algoritmu pro dvourozměrný normální případ

iterace	μ_1	μ_2	σ_{11}	σ_{22}	σ_{12}
0	10,00000	10,00000	20,00000	20,00000	10,00000
1	13,80000	15,15000	41,96000	26,70250	14,68000
2	14,09327	15,53875	46,82756	29,31608	19,22261
3	14,20370	15,69794	47,56940	30,29210	21,19651
4	14,24280	15,77894	47,54621	30,77354	22,11265
5	14,25518	15,82383	47,42563	31,05691	22,57445
6	14,25814	15,84967	47,32518	31,23294	22,82007
7	14,25811	15,86477	47,25637	31,34258	22,95518
8	14,25735	15,87367	47,21224	31,41019	23,03119
9	14,25659	15,87892	47,18478	31,45143	23,07464
10	14,25601	15,88204	47,16798	31,47638	23,09976
11	14,25562	15,88388	47,15781	31,49138	23,11439
12	14,25536	15,88497	47,15170	31,50036	23,12297
13	14,25520	15,88562	47,14803	31,50572	23,12802
14	14,25511	15,88601	47,14585	31,50891	23,13099
15	14,25505	15,88623	47,14454	31,51081	23,13275
16	14,25501	15,88637	47,14377	31,51194	23,13380
17	14,25499	15,88645	47,14330	31,51262	23,13441
18	14,25498	15,88650	47,14303	31,51301	23,13478
19	14,25497	15,88653	47,14287	31,51325	23,13500
20	14,25496	15,88654	47,14277	31,51339	23,13512
21	14,25496	15,88655	47,14271	31,51348	23,13520
22	14,25496	15,88656	47,14268	31,51352	23,13525
23	14,25496	15,88656	47,14266	31,51355	23,13527
24	14,25496	15,88656	47,14265	31,51357	23,13529
25	14,25496	15,88656	47,14264	31,51358	23,13530
26	14,25496	15,88657	47,14263	31,51359	23,13530
27	14,25496	15,88657	47,14263	31,51359	23,13531
28	14,25496	15,88657	47,14263	31,51359	23,13531
29	14,25496	15,88657	47,14263	31,51360	23,13531
30	14,25496	15,88657	47,14263	31,51360	23,13531

Tabulka 2.3 ukazuje, že konvergence je pro tento příklad trochu pomalejšího rázu než pro Příklad 1. Dále je patrné, že při vícerozměrném $\boldsymbol{\theta}$ dochází k rozdílně rychlému postupu pro jednotlivé složky, při výpočtu na pět desetinných míst jsou například u μ_1 přírůstky po kroku 20 již zanedbatelné, avšak σ_{22} dosáhne podobné úrovně až při kroku 29. V tabulce 2.3 jsou tyto hodnoty podtrženy u všech parametrů.

2.4.2 Přímý výpočet metodou maximální věrohodnosti

Nyní spočteme hodnoty parametrů, ke kterým bychom došli metodou maximální věrohodnosti, pokud bychom vynechali z Tabulky 2.2 pozorování s chybějícími složkami. Jinými slovy pracujeme s vektorem $\boldsymbol{v}^* = (\boldsymbol{\omega}_1^T, \ldots, \boldsymbol{\omega}_6^T)^T$. Hustota dvourozměrného normálního rozdělení je ve tvaru (2.16). Vynecháme členy neobsahující $\boldsymbol{\theta}$ a sestavíme logaritmickou věrohodnostní funkci:

$$L(\boldsymbol{v}^* \mid \boldsymbol{\theta}) = -\frac{n}{2}\log(\det \boldsymbol{V}) \; - \; \frac{\sigma_{11}\sigma_{22}}{2\det \boldsymbol{V}} \; \times \\ \times \; \sum_{i=1}^{n}\left[\frac{(x_i-\mu_1)^2}{\sigma_{11}} - \frac{2\sigma_{12}}{\sigma_{11}\sigma_{22}}(x_i-\mu_1)(y_i-\mu_2) + \frac{(y_i-\mu_2)^2}{\sigma_{22}}\right] \; . \quad (2.19)$$

Dle známých vlastností normálního rozdělení jsou maximálně věrohodné odhady středních hodnot a rozptylů ve tvaru:

$$\hat{\mu_1} = \frac{1}{n}\sum_{i=1}^{n} X_i = 10,5 \qquad \hat{\mu_2} = \frac{1}{n}\sum_{i=1}^{n} Y_i = 13,16667 \quad (2.20)$$

$$\hat{\sigma_{11}} = \frac{1}{n}\sum_{i=1}^{n}(X_i-\hat{\mu_1})^2 = 25,91667 \qquad \hat{\sigma_{22}} = \frac{1}{n}\sum_{i=1}^{n}(Y_i-\hat{\mu_2})^2 = 25,47222 \quad (2.21)$$

Nyní spočteme maximálně věrohodný odhad σ_{12}, tedy převedeme (2.19) na tvar:

$$L(\boldsymbol{v}^* \mid \boldsymbol{\theta}) = -\frac{n}{2}\log(\det \boldsymbol{V}) \; - \; \frac{\sigma_{22}}{2\det \boldsymbol{V}}\sum_{i=1}^{n}(x_i-\mu_1)^2 \; + \\ + \; \frac{\sigma_{12}}{\det \boldsymbol{V}}\sum_{i=1}^{n}(x_i-\mu_1)(y_i-\mu_2) \; - \; \frac{\sigma_{11}}{2\det \boldsymbol{V}}\sum_{i=1}^{n}(y_i-\mu_2)^2 \; .$$

