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

DIPLOMOVÁ PRÁCE

Modelování směsí distribučních funkcí

Vedoucí diplomové práce:
doc. RNDr. Eva Fišerová, Ph.D.
Rok odevzdání: 2014

Vypracovala:
Kristýna Vaňkátová
AME, II. ročník

Prohlášení

Prohlašuji, že jsem diplomovou práci zpracovala samostatně pod vedením paní doc. RNDr. Evy Fišerové, Ph.D. s použitím uvedené literatury.

V Olomouci dne 10. března 2014

Poděkování

Na tomto místě bych chtěla poděkovat především své vedoucí diplomové práce doc. RNDr. Evě Fišerové Ph.D. za cenné připomínky a odborné rady, kterými mi pomohla dovést tuto práci ke zdárnému konci. Velký dík patří také Dipl.-Ing. Dr.techn. Bettině Grün, která mi pomohla se závěrečnou částí analýzy dat a poskytla mi důležité rady ohledně práce v programu R.

Dále si zaslouží poděkování má rodina, která mě po celou dobu studia psychicky i finančně podporovala. V neposlední řadě bych ráda vyjádřila svůj vděk tvůrcům sci-fi a fantasy seriálů a autorům mých oblíbených knih, jelikož to byla právě jejich díla, co mi v nejtěžších chvílích poskytla kýžený únik z reality.

Obsah

Úvod

Smíšené distribuce jsou poměrně mladou oblastí matematické statistiky. První zmínka o nich se datuje do konce 19. století, kdy je využil kanadsko-americký matematik Simon Newcomb k detekci odlehlých pozorování. V posledních letech se rozsah a potenciál použití smíšených distribucí široce rozšířil. Aplikace smíšených distribucí se uplatnily například v astronomii, biologii, genetice, medicíně, psychiatrii, ekonomii či marketingu.

Směsi distribucí jsou velmi flexibilní metodou analýzy dat, především protože umožňují pracovat s daty, jejichž tvar distribuční funkce neodpovídá žádnému známému rozdělení pravděpodobností. To jen doplňuje jejich hlavní použití v případech, kdy jsou data nějakým způsobem rozdělena na skupiny, nebo pokud takovou strukturu chceme v datech zkoumat. Smíšené distribuce jsou také vhodné pro posuzování sensitivity a specificity diagnostických testů v případech, kde není k dispozici zlatý standard. Stejně tak lze vhodně využít směs distribucí v problematice neuronových sítí.

První část této práce bude věnována obecným postupům uplatňovaným při modelování smíšených distribucí a problémům s tímto modelováním souvisejícím. Bude zde tedy obsažena základní definice smíšené distribuce, ukázkové příklady, její identifikovatelnost a teoretický zálad EM algoritmu, který je v současné době nejpoužívanějším přístupem k odhadu parametrů směsi distribucí. Pozornost bude mimo jiné věnována stanovení počtu komponent či způsobům volby počátečních odhadů.

Později se budeme věnovat problematice smíšených distribucí v kontextu regresní analýzy, která patří k základním a nejčastějším metodám statistické analýzy vztahů. Je to metoda hojně užívaná v praxi a je obsažena ve většině standardních statistických programovacích systémů. Navíc ji lze použít v kterékoli vědní oblasti, kde sledujeme závislost mezi veličinami. V rámci této kapitoly bude představen EM algoritmus aplikovaný na směsi regresních modelů. Pozornost bude věnována také alternativním metodám, jako jsou klasifikační EM algoritmus a

stochastický EM algoritmus.

V poslední části práce svou pozornost obrátíme na praktické využití doposud shrnutých znalostí. Předmětem zkoumaní budou data firmy Seco GOUP a.s., na kterých budeme demonstrovat nejen modelování smíšené distribuce, ale též odhadování regresních parametrů pomocí směsi regresních modelů. Ke všem výpočtům využijeme program R a funkce balíčků `mixtools`, `fpc` a `flexmix`.

1 Úvod do problematiky smíšených distribucí

1.1 Stručná historie

Smíšené distribuce jsou poměrně mladou oblastí matematické statistiky. Jejich první objev ve statistické literatuře se datuje do konce 19. století, kdy je využil kanadsko-americký matematik Simon Newcomb k detekci odlehlých pozorování.

První významnou analýzu, která zahrnovala použití smíšených distribucí, provedl o pár let později známý anglický biometrik Karl Pearson a založil ji na metodě momentů. Data ke zkoumání mu poskytl přítel W.F.R. Weldon, kerý měl k dispozici 1000 měření poměru čela k celkové délce těla krabů ze zátoky Naples. Weldon si všiml značné šikmosti v histogramu těchto dat a přemýšlel nad tím, zda by se ve skutečnosti nemohlo jednat o populaci skládající se z dvou nových poddruhů. Jelikož však jeho znalosti matematiky nebyly dostačující, obrátil se s problémem na Pearsona, který se rozhodl ho řešit pomocí modelování směsi dvou normálních distribucí s rozdílnými středními hodnotami μ_1 a μ_2 a rozptyly σ_1 a σ_2 s proporcemi π_1 a π_2. Pearson se tedy potýkal s odhadem pěti parametrů a momentovou metodou se dostal až k řešení polynomu 9. stupně, což na danou dobu byla velmi složitá výpočetní operace.

Pearsonovu metodu založenou na velkém množství algebraických výpočtů se na začátku 19. století pokusilo několik matematiků zjednodušit. Přesto však bylo dalších 30 let ve zkoumání smíšených distribucí využíváno převážně metody momentů a došlo pouze k rozšíření této metody o případ dvou hustot z dvourozměrného normálního rozdělení a o případ více než dvou jednorozměrných normálních hustot.

Kvůli složitosti výpočtů byl důraz kladen na grafické techniky modelování smíšených distribucí, avšak s nástupem výkonnějších a rychlejších počítačů se pozornost obrátila k odhadu pomocí maximální věrohodnosti. Oficiálně tuto metodu zcela poprvé v kontextu smíšených distribucí použil matematik a statistik C.R. Rao v roce 1948, ačkoliv podle některých statistiků (R.W. Butler) to byl již

Newcomb, který vymyslel iterační metodu, jejíž podstata by se dala označit jako EM algoritmus, který iteračně řeší maximalizaci věrohodnostní funkce.

Po roce 1948 nebyla metoda maximální věrohodnost dále rozšiřována až do období 60. let, kdy byla problematika smíšených distribucí obohacena o k-složkovou směs normálních distribucí se stejným rozptylem a následně o směs exponenciálních rozdělení. Až v 70. letech bylo několika statistiky demonstrováno, že je v případě smíšených distribucí místo momentové metody vhodnější používat odhad pomocí maximální věrohodnosti, a roku 1977 byl A.P. Dempsterem položen teoretický základ pro EM algoritmus a jeho použití při modelování smíšených distribucí.

V posledních letech se rozsah a potenciál použití smíšených distribucí široce rozšířil. Aplikace smíšených distribucí se uplatnily například v astronomii, biologii, genetice, medicíně, psychiatrii, ekonomii či marketingu.

Jedním z mnoha využití je například shluková analýza, což je snadno odvoditelné již ze samotné podstaty smíšených distribucí. Dále jsou smíšené distribuce vhodné pro posuzování sensitivity a specificity diagnostických testů v případech, kde není k dispozici zlatý standard. Stejně tak lze vhodně využít směs normálních distribucí se společným rozptylem pro aproximaci libovolných spojitých funkcí. Díky své flexibilitě našly smíšené distribuce uplatnění také v problematice neuronových sítí. [13, 8]

1.2 Formulace problému smíšených distribucí

Smíšené distribuce poskytují 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.

Nejprve zmiňme, že uvažujeme náhodný výběr $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)^{\mathrm{T}}$ rozsahu n, kde $\mathbf{Y}_j$ je p-dimenzionální náhodný vektor s funkcí hustoty $f(\mathbf{y}_j)$ definovanou na $\mathbb{R}^p$. Vektor $\mathbf{Y}$ tím pádem reprezentuje celý výběr, což je n-tice bodů v $\mathbb{R}^p$. Ozna-

čením $\mathbf{y}= (\mathbf{y}_1, \ldots, \mathbf{y}_n)^{\mathrm{T}}$ budeme v dalším textu rozumět napozorované hodnoty náhodného vektoru $\mathbf{Y}$.

Definice 1.1. Nechť $f_i(\mathbf{y_j})$ je hustota rozdělení pravděpodobností pro každé $i = 1, \ldots, c$ a $0 \leq \pi_i \leq 1$, $i = 1, 2, \cdots, c$, jsou konstanty takové, že platí $\sum_{i=0}^{c} \pi_i = 1$. Potom hustotu

$$f(\mathbf{y_j}) = \sum_{i=1}^{c} \pi_i f_i(\mathbf{y_j}) \tag{1}$$

nazveme *smíšenou hustotou*. Jí odpovídající distribuční funkci $F(\mathbf{y_j})$ nazveme *smíšenou distribucí*.

Vzhledem ke skutečnosti, že tato práce bude zaměřena na smíšené hustoty normálního rozdělení, je na tomto místě vhodné uvést také parametrickou formulaci.

Definice 1.2. Nechť pro každé $i = 1, \ldots, c$ je $f_i(\mathbf{y_j}; \boldsymbol{\theta}_i)$ hustota rozdělení pravděpodobností náležící do známé třídy hustot a $0 \leq \pi_i \leq 1$, $i = 1, 2, \cdots, c$, jsou konstanty takové, že platí $\sum_{i=0}^{c} \pi_i = 1$. Potom hustotu

$$f(\mathbf{y_j}, \boldsymbol{\Psi}) = \sum_{i=1}^{c} \pi_i f_i(\mathbf{y_j}; \boldsymbol{\theta}_i) \tag{2}$$

nazveme *smíšenou hustotou*. Jí odpovídající distribuční funkci $F(\mathbf{y_j}, \boldsymbol{\Psi})$ nazveme *smíšenou distribucí*.

Vektor $\boldsymbol{\Psi}$ obsahuje všechny neznámé parametry smíšené distribuce a můžeme ho zapsat jako

$$\boldsymbol{\Psi} = (\pi_1, \ldots, \pi_{c-1}, \boldsymbol{\xi}^{\mathrm{T}})^{\mathrm{T}}.$$

Symbolem $\boldsymbol{\Omega}$ označíme parametrický prostor pro $\boldsymbol{\Psi}$. Parametry $\pi_1, \cdots, \pi_c$ udávají váhy složek smíšené distribuce a vektor $\boldsymbol{\xi}$ obsahuje všechny parametry

$$(\boldsymbol{\theta}_1^{\mathrm{T}}, \boldsymbol{\theta}_2^{\mathrm{T}}, \ldots, \boldsymbol{\theta}_c^{\mathrm{T}})^{\mathrm{T}}.$$

Značení $f_i(\mathbf{y}; \boldsymbol{\theta}_i)$ reprezentuje hustotu i-té složky, jejíž parametry jsou zastoupeny vektorem $\boldsymbol{\theta}_i$. Počet složek smíšené distribuce je dán parametrem c.

Jelikož funkce $f_1(\mathbf{y}), \ldots, f_c(\mathbf{y})$ jsou hustoty známých rozdělení pravděpodobností, platí také, že funkce $f(\mathbf{y})$ ze vztahu (2) je hustotou. A protože váhy složek smíšené distribuce π_i jsou normované, jejich součtem musíme dostat číslo 1. Díky tomu v definici vektoru $\boldsymbol{\Psi}$ můžeme vynechat parametr π_c.

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

Funkční hodnotu $f(\mathbf{y_j})$ potom spočítáme tak, že určíme funkční hodnoty dílčích hustot $f_i(\mathbf{y}_j; \boldsymbol{\theta}_i)$ v bodě $\mathbf{y_j}$, vynásobíme je příslušnými vahami a sečteme je. [13]

1.3 Koncept neúplných dat

Nechť je Z_j kategoriální náhodná proměnná, jejíž dosažitelné hodnoty jsou $1, \ldots, c$ s jim náležícími pravděpodobnostmi $\pi_1, \ldots, \pi_c$. Potom můžeme říci, že hustota vektoru $\mathbf{Y}_j$ je $f_i(\mathbf{y}_j)$ za předpokladu, že $Z_j = i$, kde $i = 1, \ldots, c$. Marginální (nepodmíněná) hustota vektoru $\mathbf{Y}_j$ je dána jako $f(\mathbf{y}_j)$. Z těchto informací je zřetelné, že proměnná Z_j může být brána jako označení složky, do které bude zařazen vektor $\mathbf{Y}_j$. V dalším textu pro nás však bude výhodnější uvažovat namísto kategoriální proměnné vektor $\mathbf{Z}_j$, jehož složky jsou $Z_{ij} = (\mathbf{Z}_j)_i$ a mohou nabývat hodnot 0 či 1 - dle toho, zda vektor $\mathbf{Y}_j$ v smíšené distribuci náleží či nenáleží do složky i. Je tedy zřejmé, že vektor $\mathbf{Z}_j$ podléhá multinomickému rozdělení odpovídajícímu tahu z c kategorií s pravděpodobnostmi $\pi_1, \ldots, \pi_c$. Jinak řečeno

$$\mathbf{Z}_j \sim \mathrm{Mult}_c(1, \boldsymbol{\pi}).$$

Interpretace takto zapsaného modelu je následující: řekněme, že smíšená distribuce se skládá z c složek a vektor $\mathbf{Y}_j$ je člen náhodného výběru z populace G, která je tvořena c skupinami $G_1, \ldots, G_c$, jejichž proporce v populaci jsou

$\pi_1, \ldots, \pi_c$. Dále pokud hustota vektoru $\mathbf{Y}_j$ ve skupině G_i je dána jako $f_i(\mathbf{y}_j)$ pro $i = 1, \ldots, c$, pak hustota vektoru $\mathbf{Y}_j$ je dána vztahem (2).

Ukazuje se, že tento koncept je výhodný především z hlediska následujícího řešení maximalizace věrohodnostní funkce skrze EM algoritmus, v jehož rámci nahlížíme na napozorovaná data $\mathbf{y}_1, \ldots, \mathbf{y}_n$ jako na neúplná. K nim jsou přidruženy vektory udávající příslušnost k jednotlivým složkám $\mathbf{z}_1, \ldots, \mathbf{z}_n$, které jsou z našeho pohledu nedostupné. Kompletní data potom vypadají následovně

$$\mathbf{y_k} = (\mathbf{y}, \mathbf{z}),$$

kde index k vyjadřuje, že data jsou kompletní, a

$$\mathbf{y} = (\mathbf{y_1}, \ldots, \mathbf{y_n})^{\mathrm{T}}$$

jsou napozorovaná neúplná data a

$$\mathbf{z} = (\mathbf{z_1}, \ldots, \mathbf{z_n})^{\mathrm{T}}$$

je nepozorovatelná matice udávající příslušnost k jednotlivým složkám.