Poté dosadíme $\hat{\mu_1}$ a $\hat{\mu_2}$ a obecné vyjádření rozptylů z (2.21), čímž v důsledku získáme funkci jediné proměnné σ_{12}:

$$L(\boldsymbol{v}^* \mid \boldsymbol{\theta}) = -\frac{n}{2}\log(\det \boldsymbol{V}) \; - \; \frac{1}{n\det \boldsymbol{V}}\sum_{i=1}^{n}[(x_i-\hat{\mu_1})^2(y_i-\hat{\mu_2})^2] \; + \\ + \; \frac{\sigma_{12}}{\det \boldsymbol{V}}\sum_{i=1}^{n}(x_i-\hat{\mu_1})(y_i-\hat{\mu_2}) \; .$$

Než přistoupíme k derivaci dle σ_{12}, je vhodné zdůraznit, že $\det \boldsymbol{V}$ je v tomto případě funkcí proměnné σ_{12}.

$$\frac{\partial}{\partial\sigma_{12}}L(\boldsymbol{v}^* \mid \boldsymbol{\theta}) = \frac{n\sigma_{12}}{2\det \boldsymbol{V}} - \frac{2\sigma_{12}}{n(\det \boldsymbol{V})^2}A + \frac{\det \boldsymbol{V} + 2\sigma_{12}}{(\det V)^2}B \; ,$$

kde $A = \sum_{i=1}^{n}[(x_i-\hat{\mu_1})(y_i-\hat{\mu_2})]$ a $B = \sum_{i=1}^{n}[(x_i-\hat{\mu_1})^2]\sum_{i=1}^{n}[(y_i-\hat{\mu_2})^2]$.

Tuto funkci položíme rovnu nule, dosadíme $\det \boldsymbol{V} = \hat{\sigma_{11}}\hat{\sigma_{22}} - \sigma_{12}^2$ a elementárními úpravami dosáhneme rovnice třetího stupně.

$$-n\sigma_{12}^3 + B\sigma_{12}^2 + (n\hat{\sigma_{11}}\hat{\sigma_{22}} - \frac{1}{3}A)\sigma_{12} + \hat{\sigma_{11}}\hat{\sigma_{22}}B = 0$$

Po dosazení a vyčíslení vypadá rovnice následovně:

$$\sigma_{12}^3 - 361,303\sigma_{12}^2 + 606,155\sigma_{12} - 238516,174 = 0 \ .$$

Tato rovnice má tři kořeny: $27,7488$; $-24,03740$ a $357,592$. Z druhé derivace vypsané níže plyne, že $27,7488$ je jediným maximem. Zároveň je k našemu odhadu z Tabulky 2.3 také nejbližší.

$$\frac{\partial^2}{(\partial\sigma_{12})^2} L(\boldsymbol{v}^* \mid \boldsymbol{\theta}) = 3\sigma_{12}^2 - 722,606\sigma_{12} + 606,155$$

2.4.3 Srovnání výsledků

Tabulka 2.4 porovnává hodnoty získané pomocí EM algoritmu a hodnoty vypočtené přímým způsobem jen z úplných pozorování. U středních hodnot je chyba menší, zato u σ_{11} je dosti podstatná. Je tedy patrné, že vynecháme-li při výpočtu maximálně věrohodného odhadu neúplná pozorování, dopouštíme se tím velmi závažné chyby.

Tabulka 2.4: Srovnání hodnot parametrů získaných různými postupy

	EM algoritmus	**přímý výpočet**	**rozdíl**
μ_1	14,25496	10,50000	3,75496
μ_2	15,88657	13,16667	2,71990
σ_{11}	47,14263	25,91667	21,22596
σ_{22}	31,51360	25,47222	6,04138
σ_{12}	23,13531	27,74880	4,61349

2.5 Zobecněný EM algoritmus

Výše uvedený popis algoritmu je značně obecný a platí tedy pro velmi širokou škálu problémů, avšak jeho základním nedostatkem je předpoklad, že se výsledek kroku M dá vyjádřit pomocí omezeného počtu elementárních funkcí se základními aritmetickými operacemi (orig. closed-form expression). Obecně to však předpokládat nelze, čímž bychom se při maximalizaci v kroku M mohli dostat do značných potíží. Tedy Dempster a kol. [1977, str. 7] pro dosažení maximální obecnosti zavedli tzv. Zobecněný EM algoritmus (orig. Generalized EM algorithm - dále již jen GEM).

McLachlan a Krishnan [1997, str. 28] uvádí, že se od EM algoritmu liší podstatou maximalizačního kroku - zde hledáme $\boldsymbol{\theta}^{(k+1)}$ navyšující hodnotu funkce $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})$ nad její hodnotu v bodě $\boldsymbol{\theta} = \boldsymbol{\theta}^{(k)}$ místo maximalizace přes všechna $\boldsymbol{\theta} \in \Omega$. Jinými slovy hledáme takové $\boldsymbol{\theta}^{(k+1)}$, aby platilo:

$$Q(\boldsymbol{\theta}^{(k+1)} \mid \boldsymbol{\theta}^{(k)}) \geq Q(\boldsymbol{\theta}^{(k)} \mid \boldsymbol{\theta}^{(k)}) \ .$$

Tato podmínka je dostatečná pro zajištění toho, aby posloupnost $\{L(\boldsymbol{\theta}^{(k)})\}$ byla neklesající a tedy (za předpokladu, že je omezená shora) jsme měli zajištěnou konvergenci.