V tomto kontextu lze na proporci π_i nahlížet jako na původní pravděpodobnost, že subjekt patří do i-té složky smíšené distribuce. Pravděpodobnost $\tau_i(\mathbf{y}_j)$, že subjekt náleží do i-té složky, avšak za podmínky, že máme k dispozici napozorovaná data $\mathbf{y}_j$, je potom dána vztahem [13]

$$\tau_i(\mathbf{y}_j) = \mathrm{P}(Z_{ij} = 1|\mathbf{y}_j) = \frac{\pi_i f_i(\mathbf{y}_j)}{f(\mathbf{y}_j)}, \quad i = 1, \ldots, c, \quad j = 1, \ldots, n.$$

1.4 Smíšené distribuce s normálně rozdělenými složkami

Tato práce se bude zabývat problematikou modelování smíšených distribucí, kde jednotlivé složky mají normální rozdělení. Proto si v této podkapitole představíme teorii potřebnou k pochopení jednorozměrných i vícerozměrných smíšených distribucí složených z hustot náhodných vektorů podléhajících normálnímu rozdělení. Tato kapitola je zpracována především pomocí zdrojů [13, 10]

Jednorozměrné smíšené distribuce normálních rozdělení

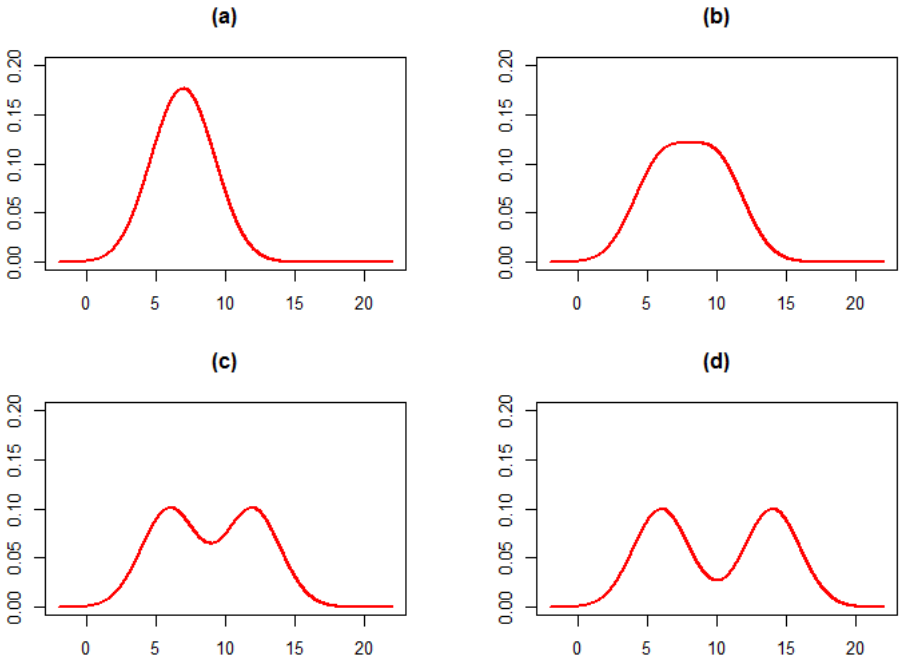
Pro lepší představu významu definice 1.1 si v následující části představíme některé poznatky o smíšených distribucích složených ze dvou složek s normálním rozdělením. Nejprve si ukážeme, jak vypadají vybrané hustoty dvousložkových smíšených distribucí normálních rozdělení se stejným rozptylem σ^2, za předpokladu proporcí π_1 a π_2 a středních hodnot μ_1 a μ_2. Odpovídající hustota má tvar

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

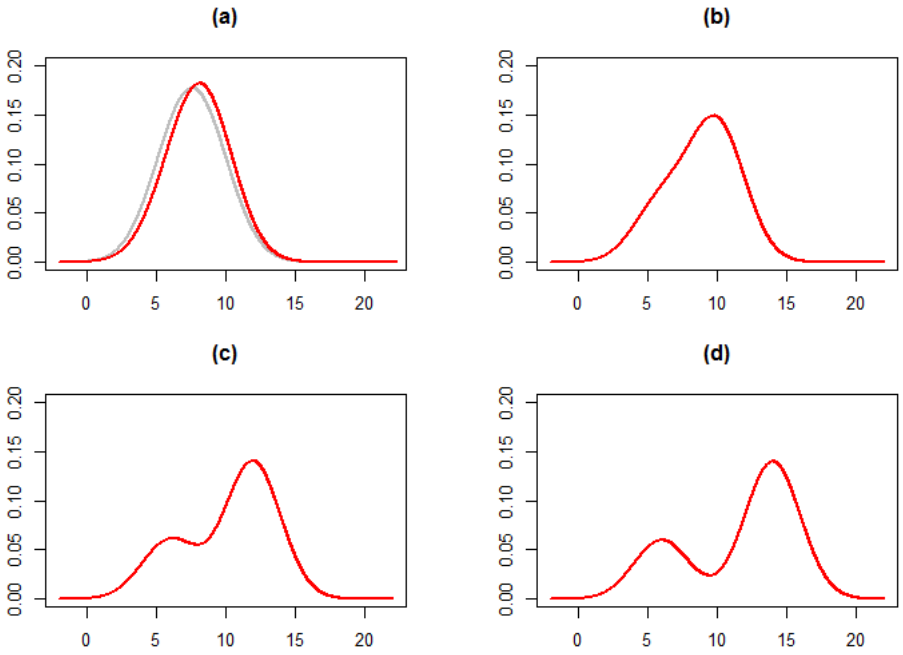
kde

$$\phi(y; \mu, \sigma^2) = (2\pi)^{-\frac{1}{2}} \sigma^{-1} \exp\{-\frac{1}{2}(y-\mu)^2/\sigma^2\}$$

vyjadřuje hustotu normálního rozdělení se střední hodnotou μ a rozptylem σ^2.



Obr. 1.1: Grafy hustot smíšených distribucí dvou normálních homoskedastických rozdělení s rozptylem $\sigma = 2$, stejnými vahami složek a středními hodnotami $\mu_1 = 6$ a $\mu_2 = 6 + 2\Delta$, kde je Δ rovno: (a)$\Delta = 1$, (b)$\Delta = 2$, (c)$\Delta = 3$, (c)$\Delta = 4$.



Obr. 1.2: Grafy hustot smíšených distribucí dvou normálních homoskedastických rozdělení s rozptylem $\sigma = 2$, vahami složek $\pi_1 = 0.3$ a $\pi_2 = 0.7$ a středními hodnotami $\mu_1 = 6$ a $\mu_2 = 6 + 2\Delta$, kde je Δ rovno: (a)$\Delta = 1$, (b)$\Delta = 2$, (c)$\Delta = 3$, (d)$\Delta = 4$. Na obrázku (a) je navíc šedě znázorněna hustota pro $\pi_1 = \pi_2 = 0.5$.