Kapitola 3

Vlastnosti

Pro začátek si připomeňme základní pojmy. Funkce $f(\boldsymbol{x} \mid \boldsymbol{\theta})$ je hustotou vektoru úplných dat $\boldsymbol{x}$ a funkce $g(\boldsymbol{y} \mid \boldsymbol{\theta})$ vektoru dat neúplných, tedy $\boldsymbol{y}$. Obě dvě jsou závislé na parametru $\boldsymbol{\theta} \in \Omega$, kde Ω je parametrický prostor. Funkce Q je základní funkce EM algoritmu, konkrétně kroku E, a je definována v (2.11). $L(\boldsymbol{\theta})$ je logaritmická věrohodnostní funkce a $\{\boldsymbol{\theta}^{(k)}\}$ je posloupnost odhadů získaná EM algoritmem.

3.1 Monotonie

Mějme zobrazení G a M, která zobrazují $\boldsymbol{\theta} \in \Omega$ na podmnožiny Ω, definovaná následovně:

$$G(\boldsymbol{\theta}) = \{\boldsymbol{\psi} \mid Q(\boldsymbol{\psi} \mid \boldsymbol{\theta}) \geq Q(\boldsymbol{\theta} \mid \boldsymbol{\theta})\} \tag{3.1}$$

$$M(\boldsymbol{\theta}) = \left\{\boldsymbol{\phi} \mid Q(\boldsymbol{\phi} \mid \boldsymbol{\theta}) = \max_{\hat{\boldsymbol{\theta}} \in \Omega} Q(\hat{\boldsymbol{\theta}} \mid \boldsymbol{\theta})\right\} . \tag{3.2}$$

Máme-li iterativní algoritmus, pro jehož prvky platí $\boldsymbol{\theta}^{(k+1)} \in G(\boldsymbol{\theta}^{(k)})$, pak hovoříme o GEM algoritmu. Je-li však $\boldsymbol{\theta}^{(k+1)} \in M(\boldsymbol{\theta}^{(k)})$, pak se jedná o EM algoritmus. Obecně nemůžeme říci nic o mohutnosti těchto množin, avšak v obou našich příkladech (pracujeme se zobrazením M), je množina $M(\boldsymbol{\theta}^{(k)})$ jednoprvková pro všechna k, neboť zobrazení M převádí body na body. Třeba pro Příklad 1 platí: $M(\theta^{(k+1)}) = \frac{4560\ \theta^{(k)}+3520}{5480\ \theta^{(k)}+5360}$ (viz (2.9)).

Dále je výhodné si uvědomit, že $M \subset G$, což je patrné nejen z definic těchto množin, ale zároveň i ze skutečnosti, že GEM algoritmus je zobecněním EM. Tedy jsou-li věty v této kapitole definovány pro GEM, pak převedení na využití pro EM je triviální záležitostí, ale v opačném směru tomu tak být nemusí (resp. je-li to vůbec možné).

Označíme si podmíněnou hustotu $\boldsymbol{x}$ ku $\boldsymbol{y}$ a $\boldsymbol{\theta}$ jako $k(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\theta}) = \frac{f(\boldsymbol{x}|\boldsymbol{\theta})}{g(\boldsymbol{y}|\boldsymbol{\theta})}$. Pak tedy pro monotonii $\{L(\boldsymbol{\theta}^{(k)})\}$ platí následující tvrzení (pro důkaz viz Dempster a kol. [1977, str. 7]):

Věta 1. *Pro každý GEM algoritmus platí* $L(G(\boldsymbol{\theta})) \geq L(\boldsymbol{\theta})$ *pro všechna* $\boldsymbol{\theta} \in \Omega$ *a že rovnost nastává pouze pokud platí (skoro všude):*

$$Q(G(\boldsymbol{\theta}) \mid \boldsymbol{\theta}) = Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}) \text{ a } k(\boldsymbol{x} \mid \boldsymbol{y}, G(\boldsymbol{\theta})) = k(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\theta}) \ .$$

Nyní ověříme platnost tvrzení pro Příklad 1. Funkci $L(\theta)$ zde odpovídá (2.3) a $M(\theta)$ je v souladu s definicí (2.9). Dosadíme-li číselné údaje, získáme rovnici:

$$L(\theta) = \log C(\boldsymbol{y}) + 1760 \log \theta + 920 \log(1 - \theta) + 2800 \log(2 + \theta)$$

Pro vektor neúplných dat $\boldsymbol{y} = (1700, 2800, 450, 470, 60)^T$ nabývá výraz $\log C(\boldsymbol{y})$ hodnoty přibližně $6409,413269031$. Jelikož platí $M(\theta^{(k)}) = \theta^{(k+1)}$, pak doplněním Tabulky 2.1 získáme Tabulku 3.1. Z ní je patrné, že pro tento příklad Věta 1 bezpodmínečně platí.

Co se Příkladu 2 týče, tak logaritmickou věrohodnost získáme obvyklým postupem z (2.12). Tato funkce pro vektor neúplných dat $\boldsymbol{y}$ je ve tvaru:

$$L(\boldsymbol{\theta}) = c(\boldsymbol{\theta})^T t(\boldsymbol{y}) - \log a(\boldsymbol{\theta}) + \log b(\boldsymbol{x}) \ .$$

Nyní, stejně jako pro ověření předchozího příkladu, vyjdeme z Tabulky 2.3 a získáme Tabulku 3.2. I zde platí $M(\boldsymbol{\theta}^{(k)}) = \boldsymbol{\theta}^{(k+1)}$.