Hustotu $f(y)$ ze vztahu (3) označíme za bimodální, pokud se střední hodnoty μ_1 a μ_2 od sebe dostatečně liší. V tomto případě bude hustota $f(y)$ vypadat jako dvě hustoty položené vedle sebe. Naopak unimodální hustotu dostaneme, jestliže jsou si střední hodnoty dostatečně blízko. Obecně je vzdálenost středních hodnot definována vztahem

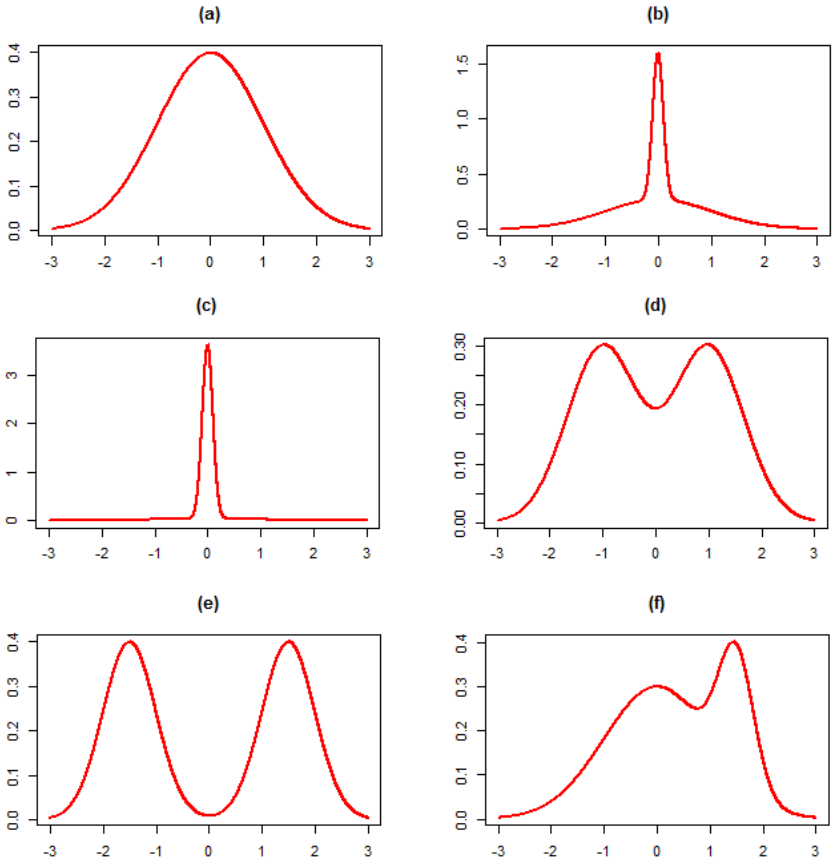
$$\Delta = \frac{|\mu_1 - \mu_2|}{\sigma}.$$

Nazýváme ji Mahalanobisova vzdálenost mezi homoskedastickými složkami smíšené distribuce normálních rozdělení.

Příklady smíšených distribucí o dvou normálně rozdělených složkách se stejným rozptylem lze spatřit na obrázku 1.1. Jedná se o čtyři různé smíšené distribuce lišící se rozdílem středních hodnot jejich složek tak, že výsledná hustota

přechází z unimodální do bimodální.

Pomocí smíšené distribuce lze také dosáhnout asymetrické hustoty. Stane se tak v případě, že střední hodnoty obou složek jsou blízké, avšak každá složka má jinou váhu. Pro názornost v obrázku 1.2 uvedeme smíšené distribuce z grafů v 1.1, avšak tentokrát změníme proporce na $\pi_1 = 0.3$ a $\pi_2 = 0.7$.



Obr. 1.3: Grafy dvousložkových normálních distribucí z tabulky 1.1.

Šikmost a špičatost $f(y)$ lze v případě stejného rozptylu obou složek vypočítat dle následujících vzorců

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

a

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

kde a je poměr větší váhy složky k menší a Δ je Mahalanobisova vzdálenost mezi složkami.

graf	název hustoty	hustota $f(y)$
(a)	Gaussova	$N(0,1)$
(b)	Špičatá unimodální	$\frac{2}{3}N(0,1) + \frac{1}{3}N(0,(\frac{1}{10})^2)$
(c)	Odlehlá	$\frac{1}{10}N(0,1) + \frac{9}{10}N(0,(\frac{1}{10})^2)$
(d)	Bimodální	$\frac{1}{2}N(-1,(\frac{2}{3})^2) + \frac{1}{2}N(1,(\frac{2}{3})^2)$
(e)	Oddělená bimodální	$\frac{1}{2}N(-\frac{3}{2},(\frac{1}{3})^2) + \frac{1}{2}N(\frac{3}{2},(\frac{2}{3})^2)$
(f)	Šikmá bimodální	$\frac{3}{4}N(0,1) + \frac{1}{4}N(\frac{3}{2},(\frac{1}{3})^2)$

Tabulka 1.1: Dvousložkové normální smíšené distribuce.

S rozdílným rozptylem složek se samozřejmě celá situace komplikuje. Užitečnou publikaci o této problematice napsal I. Eisenberger, který například definoval nerovnost, podle které je možné rozhodnout o bimodalitě distribuce. Hustota smíšené distribuce $f(y)$ není bimodální, pokud platí

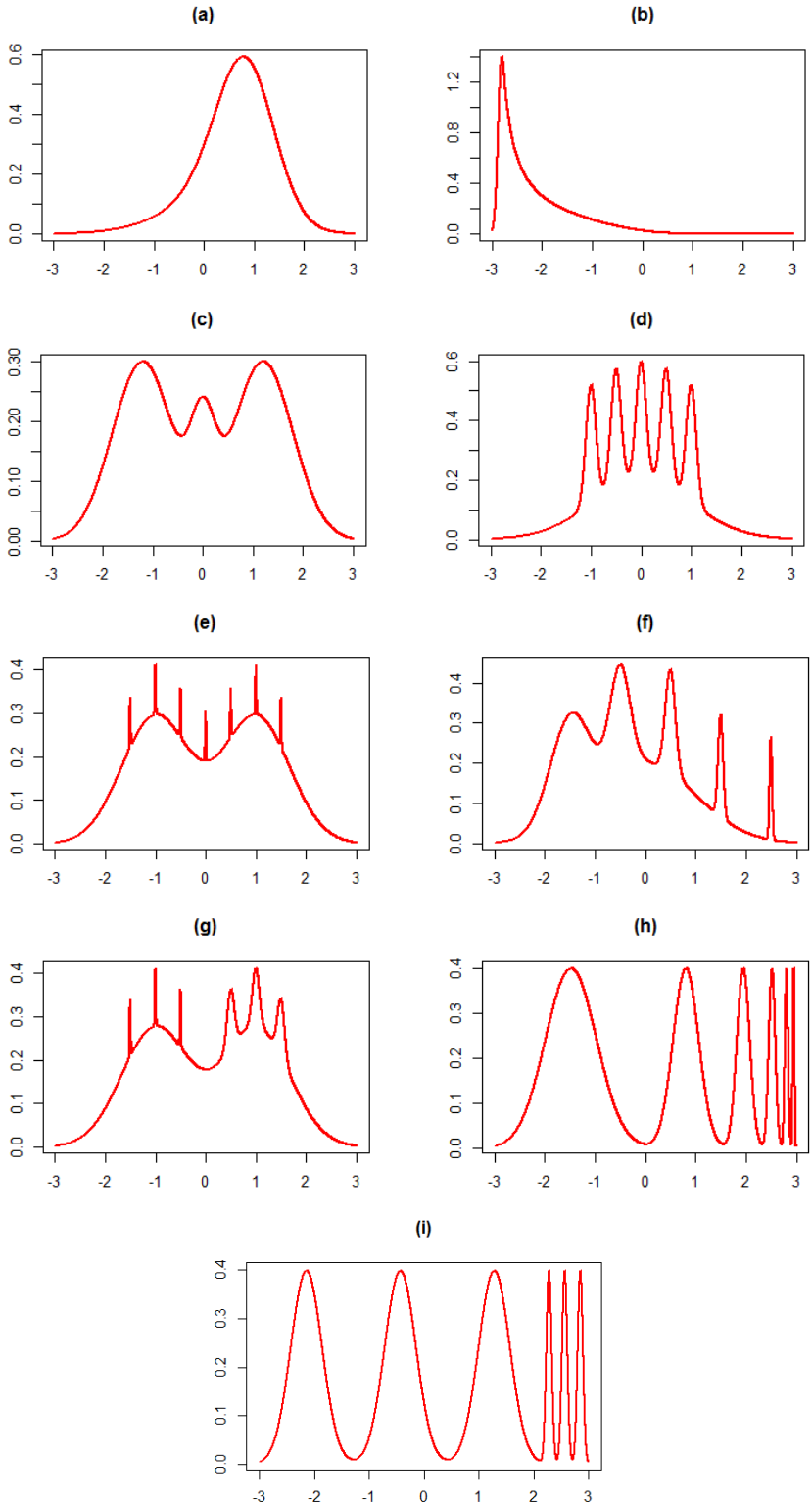
$$\Delta^2 < (27\sigma_2^2)/[4(1+k)],$$

za předpokladu, že $k = \sigma_2^2/\sigma_1^2$. Naopak opačná nerovnost

$$\Delta^2 > (27\sigma_2^2)/[4(1+k)],$$

znamená, že existuje takové π, pro které je hustota $f(y)$ bimodální.

Příklady několika reprezentativních dvousložkových smíšených distribucí nalezneme na obrázku 1.3 a v tabulce 1.1 jsou uvedeny tvary těchto distribucí.



Obr. 1.4: Grafy vícesložkových smíšených distribucí z tabulky 1.2.

Přesuňme se k problematice c-složkových smíšených distribucí, která je však natolik rozmanitá, že se v této práci omezíme pouze na představení několika vybraných příkladů grafů smíšených hustot pro různé počty složek, střední hodnoty a rozptyly. Grafické zobrazení nalezneme v obrázku 1.4 a zadání smíšených funkcí v tabulce 1.2.

graf	název hustoty	hustota $f(y)$
(a)	Šikmá unimodální	$\frac{1}{5}N(0,1) + \frac{1}{5}N(\frac{1}{2},(\frac{2}{5})^2) + \frac{3}{5}N(\frac{13}{15},(\frac{5}{9})^2)$
(b)	Silně zešikmená	$\sum_{i=0}^{7} \frac{1}{8}N(3[(\frac{2}{3})^i - 1],(\frac{2}{3})^{2i})$
(c)	Trimodální	$\frac{9}{20}N(-\frac{6}{5},(\frac{3}{5})^2) + \frac{9}{20}N(\frac{6}{5},(\frac{3}{5})^2) + \frac{1}{10}N(0,(\frac{1}{4})^2)$
(d)	Dráp	$\frac{1}{2}N(0,1) + \sum_{i=0}^{4} \frac{1}{10}N(\frac{i}{2} - 1,(\frac{1}{10})^2)$
(e)	Dvojitý dráp	$\frac{49}{100}N(-1,(\frac{2}{3})^2) + \frac{49}{100}N(1,(\frac{2}{3})^2)$ $+ \sum_{i=0}^{6} \frac{1}{350}N((i-3)/2,(\frac{1}{100})^2)$
(f)	Asymetrický dráp	$\frac{1}{2}N(0,1) + \sum_{i=-2}^{2}(2^{1-i}/31)N(i + \frac{1}{2},(2^{-i}/10)^{2i})$
(g)	Asymetrický dvojitý dráp	$\sum_{i=0}^{1} \frac{46}{100}N(2i-1,(\frac{2}{3})^2) + \sum_{i=1}^{3} \frac{1}{300}N(-\frac{i}{2},(\frac{1}{100})^2)$ $+ \sum_{i=1}^{3} \frac{7}{300}N(\frac{i}{2},(\frac{7}{100})^2)$
(h)	Hladký hřeben	$\sum_{i=0}^{5} 2^{5-i}/63N((65 - 96(\frac{1}{2})^i/21),(\frac{32}{63})^2/2^{2i})$
(i)	Diskrétní hřeben	$\sum_{i=0}^{2} \frac{2}{7}N((12i-15)/7,(\frac{2}{7})^2) + \sum_{i=8}^{10} \frac{1}{21}N(2i/7,(\frac{1}{21})^2)$

Tabulka 1.2: Vícesložkové normální smíšené distribuce.

Vícerozměrné smíšené distribuce

V této části si stručně představíme základy vícerozměrných smíšených distribucí, v našem případě se omezíme na smíšené distribuce složené z p-rozměrných normálních rozdělení. Obecně je takováto hustota vyjádřena v následujícím tvaru

$$f(\mathbf{y}) = \pi_1 f(\mathbf{y};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1) + \pi_2 f(\mathbf{y};\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2) + \cdots + \pi_c f(\mathbf{y};\boldsymbol{\mu}_c,\boldsymbol{\Sigma}_c), \mathbf{y} \in \mathbb{R}^p,$$

kde opět předpokládáme, že $\sum_{i=0}^{c} \pi_i = 1, \pi > 0$. Vektor $\boldsymbol{\mu}_i$ reprezentuje vektor středních hodnot i-té složky a matice $\boldsymbol{\Sigma}_i$ je její varianční matice.

2 Odhady parametrů smíšených distribucí

Jak již bylo stručně zmíněno v sekci 2.1, v průběhu let bylo použito mnoho různých přístupů k odhadu parametrů ve smíšených distribucích. Tyto přístupy zahrnovaly například grafické metody, momentové metody, metody založené na minimální vzdálenosti, maximální věrohodnost a bayesovské přístupy. Pravděpodobně hlavní důvod velkého počtu prací zabývajících se touto tématikou je skutečnost, že prozatím není k dispozici žádná explicitní formule pro odhad parametrů. Například odhad pomocí metody maximální věrohodnosti nelze počítat analyticky a je nutné použít iterační postup. Tím je v našem případě EM algoritmus.

Také již bylo řečeno, že od zavedení EM algoritmu je metoda maximální věrohodnosti nejpoužívanější metodou v této oblasti, proto se nadále budeme zabývat jen touto metodou a EM algoritmem.

2.1 Identifikovatelnost

Předpoklad identifikovatelnosti je základem většiny statistické teorie a praxe. Především v praxi je velmi důležité zvážit identifikovatelnost, protože bez ní by procedury umožňující odhady nemusely být správně definovány.

Pro odhady parametrů smíšené distribuce je klíčové vědět, zda je vektor $\mathbf{\Psi}$ identifikovatelný a zda je tím pádem smysluplné snažit se získat jeho odhad. Obecně můžeme říci, že třída hustot $f(\mathbf{y}_j, \mathbf{\Psi})$ je identifikovatelná, pokud jednotlivé hodnoty parametru $\mathbf{\Psi}$ určují jednotlivé členy třídy hustot.

Definice 2.1. Nechť $f(\mathbf{y}_j, \mathbf{\Psi})$ je třída hustot, pro kterou platí, že odlišné hodnoty parametru $\mathbf{\Psi}$ určují odlišné členy třídy hustot

$$\{f(\mathbf{y}_j, \mathbf{\Psi}) : \mathbf{\Psi} \in \mathbf{\Omega}\}.$$

Jinak řečeno,

$$f(\mathbf{y}_j, \mathbf{\Psi}) = f(\mathbf{y}_j, \mathbf{\Psi}^*),$$

tehdy a jen tehdy, když

$$\boldsymbol{\Psi} = \boldsymbol{\Psi}^*.$$

Potom řekneme, že je tato třída hustot identifikovatelná.

Obecně u smíšených distribucí rozeznáváme tři hlavní druhy neidentifikovatelnosti. První dva jsou považovány za lehce řešitelné a lze jim zabránit pomocí vhodných omezení. Mnohem podstatnější problém je nazýván generická identifikovatelnost. [13, 16]

2.1.1 Neidentifikovatelnost způsobená invariancí vzhledem ke změně označení komponent

Tento druh neidentifikovatelnosti byl poprvé zaznamenán Renderem a Walkerem v roce 1984. Můžeme ho definovat jako možnost prohodit pořadí a označení jednotlivých složek, aniž by se změnila výsledná smíšená distribuce.