Pro přehlednost tabulek uvádíme ve třetím sloupci veličinu:

$$\Delta_k = L(\boldsymbol{\theta}^{(k+1)}) - L(\boldsymbol{\theta}^{(k)}) \ ,$$

kterou jsme získali z $L(M(\boldsymbol{\theta}^{(k)})) \geq L(\boldsymbol{\theta}^{(k)})$ (jelikož platí $M(\boldsymbol{\theta}^{(k)}) = \boldsymbol{\theta}^{(k+1)}$). V tabulkách zkoumáme, jestli je $\Delta_k > 0$ pro všechna k a jestli je posloupnost $\{\Delta_k\}$ klesající.

Tabulka 3.1: Vývoj hodnot věrohodnostní funkce pro Příklad 1

iterace	$L(\theta^{(k)})$	Δ_k
0	7117,3928743781	343,6305340512
1	7461,0234084293	1,9757345347
2	7462,9991429640	0,0071051705
3	7463,0062481344	0,0000245288
4	7463,0062726633	0,0000000845
5	7463,0062727477	0,0000000003
6	7463,0062727480	0,0000000000
7	7463,0062727480	0,0000000000

Tabulka 3.2: Vývoj hodnot věrohodnostní funkce pro Příklad 2

iterace	$L(\boldsymbol{\theta}^{(k)})$	Δ_k
0	-64,60658	10,16634
1	-54,44024	0,14144
2	-54,29880	0,03149
3	-54,26730	0,00967
4	-54,25764	0,00310
5	-54,25454	0,00102
6	-54,25352	0,00035
7	-54,25317	0,00012
8	-54,25305	0,00004
9	-54,25301	0,00001
10	-54,25300	0,00001
11	-54,25299	0,00000
12	-54,25299	0,00000

3.2 Konvergence $\{L(\boldsymbol{\theta}^{(k)})\}$

V minulém oddíle bylo pojednáváno o konvergenci posloupnosti $\{L(\boldsymbol{\theta}^{(k)})\}$ a plyne z něj, že pokud je tato posloupnost shora omezená, tak konverguje k nějakému L^*. Nyní zjistitíme, za jakých podmínek je bod L^* maximem funkce $L(\boldsymbol{\theta})$ (resp. stacionárním bodem).

Nadále budeme předpokládat následující podmínky (viz Wu [1983, str. 96-97]):

1. Ω je podmnožinou n-rozměrného Euklidovského prostoru R^n
2. $\Omega_{\boldsymbol{\theta}^0} = \{\boldsymbol{\theta} \in \Omega : L(\boldsymbol{\theta}) \geq L(\boldsymbol{\theta}^0)\}$ je kompaktní $\forall \boldsymbol{\theta}^0 \in \Omega$ taková, že $L(\boldsymbol{\theta}^0) > -\infty$
3. L je spojitá na celém Ω a diferencovatelná na $\operatorname{int}\Omega$

Tyto podmínky předpokládáme proto, abychom měli pro posloupnost $\{L(\boldsymbol{\theta}^{(k)})\}$ zajištěnou omezenost shora, což spolu s poznatky z předchozího oddílu dává monotónní konvergenci $\{L(\boldsymbol{\theta}^{(k)})\}$ k L^*. Existují však i případy, kdy je druhá podmínka, tedy předpoklad kompaktnosti $\Omega_{\boldsymbol{\theta}^0}$, značně problematická, neboť neexistuje žádný způsob zkompaktnění tohoto parametrického prostoru.

Náš druhý příklad ($\Omega = \mathbb{R} \times \mathbb{R} \times (0,\infty) \times (0,\infty) \times \mathbb{R}$) je toho důkazem, neboť blížíme-li se středními hodnotami (μ_1, μ_2) k nějaké reálné hodnotě a zároveň rozptyly $(\sigma_{11}, \sigma_{22})$ směřují k 0, pak se hodnota věrohodnostní funkce blíží nekonečnu. Jinými slovy máme značné problémy při volbě $\boldsymbol{\theta}^0$ na hranici parametrického prostoru. McLachlan a Krishnan [1997, str. 87] však podotýkají, že prostor $\Omega_{\boldsymbol{\theta}^0}$ můžeme zkompaktnit, zavedeme-li pro Ω omezení $\sigma_{ii} \geq \varepsilon$ pro nějaké malé $\varepsilon > 0$ a $i = 1, 2$. Je však důležité si uvědomit, že pokud bychom se s odhady $\boldsymbol{\theta}^{(k)}$ dostávali blízko k hranici Ω, bylo by pro zkompaktnění třeba využít silnějších prostředků. To však vzhledem k hodnotám v Tabulce 2.3 není nutné.

Abychom v těchto případech předešli problémům, budeme dále předpokládat, že pro každou posloupnost $\{\boldsymbol{\theta}^{(k)}\}$, pro jejíž počáteční hodnotu $\boldsymbol{\theta}^0$ platí $L(\boldsymbol{\theta}^0) > -\infty$, leží $\boldsymbol{\theta}^{(k)}$ v $\operatorname{int}\Omega$ pro všechna k. Toto lze jednoduše zajistit pomocí následující podmínky:

$$\Omega_{\boldsymbol{\theta}^0} \subset \operatorname{int}\Omega \textit{ pro libovolné } \boldsymbol{\theta}^0 \in \Omega \text{ (viz Wu [1983, str. 97])}.$$

Touto jednoduchou změnou podmínky 2 předejdeme problémům hlavně při derivování v kroku M.

Jak již bylo dříve řečeno, množina M (viz (3.2)) nemusí být jednoprvková, tedy pro další postup zavedeme následující definici:

Řekněme, že zobrazení A zobrazující body z X na podmnožiny X je uzavřené v $x \in X$, pokud $x_k \longrightarrow x, x_k \in X$ a $y_k \longrightarrow y, y_k \in A(x_k)$ implikuje $y \in A(x)$. (3.3)

V obou našich příkladech ale M zobrazuje body na body, tedy uzavřenost plyne přímo ze spojitosti tohoto zobrazení.

Pro potřeby tvrzení uvedených v této kapitole je vhodné upřesnit, že je-li řeč o posloupnosti $\{\boldsymbol{\theta}^{(k)}\}$ jako o posloupnosti získané pomocí GEM (respektive EM) algoritmu, pak je tímto myšleno, že platí $\boldsymbol{\theta}^{(k+1)} \in G(\boldsymbol{\theta}^{(k)})$ (respektive $M(\boldsymbol{\theta}^{(k)})$) pro všechna k. Zobrazení G a M jsou definovány v (3.1) a (3.2).

Věty 2 a 3 uvedl Wu [1983, str. 97-98], jsou odvozeny z globální věty o konvergenci a směřují k aplikaci pro posloupnost $\{L(\boldsymbol{\theta}^{(k)})\}$ s prvky $\theta^{(k)}$ získanými EM nebo GEM algoritmem. Pro následující věty předpokládáme tři podmínky uvedené v začátku tohoto oddílu. Důkazy těchto vět uvádí Wu [1983, str. 98] nebo McLachlan a Krishnan [1997, str. 87-88].

Věta 2. *Nechť $\{\boldsymbol{\theta}^{(k)}\}$ je posloupnost získaná pomocí GEM algoritmu. Množinu všech stacionárních bodů (resp. lokálních maxim) funkce L na* $\operatorname{int}\Omega$ *označíme jako $\mathcal{S}$ (resp. $\mathcal{M}$). Dále předpokládejme:*

$$M \text{ je uzavřené zobrazení zobrazující body do podmnožin } \Omega \setminus \mathcal{S}(\text{resp. } \mathcal{M}); \tag{3.4}$$

$$L(\boldsymbol{\theta}^{(k+1)}) > L(\boldsymbol{\theta}^{(k)}) \quad \forall \boldsymbol{\theta}^{(k)} \notin \mathcal{S} \ (\text{resp. } \mathcal{M}) \ . \tag{3.5}$$

Pak všechny limitní body $\{\boldsymbol{\theta}^{(k)}\}$ jsou stacionárními body (resp. lokálními maximy) L a $\{L(\boldsymbol{\theta}^{(k)})\}$ konverguje monotónně k $L^* = L(\boldsymbol{\theta}^*)$ pro nějaké $\boldsymbol{\theta}^* \in \mathcal{S}$ (resp. $\mathcal{M}$).

Jinými slovy nám tato věta, platí-li všechny předpoklady, značně usnadňuje výpočet kroku M, neboť nemá smysl hledat maximum věrohodnostní funkce v jiných bodech než těch stacionárních (respektive v lokálních maximech). Takový bod je v našich příkladech vždy právě jeden. O této větě řekli, že je "hlavní větou o konvergenci posloupností odhadů GEM algoritmu" (citace, McLachlan a Krishnan [1997, str. 88]) nebo, že "je nejobecnější výsledek pro EM a GEM algoritmy." (citace, Wu [1983, str. 99]).

Wu [1983, str. 98] pro zajištění uzavřenosti zobrazení M pro EM algoritmus definuje velmi jednoduchou postačující podmínku, a to:

$$\text{nechť je } Q(\boldsymbol{\psi} \mid \boldsymbol{\theta}) \text{ spojitá v obou proměnných.} \tag{3.6}$$

Tato podmínka je značně slabá, a tedy je v mnoha případech splněna. Pro nás je nejdůležitější to, že mezi ně patří problémy zahrnující regulární exponenciální rodinu rozdělení. Věta 3 je přímým důsledkem věty předchozí, kde je podmínka (3.4) splněna díky (3.6), a (3.5) plyne přímo z definice EM algoritmu.

Věta 3. *Nechť Q splňuje podmínku* (3.6)*. Pak limitní body všech posloupností prvků $\{\boldsymbol{\theta}^{(k)}\}$ získaných EM algoritmem jsou stacionární body L a $\{L(\boldsymbol{\theta}^{(k)})\}$ konverguje monotónně k $L^* = L(\boldsymbol{\theta}^*)$ pro nějaký stacionární bod $\boldsymbol{\theta}^* \in \Omega$.*

V Příkladu 1 platí $\mathcal{S} = \mathcal{M} = \{\theta^*\}$, protože rovnice pro získání maximálně věrohodného odhadu θ ve tvaru

$$\frac{1760}{\theta} - \frac{920}{1-\theta} + \frac{2800}{2+\theta} = 0$$

má právě jeden přípustný kořen (druhý je záporný, ale θ jsme již v zadání předpokládali mezi 0 a 1). Dále pak platnost (3.4) plyne přímo z definice uzavřeného zobrazení (viz (3.3)), v níž stačí označit $x_k = \theta^{(k)}$, $x = \theta^*$, $y_k = L(\theta^{(k)})$, $y = L(\theta^*)$. Podmínka (3.5) je splněna pro všechna $\theta \neq \theta^*$ (vyžaduje sice přepočet Tabulky 2 na mnoho desetinných míst, ale platí).