Jako příklad můžeme uvést smíšenou distribuci skládající se z dvou normálně rozdělených složek, $f(\mathbf{y}_j, \boldsymbol{\Psi}) = \pi_1\phi(\mathbf{y_j}; \mu_1, \sigma_1^2) + \pi_2\phi(\mathbf{y_j}; \mu_2, \sigma_2^2)$. V tomto případě $\boldsymbol{\theta}_k = (\mu_k, \sigma_k^2)^{\mathrm{T}}$ a $\boldsymbol{\Psi} = (\boldsymbol{\theta}_1^{\mathrm{T}}, \boldsymbol{\theta}_2^{\mathrm{T}}, \pi_1, \pi_2)^{\mathrm{T}} \in \boldsymbol{\Omega}$. Potom také existuje vektor $\boldsymbol{\Psi}^* = (\boldsymbol{\theta}_2^{\mathrm{T}}, \boldsymbol{\theta}_1^{\mathrm{T}}, \pi_2, \pi_1)^{\mathrm{T}}$, který vznikne přehozením pořadí jednotlivých komponent. Je zřejmé, že výsledné smíšené distribuce vytvořené pomocí $\boldsymbol{\Psi}$ a $\boldsymbol{\Psi}^*$ jsou identické

$$f(\mathbf{y}_j, \boldsymbol{\Psi}^*) = \pi_2 f(\mathbf{y_j}; \mu_2, \sigma_2^2) + \pi_1 f(\mathbf{y_j}; \mu_1, \sigma_1^2) =$$

$$\pi_1 f(\mathbf{y_j}; \mu_1, \sigma_1^2) + \pi_2 f(\mathbf{y_j}; \mu_2, \sigma_2^2) = f(\mathbf{y}_j, \boldsymbol{\Psi})$$

V tomto pojetí tedy tato smíšená distribuce není identifikovatelná. [16]

2.1.2 Neidentifikovatelnost způsobená potenciálním přeurčením

Dalším problémem s identifikovatelností je možnost přeurčení modelu, kterého dosáhneme, když do smíšené distribuce zahrneme více složek, než je potřeba.

Uvažujme konečnou smíšenou distribuci s c složkami, definovanou jako

$$f(\mathbf{y_j}, \boldsymbol{\Psi}) = \sum_{i=1}^{c} \pi_i f_i(\mathbf{y_j}; \boldsymbol{\theta}_i),$$

kde $\boldsymbol{\Psi} = (\pi_1, \ldots, \pi_{c-1}, \boldsymbol{\theta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\theta}_c^{\mathrm{T}})^{\mathrm{T}} \in \boldsymbol{\Omega}_c$. Dále mějme smíšenou distribuci ze stejné parametrické třídy distribucí, avšak s pouhými c-1 složkami. Jak Crawford v roce 1994 dokázal, jakákoli smíšená distribuce s c-1 komponentami určuje neidentifikovatelnou podmnožinu v parametrickém prostoru $\boldsymbol{\Omega}_c$, který odpovídá smíšené distribuci s c komponentami, kde jedna z nich má nulovou váhu nebo jsou dvě komponenty shodné.

Pokud se vrátíme k příkladu uvedenému v předchozí podkapitole, vektor parametrů je $\boldsymbol{\Psi} = (\pi_1, \mu_1, \sigma_1^2, \pi_2, \mu_2, \sigma_2^2)^{\mathrm{T}} \in \boldsymbol{\Omega}_2$ a uvažujeme smíšenou distribuci skládající se z dvou hustot normálního rozdělení. Jakákoli smíšená distribuce dvou normálních hustot může být zapsána jako třísložková distribuce přidáním složky π_3 s váhou 0, což můžeme vyjádřit jako

$$f(\mathbf{y}_j, \boldsymbol{\Psi}) = \pi_1 f(\mathbf{y_j}; \mu_1, \sigma_1^2) + \pi_2 f(\mathbf{y_j}; \mu_2, \sigma_2^2) =$$

$$\pi_1 f(\mathbf{y_j}; \mu_1, \sigma_1^2) + \pi_2 f(\mathbf{y_j}; \mu_2, \sigma_2^2) + 0 \cdot f(\mathbf{y_j}; \mu_3, \sigma_3^2).$$

Ke stejné neidentifikovatelnosti dospějeme, pokud vytvoříme smíšenou distribuci rozdělením jedné složky původní identifikovatelné dvousložkové smíšené distribuce:

$$f(\mathbf{y}_j, \boldsymbol{\Psi}) = \pi_1 f(\mathbf{y_j}; \mu_1, \sigma_1^2) + \pi_2 f(\mathbf{y_j}; \mu_2, \sigma_2^2) =$$

$$\pi_1 f(\mathbf{y_j}; \mu_1, \sigma_1^2) + (\pi_2 - \pi_3) f(\mathbf{y_j}; \mu_2, \sigma_2^2) + \pi_3 f(\mathbf{y_j}; \mu_2, \sigma_2^2),$$

kde hustota $f(\mathbf{y}_j, \boldsymbol{\Psi})$ bude stejná pro všechny hodnoty π_3, pro které platí $0 \leq \pi_3 \leq \pi_2$. [16]

2.1.3 Omezení vektoru parametrů

V mnoha případech shledáme smíšenou distribuci neidentifikovatelnou ve smyslu výše uvedených druhů neidentifikovatelnosti, avšak tento nedostatek není překážkou pro obyčejné odhady parametrů smíšených distribucí pomocí EM algoritmu, problém představuje až v bayesovském pojetí smíšených distribucí. Navíc

lze neidentifikovatelnost těchto druhů snadno zvládnout pomocí restrikce vektoru parametrů $\boldsymbol{\Psi}$ takovým způsobem, že smíšená distribuce bude obsahovat c rozdílných neprázdných složek.

Nejdříve je nutné omezit váhy složek tak, aby každá z nich byla kladná a nedošlo tak k neidentifikovatelnosti způsobené prázdnými složkami. Jednoduchá podmínka je následující

$$\pi_i > 0, \qquad i = 1, \ldots, c. \tag{4}$$

Dále je třeba omezit parametry tak, abychom se vyhnuli neidentifikovatelnosti způsobené rovností složek směsí. V případě jednoparametrických složek budeme tedy požadovat, aby

$$\theta_i \neq \theta_{i'}, \quad \forall i \neq i', \quad i, i' = 1, \ldots, c.$$

U smíšených distribucí s víceparametrickými složkami jsou k dispozici dvě obdobné podmínky, jedna velmi silná a jedna slabší. Místo jednorozměrného parametru θ_i teď totiž pracujeme s celým vektorem parametrů $\boldsymbol{\theta}_i$, tudíž můžeme porovnávat více složek vektoru. Silnější z podmínek říká, že pokud uvažujeme $\boldsymbol{\theta}_i$ a $\boldsymbol{\theta}_{i'}$, pak se složky s těmito dvěma vektory parametrů liší pouze tehdy, když se vektory liší ve všech svých složkách. To je ovšem poněkud silný předpoklad, protože po jeho aplikaci bychom nemohli vytvořit například smíšenou distribuci o dvou normálních složkách, kde rozptyly složek jsou shodné, avšak střední hodnoty se liší. Jako vhodnější se tedy jeví používat slabší podmínku, která požaduje, aby se vektory parametrů lišily alespoň v jedné složce, což musí platit pro všechny dvojice komponent smíšené distribuce.