Aplikace Věty 2 na Příklad 2 probíhá ve stejném duchu. Rovnice (2.15) má rovněž právě jeden kořen, tedy i zde platí $\mathcal{S} = \mathcal{M} = \{\theta^*\}$. Pro (3.4) máme zobrazení definované stejně jako v předchozím odstavci a uzavřené je rovněž, neboť zobrazuje body na body a je spojité. Podmínka (3.5) rovněž platí a lze ji ověřit obdobně jako pro Příklad 1.

Jelikož rozdělení užitá v obou příkladech této práce jsou prvky regulární exponenciální rodiny, pak je pro ně Věta 3 přímým důsledkem Věty 2 dle úvah plynoucích z (3.6) před Větou 3.

3.3 Konvergence $\{\boldsymbol{\theta}^{(k)}\}$

V minulém oddíle bylo pojednáno o podmínkách konvergence hodnot věrohodnostní funkce $\{L(\boldsymbol{\theta}^{(k)})\}$ k nějakému L^* (ať už lokálnímu maximu či stacionárnímu bodu), což však obecně neimplikuje konvergenci korespondující posloupnosti iteračních členů $\{\boldsymbol{\theta}^{(k)}\}$ k bodu $\boldsymbol{\theta}^*$ takovému, že $L^* = L(\boldsymbol{\theta}^*)$. Tato konvergence "není z numerického hlediska tak důležitá jako konvergence $\{L(\boldsymbol{\theta}^{(k)})\}$ k nějakému lokálnímu maximu či stacionárnímu bodu" (citace, Wu [1983, str. 99]). Je však vhodné o ní pro úplnost pojednat (důkazy uvedených tvrzení viz Wu [1983, str. 99-101] nebo McLachlan a Krishnan [1997, str. 89-91]).

Pro následující věty si označíme $S(a) = \{\boldsymbol{\theta} \in \mathcal{S} : L(\boldsymbol{\theta}) = a\}$ (jedná se o podmnožinu $\mathcal{S}$, tedy množiny stacionárních bodů $\operatorname{int}\Omega$). Dále předpokládáme splnění podmínek (3.4) a (3.5) (viz předchozí oddíl).

Věta 4. *Nechť $\{\boldsymbol{\theta}^{(k)}\}$ je posloupností získanou pomocí GEM algoritmu, která splňuje podmínky* (3.4) *a* (3.5)*. Dále buď $S(L^*) = \{\boldsymbol{\theta}^*\}$ jednoprvková množina, kde L^* je limita $\{L(\boldsymbol{\theta}^{(k)})\}$. Pak $\boldsymbol{\theta}^{(k)} \to \boldsymbol{\theta}^*$.*

V obou našich příkladech jsou předpoklady (3.4) a (3.5) pro posloupnost $\{\boldsymbol{\theta}^{(k)}\}$ splněny přímo z definice EM algoritmu (dle úvah v závěru předchozího oddílu). Dále z vlastností logaritmické věrohodnostní funkce exponenciální rodiny víme, že $S(L^*)$ je jednoprvková. Což znamená (zjednodušeně řečeno), že posloupnost iteračních členů nemá jinou možnost, než směřovat k oné hodnotě $\boldsymbol{\theta}^*$, o níž víme, že maximalizuje věrohodnostní funkci.

Wu [1983, str. 99] dále uvádí, že předpoklad $S(L^*) = \{\boldsymbol{\theta}^*\}$ lze značně vylepšit, pokud předpokládáme následující jako nutnou podmínku pro $\boldsymbol{\theta}^{(k)} \to \boldsymbol{\theta}^*$:

$$\| \boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^{(k)} \| \to 0 \quad \text{pro } k \to \infty. \tag{3.7}$$

Věta 5. *Nechť $\{\boldsymbol{\theta}^{(k)}\}$ je posloupností získanou pomocí GEM algoritmu, která splňuje podmínky* (3.4) *a* (3.5)*. Splňuje-li tato posloupnost rovněž podmínku* (3.7)*, pak se všechny limitní body $\{\boldsymbol{\theta}^{(k)}\}$ nacházejí v souvislé a kompaktní podmnožině $S(L^*)$. Speciálně, je-li $S(L^*)$ diskrétní, pak $\{\boldsymbol{\theta}^{(k)}\}$ konverguje k nějakému $\boldsymbol{\theta}^* \in S(L^*)$.*

Uvedené podmínky zajišťují konvergenci posloupnosti iteračních členů a navíc o mohutnosti množiny $S(L^*)$ nic nepředpokládáme, což značně přidává na užitečnosti. Jinými slovy Věta 5 říká, že pokud jsou splněny všechny předpoklady, tak $\boldsymbol{\theta}^{(k)}$ s rostoucím k směřuje k takové hodnotě $\boldsymbol{\theta}^*$, jejíž funkční hodnota je rovná oné L^* (která je buď lokální maximum nebo stacionární bod). Tedy najdeme-li hodnotu parametru, v němž má L stacionární bod (resp. lokální maximum), pak naše posloupnost $\boldsymbol{\theta}^{(k)}$ k této hodnotě konverguje. V našich příkladech jsou předpoklady splněny z definice EM algoritmu a díky skutečnosti, že M zobrazuje body na body.