Toto byly podmínky zabraňující přeurčenosti smíšené distribuce. Následující omezení nám naproti tomu pomůže vyvarovat se neidentifikovatelnosti způsobené invariancí vzhledem ke změně označení komponent. Zároveň lze podmínku upravit tak, aby zahrnovala i vztah (4):

$$0 < \pi_1 < \ldots < \pi_c.$$

[13, 16]

2.1.4 Generická identifikovatelnost

Navzdory tomu, že provedeme všechna potřebná omezení vektoru parametrů $\mathbf{\Psi}$, může smíšená distribuce zůstat neidentifikovatelnou ve smyslu generické identifikovatelnosti.

Definice 2.2. Nechť

$$f(\mathbf{y_j}, \mathbf{\Psi}) = \sum_{i=1}^{c} \pi_i f_i(\mathbf{y_j}; \boldsymbol{\theta}_i)$$

a

$$f(\mathbf{y_j}, \mathbf{\Psi}^*) = \sum_{i=1}^{c^*} \pi_i^* f_i(\mathbf{y_j}; \boldsymbol{\theta}_i^*)$$

jsou libovolné dvě smíšené distribuce náležící do dané třídy smíšených distribucí. Potom o této třídě řekneme, že je identifikovatelná pro $\mathbf{\Psi} \in \mathbf{\Omega}$, pokud

$$f(\mathbf{y}_j, \mathbf{\Psi}) \equiv f(\mathbf{y}_j, \mathbf{\Psi}^*),$$

a to tehdy a jen tehdy, když $c = c^*$ a je možné permutovat označení složek tak, že

$$\pi = \pi^* \quad \text{a} \quad f_i(\mathbf{y}_j, \boldsymbol{\theta}_i) = f_i(\mathbf{y}_j, \boldsymbol{\theta}_i^*) \quad \text{pro každé } i = (1, \ldots, c).$$

Symbol $\equiv$ značí rovnost hustot pro téměř všechna $\mathbf{y}_j$. [13]

Věta 1. *Nechť*

$$\mathcal{F} = \{F(\mathbf{x}, \boldsymbol{\theta}), \boldsymbol{\theta} \in \mathbf{\Theta}, \mathbf{x} \in \mathbb{R}^d\}$$

je třída d-dimenzionálních distribučních funkcí, ze kterých budou smíšené distribuce sestaveny. Dále mějme třídu $\mathcal{H}$ konečných smíšených distribucí o složkách z $\mathcal{F}$,

$$\mathcal{H} = \{H(\mathbf{x}) : H(\mathbf{x}) = \sum_{i=1}^{c} \pi_i F(\mathbf{x}; \boldsymbol{\theta}_i), \pi_i > 0, \sum_{i=1}^{c} \pi_i = 1, F(\cdot; \boldsymbol{\theta}_i) \in \mathcal{F}, \mathbf{x} \in \mathbb{R}^d\}.$$

Takto definovaná třída distribučních funkcí $\mathcal{H}$ je konvexním obalem $\mathcal{F}$. [16]

Důkaz. Nejprve dokážeme, že je množina $\mathcal{H}$ konvexní. Mějme libovolné $G(\mathbf{x}), H(\mathbf{x}) \in \mathcal{H}$, pro které platí

$$G(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i F(\mathbf{x}; \boldsymbol{\mu}_i), \qquad H(\mathbf{x}) = \sum_{i=1}^{n} \beta_i F(\mathbf{x}; \boldsymbol{\nu}_i),$$

kde $F(\mathbf{x}; \boldsymbol{\mu}_i),\ F(\mathbf{x}; \boldsymbol{\nu}_i) \in \mathcal{F}, \alpha_i, \beta_i \geq 0, \sum_{i=1}^{m} \alpha_i = \sum_{i=1}^{n} \beta_i = 1$. Dále nechť $\lambda \in [0, 1]$ je libovolnělené. Potom

$$I(\mathbf{x}) = \lambda G(\mathbf{x}) + (1 - \lambda) H(\mathbf{x}) = \sum_{i=1}^{m} \lambda \alpha_i F(\mathbf{x}; \boldsymbol{\mu}_i) + \sum_{i=1}^{n} (1 - \lambda) \beta_i F(\mathbf{x}; \boldsymbol{\nu}_i)$$

Tato konvexní kombinace náleží do množiny $\mathcal{H}$, neboť $\lambda\alpha_i, (1 - \lambda)\beta_i \geq 0$ a $\lambda \sum_{i=1}^{m} \alpha_i + (1 - \lambda) \sum_{i=1}^{n} \beta_i = 1$. Z toho plyne, že $\mathcal{H}$ je konvexní množina.

V dalším kroku je nutné dokázat, že $\mathcal{H}$ je konvexním obalem množiny $\mathcal{F}$, tedy že $\mathcal{H}$ je nejmenší konvexní množina, která obsahuje $\mathcal{F}$. Nechť $\mathbf{Y}$ je libovolná konvexní množina taková, že $\mathcal{F} \subseteq \mathbf{Y}$. Jelikož je $\mathbf{Y}$ konvexní množina, tak platí, že každá konvexní kombinace prvků z $\mathbf{Y}$ je také prvkem $\mathbf{Y}$, tudíž množina $\mathcal{H}$, jakožto množina tvořená konvexními kombinacemi prvků z $\mathcal{F}$, která navíc obsahuje restrikce vah prvků, musí být podmnožinou takové množiny $\mathbf{Y}$. Protože $\mathbf{Y}$ je zcela libovolná,

$$\text{conv}(\mathcal{F}) = \bigcap_{Y \supseteq \mathcal{F}} \mathbf{Y} \supseteq \mathcal{H}, \tag{5}$$

kde $\text{conv}(\mathcal{F})$ značí konvexní obal množiny $\mathcal{F}$. Ze vztahu 5 plyne, že $\text{conv}(\mathcal{F}) \supseteq \mathcal{H}$. Avšak pokud je $\text{conv}(\mathcal{F})$ nejmenší konvexní množina obsahující $\mathcal{F}$ a pokud zároveň víme, že $\mathcal{H}$ je konvexní, potom zřejmě $\text{conv}(\mathcal{F}) \subseteq \mathcal{H}$. Dohromady tedy dostaneme [15]

$$\mathcal{H} = \text{conv}(\mathcal{F}).$$

□

Věta 2. *(Yakowitz-Spraginsova) Třída konečných smíšených distribucí $\mathcal{H}$ je identifikovatelná, právě pokud je $\mathcal{F}$ lineárně nezávislou množinou, tj. pokud jsou všechny její prvky lineárně nezávislé na oboru reálných čísel $\mathbb{R}$. Tato podmínka je nutná a zároveň postačující.*

Důkaz. Nejprve dokážeme nutnost této podmínky. Předpokládejme, že $\mathcal{F}$ není lineárně nezávislá množina na $\mathbb{R}$. Potom pro nějaké $k > 0$ existuje nulová kombinace různých členů z $\mathcal{F}$ taková, že pro libovolné $m, 0 < m < k, \sum_{h=1}^{k} \vartheta_h F_h = 0$, platí

$$\vartheta_h < 0 \qquad \text{pro} \qquad h \leq m$$

$$\vartheta_h > 0 \qquad \text{pro} \qquad h > m.$$

Potom

$$\sum_{h=1}^{m} |\vartheta_h| F_h = \sum_{h=m+1}^{k} |\vartheta_h| F_h. \tag{6}$$

Na základě výše uvedeného vztahu a faktu, že $\{F_h\}$ jsou distribuční funkce, platí

$$\sum_{h=1}^{m} |\vartheta_h| = \sum_{h=m+1}^{k} |\vartheta_h| = b,$$

pokud uvažujeme nastavení všech F_h na hodnotu 1. Jestliže dále zadefinujeme

$$\lambda_h = \frac{|\vartheta_h|}{b} \qquad \text{pro} \qquad h = 1, \ldots, k,$$

můžeme vztah (6) modifikovat na

$$\sum_{h=1}^{m} |\lambda_h| F_h = \sum_{h=m+1}^{k} |\lambda_h| F_h.$$

Nyní tedy máme k dispozici dvě rozdílné reprezentace stejné konečné smíšené distribuce, což implikuje, že $\mathcal{H}$ není identifikovatelná.