Věta 6. *Nechť* $\{\boldsymbol{\theta}^{(k)}\}$ *je posloupností získanou pomocí GEM algoritmu, která navíc splňuje podmínku :*

$$\left[\frac{\partial}{\partial\boldsymbol{\theta}}Q(\theta \mid \boldsymbol{\theta}^{(k)})\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(k+1)}} = 0 \ . \tag{3.8}$$

Dále předpokládejme, že $\frac{\partial}{\partial\boldsymbol{\theta}}Q(\boldsymbol{\psi} \mid \boldsymbol{\theta})$ *je spojitá v obou proměnných.*

Pak $\boldsymbol{\theta}^{(k)}$ *konverguje ke stacionárnímu bodu* $\boldsymbol{\theta}^*$*, pro který platí* $L^* = L(\boldsymbol{\theta}^*)$ *kde* L^* *je limitou* $\{L(\boldsymbol{\theta}^{(k)})\}$ *právě tehdy, když je splněna jedna z následujících podmínek (pro* $\mathcal{P}(L^*) = \{\theta \in \Omega : L(\boldsymbol{\theta}) = L^*\}$*):*

1) $\mathcal{P}(L^*) = \{\boldsymbol{\theta}^*\}$

2) platí podmínka (3.7) *a* $\mathcal{P}(L^*)$ *je diskrétní.*

Tato věta je velmi důležitá z několika důvodů:

- podmínka (3.8) je za předpokládaných podmínek regularity (tedy podmínek (i)-(iii) ze začátku oddílu 3.2) automaticky splněna pro libovolnou posloupnost získanou EM algoritmem
- $\mathcal{S}(L)$ je podmnožinou $\mathcal{P}(L)$
- je nezávislá na obtížněji ověřitelných předpokladech (3.4) a (3.5).

McLachlan a Krishnan [1997, str. 90] konstatují, že vzhledem k těmto důvodům je Věta 6 mnohem silnějším (a lépe aplikovatelnějším) nástrojem pro ověřování konvergence $\{\boldsymbol{\theta}^{(k)}\}$ než Věty 4 a 5. Velmi praktický je rovněž níže uvedený důsledek této věty, avšak s ohledem na důležitost Věty 6 si ji nejprve přiblížíme pro naše příklady.

V obou našich příkladech je podmínka (3.8) pro posloupnost $\{\boldsymbol{\theta}^{(k)}\}$ splněna automaticky, neboť je podstatou kroku M. Spojitost Q v obou proměnných je rovněž splněna, tentokrát z vlastností logaritmické věrohodnosti pro rozdělení z exponenciální rodiny. Z přímých výpočtů maximálně věrohodných odhadů v našich příkladech plyne, že logaritmická věrohodnostní funkce má právě jedno maximum. Jeho konkrétní hodnota je již jen otázkou využitých dat a zvoleného postupu výpočtu. Tedy $\boldsymbol{\theta}^*$ splňující $L^* = L(\boldsymbol{\theta}^*)$ existuje právě jedno, a proto platí výsledek **1** Věty 6.

Důsledek Věty 6 se zabývá kategorií unimodálních funkcí, což jsou funkce, které mají právě jeden modus, jinými slovy hodnotu $\boldsymbol{x_0}$ takovou, že platí $f(\boldsymbol{x_0}) \geq f(\boldsymbol{x})$ pro všechna $\boldsymbol{x}$ (viz Anděl [2007, str. 17]).

Důsledek 1. *Nechť* $L(\boldsymbol{\theta})$ *je unimodální funkce na* Ω *a buď* $\boldsymbol{\theta}^*$ *její jediný stacionární bod. Dále nechť je* $\frac{\partial}{\partial\boldsymbol{\psi}}Q(\boldsymbol{\psi} \mid \boldsymbol{\theta})$ *spojitá v obou proměnných. Pak každá posloupnost prvků* $\{\boldsymbol{\theta}^{(k)}\}$ *získaných EM algoritmem konverguje k maximálně věrohodnému odhadu parametru* $\boldsymbol{\theta}$*, který je jediný - tedy k* $\boldsymbol{\theta}^*$*.*

Aplikace důsledku na naše příklady plyne z úvah k aplikacím vět v tomto oddíle.

Kapitola 4

Závěr

EM algoritmus je obecně snadné naprogramovat a nemá velké nároky na výpočetní techniku (např. nedochází k práci s informační maticí). To je velmi důležité, neboť je často zapotřebí velkého počtu iterací, což je také jednou z nejčastěji kritizovaných vlastností tohoto algoritmu. Pomalá konvergence může být způsobena například velkým množstvím chybějících dat. Avšak předností má stále mnoho - například krok M lze velmi často naprogramovat za užití standardních statistických balíčků i v případech, kdy odhady metodami maximální věrohodnosti neexistují v jednodušší podobě. V jiných případech lze využít některá snadno aplikovatelná rozšíření algoritmu, která sdílejí s původním algoritmem vlastnost stabilní monotónní konvergence (viz níže). Monotónní růst věrohodnostní posloupnosti $\{L(\boldsymbol{\theta}^{(k)})\}$ je další velmi výhodnou vlastností, neboť umožňuje velmi snadno sledovat rychlost konvergence a průběžně hledat případné chyby v naprogramování (viz Tabulky 2.1 a 2.3).

Velmi praktickou vlastností je to, že předpoklady pro konvergenci EM algoritmu jsou značně obecné, tedy z libovolné počáteční hodnoty konverguje takřka vždy k lokálnímu maximu. Největší úskalí ale spočívá v tom, že tato konvergence velmi závisí na volbě výchozího bodu a chování věrohodnostní funkce. Jsou známy případy, kdy posloupnost $\{L(\boldsymbol{\theta}^{(k)})\}$ nejenže nekonverguje k lokálnímu maximu, nýbrž k stacionárnímu bodu, nebo dokonce i k lokálnímu minimu věrohodnostní funkce. Tyto příklady hlouběji rozvedli například McLachlan a Krishnan [1997, str. 91-99].

EM algoritmus se stal velmi rychle značně oblíbeným, a tak byl (ve snaze alespoň částečně řešit výše zmíněné problémy) formulován značný počet rozšíření. Za zmínku stojí obzvláště ECM a ECME. Obě dvě jsou postaveny na myšlence takzvané podmíněné maximalizace a zachovávají podmínky konvergence původního EM algoritmu. ECM formulovali Meng a Rubin [1993], je zkratkou pro Expectation-Conditional Maximization a staví na myšlence, že v některých případech lze krok M mnohem snadněji vypočítat, podmíníme-li ho nějakou funkcí parametrů místo parametrů samotných. Jinými slovy nahrazuje M krok posloupností kroků, kdy dochází k snazšímu, ale méně účinnému výpočtu. Což v praxi znamená, že je zapotřebí více kroků, ale výpočetní nároky jsou celkově menší, neboť se jedná o matematicky jednodušší postup.

ECME (Expectation-Conditional Maximization Either) je rozšířením ECM algoritmu a formulovali ho Liu a Rubin [1994]. Tento algoritmus nahrazuje některé kroky ECM algoritmu přímou maximalizací logaritmické věrohodnosti vektoru neúplných dat (místo funkce Q) a jiné nechává ve stejné podobě jako ECM algoritmus. ECME postupuje v podstatně menším počtu kroků než EM a ECM algoritmy a navíc je prakticky vždy rychlejší.

Cílem této práce je přiblížit čtenáři, co je to EM algoritmus, na několika příkladech ukázat, jak tento algoritmus postupuje, nastínit jeho základní vlastnosti a vyzdvihnout jeho klady a zápory. Pro lepší přehlednost získaných informací je rovněž doplněna tabulkami k ilustračním příkladům.

Podoba vstupních dat o populaci tygrů nebyla zvolena náhodou, neboť vcelku nedávno skutečně došlo na základě testů DNA k zjištění, že poddruh tygr indočínský vlastně zahrnuje dva poddruhy. Ten nově objevený byl pojmenován malajský dle hlavní oblasti výskytu.

Práce je určena čtenářům, kteří již nějaké základní znalosti v oblasti pravděpodobnosti a statistiky mají, a bere si za cíl seznámit je s tímto praktickým nástrojem pro výpočet statistických problémů neúplných dat. Jelikož jsem při pátrání po zdrojích nenarazil na jiné než v anglickém jazyce, pak bych považoval za úspěch, kdyby tato práce posloužila k porozumění EM algoritmu těm, kteří nejsou v práci s matematickými publikacemi v cizích jazycích příliš zdatní.

Seznam tabulek

Tabulka 4.1: Použité symboly a zkratky

EM	Expectation-Maximization algorithm
GEM	Generalized EM algorithm
ECM	Expectation-Conditional Maximization algorithm
ECME	Expectation-Conditional Maximization Either algorithm
$\frac{\partial}{\partial \boldsymbol{\theta}}$	parciální derivace podle $\boldsymbol{\theta}$
$\times$	znak zastupující násobení při dělení řádků ve vzorcích
$\rightarrow$	"konverguje k", "směřuje k"
Q	základní funkce kroku E
L	logaritmická věrohodnostní funkce
L_c	logaritmická věrohodnostní funkce pro úplná data
max	maximum (funkce)
det	determinant (matice)
exp	základ exponenciální funkce
log	logaritmus o přirozeném základu (Eulerově čísle)
int	vnitřek množiny (množina vnitřních bodů)
$\mathbb{R}^n$	n-rozměrný Euklidovský prostor reálných vektorů
$\boldsymbol{\theta}$	vektor parametrů
Ω	parametrický prostor
$\mathcal{X}$	výběrový prostor úplných dat
$\mathcal{Y}$	výběrový prostor neúplných dat
$\mathcal{S}$	množina všech stacionárních bodů L v $\operatorname{int}\Omega$
$\mathcal{M}$	množina všech lokálních maxim L v $\operatorname{int}\Omega$
$\mathbb{E}_{\boldsymbol{\theta}}$	střední hodnota podmíněná pevnou hodnotou $\boldsymbol{\theta}$
$N(\mu, \sigma^2)$	normální rozdělení o střední hodnotě μ a rozptylu σ^2
$\operatorname{cov}(X, Y)$	kovariance náhodných veličin X a Y
$\operatorname{var}(X)$	rozptyl náhodné veličiny X

Literatura

J. Anděl. *Základy matematické statistiky*. Matfyzpress – Praha, 2007.

A. P. Dempster, N. M. Laird, a D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, 39(1), 1977.

C. Liu a D. B. Rubin. The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence. *Biometrika*, 81(4), 1994.

G. J. McLachlan a T. Krishnan. *The EM Algorithm and Extensions*. John Wiley & Sons – New York, 1997.

X. L. Meng a D. B. Rubin. Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika*, 80(2), 1993.

C. R. Rao. *Linear Statistical Inference and its Applications*. John Wiley & Sons – New York, 1965.

C. F. J. Wu. On the Convergence Properties of the EM Algorithm. *The Annals of Statistics*, 11(1), 1983.