V dalším kroku provedeme důkaz, že se jedná o postačující podmínku. Jelikož je $\mathcal{F}$ lineárně nezávislá množina na $\mathbb{R}$, tvoří také bázi pro lineární obal množiny

$\mathcal{F}$, $\langle\mathcal{F}\rangle$. Potom má každý člen z $\langle\mathcal{F}\rangle$ unikátní reprezentaci jako lineární kombinace složek z $\mathcal{F}$. Identifikovatelnost $\mathcal{H}$ je přímo odvoditelná ze skutečnosti, že $\mathcal{H} \subset \langle\mathcal{F}\rangle$. [16] □

2.2 Metoda maximální věrohodnosti

Metoda maximální věrohodnosti (zkráceně MLE z anglického Maximum Likelihood Estimation) je univerzální metoda pro konstrukci odhadů parametrů. Používáme ji pro odhad parametrů, které určují rozdělení pravděpodobností, např. střední hodnota a rozptyl normálního rozdělení, pravděpodobnost nastoupení jevu alternativního rozdělení apod. Princip metody maximální věrohodnosti spočívá v tom, že za odhad neznámého parametru vezmeme tu jeho hodnotu, ve které tzv. věrohodnostní funkce nabývá svého maxima.

Podívejme se nejprve na tuto metodu z obecného hlediska. Mějme náhodný výběr $\mathbf{X} = (X_1, \ldots, X_n)^{\mathrm{T}}$ pořízený z rozdělení, které je charakterizováno hustotou $f(\mathbf{x}, \boldsymbol{\theta})$. Toto rozdělení závisí na vektoru neznámých parametrů $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^{\mathrm{T}}$, přičemž $\boldsymbol{\theta}$ náleží do parametrického prostoru Θ, který je podmnožinou p-rozměrného prostoru reálných čísel $\mathbb{R}^p$. Protože složky náhodného výběru jsou nezávislé náhodné veličiny, sdružená hustota náhodného vektoru $\mathbf{X}$ je dána předpisem

$$f(x_1, x_2, \ldots, x_n; \boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i, \boldsymbol{\theta}).$$

Funkci $f(\mathbf{x}; \boldsymbol{\theta})$ proměnné $\boldsymbol{\theta}$ pro pevné $\mathbf{x}$ nazveme věrohodnostní funkcí a označíme ji $L(\boldsymbol{\theta})$,

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i, \boldsymbol{\theta}).$$

Analogicky v případě diskrétního rozdělení náhodného výběru charakterizovaného pravděpodobnostmi $p(x_i; \boldsymbol{\theta}), i = 1, 2, \ldots, n$, definujeme věrohodnostní funkci jako

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} p(x_i, \boldsymbol{\theta}).$$

Vektor $\widehat{\boldsymbol{\theta}} \in \Theta$ nazveme maximálně věrohodným odhadem parametru $\boldsymbol{\theta}$, jestliže pro jeho věrohodnostní funkci platí [17]

$$L(\widehat{\boldsymbol{\theta}}) \geq L(\boldsymbol{\theta}) \quad \text{pro každé } \boldsymbol{\theta} \in \Theta.$$

Tento obecný postup lze snadno aplikovat na případ odhadu parametrů smíšené distribuce, pouze zaměníme náhodný vektor $\mathbf{X}$ za vektor $\mathbf{Y}$, abychom dostáli značení z předchozích kapitol, a jako vektor odhadovaných parametrů budeme uvažovat $\boldsymbol{\Psi}$. Stále pracujeme s napozorovanými daty $\mathbf{y}= (\mathbf{y}_1, \ldots, \mathbf{y}_n)^{\mathrm{T}}$ a se smíšeným modelem

$$f(\mathbf{y_j}, \boldsymbol{\Psi}) = \sum_{i=1}^{c} \pi_i f_i(\mathbf{y_j}; \boldsymbol{\theta}_i).$$

Přičemž

$$\boldsymbol{\Psi} = (\pi_1, \ldots, \pi_{c-1}, \boldsymbol{\xi}^{\mathrm{T}})^{\mathrm{T}}$$

je vektor obsahující všechny neznámé parametry, kde $\pi_1, \cdots, \pi_c$ jsou váhy složek smíšené distribuce a vektor $\boldsymbol{\xi}$ obsahuje všechny parametry $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_c$.

Aplikací metody maximální věrohodnosti na tento model se dostáváme k úloze maximalizace rovnice

$$L(\boldsymbol{\Psi}) = \prod_{j=1}^{n} f(\mathbf{y}_j, \boldsymbol{\Psi})$$

pro pozorování $\mathbf{y}_1, \ldots, \mathbf{y}_n$. Tento vztah můžeme přepsat v logaritmické podobě jako

$$\ln L(\boldsymbol{\Psi}) = \sum_{j=1}^{n} \ln f(\mathbf{y}_j, \boldsymbol{\Psi}) = \sum_{j=1}^{n} \ln \sum_{i=1}^{c} \pi_i f_i(\mathbf{y}_j, \boldsymbol{\theta}_i).$$

Maximálně věrohodný odhad vektoru $\boldsymbol{\Psi}$ získáme řešením rovnice

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

či ekvivalentně

$$\partial \ln L(\boldsymbol{\Psi}) / \partial \boldsymbol{\Psi} = 0. \tag{7}$$

Podmíněnou pravděpodobnost, že j-té pozorování náleží do i-té složky smíšené distribuce, to vše za podmínky, že máme k dispozici vektor napozorovaných hodnot $\mathbf{y}$, vyjádříme jako

$$\tau_i(\mathbf{y}_j, \boldsymbol{\Psi}) = P\{Z_{ij} = 1|\mathbf{y}\} = \frac{\pi_i f_i(\mathbf{y}_j, \boldsymbol{\theta}_i)}{\sum_{h=1}^{c} \pi_h f_h(\mathbf{y}_j, \boldsymbol{\theta}_h)}. \tag{8}$$

Pomocí $\widehat{\boldsymbol{\Psi}}$, maximálně věrohodného odhadu množiny neznámých parametrů $\boldsymbol{\Psi}$, můžeme odhadnout váhy složek

$$\widehat{\pi}_i = \frac{\sum_{j=1}^{n} \tau_i(\mathbf{y}_j, \widehat{\boldsymbol{\Psi}})}{n}, \qquad i = 1, \ldots, c, \tag{9}$$

Jednotlivé odhady proporcí tedy vzniknou jako průměr ze všech n hodnot $\tau_i(\mathbf{y}_j, \widehat{\boldsymbol{\Psi}})$, jinak řečeno jako průměr přes všechna j ze všech podmíněných pravděpodobností $P\{Z_{ij} = 1|\mathbf{y}\}$.

Dále zavedeme následující restrikci:

$$\sum_{i=1}^{c} \sum_{j=1}^{n} \tau_i(\mathbf{y}_j, \widehat{\boldsymbol{\Psi}}) \frac{\partial \ln f_i(\mathbf{y}_j, \widehat{\boldsymbol{\theta}}_i)}{\partial \boldsymbol{\xi}} = 0. \tag{10}$$

Tento vztah zjednodušuje rovnici (7). Místo celé logaritmické věrohodnostní funkce derivujeme pouze logaritmické hustoty jednotlivých složek a to podle vektoru $\boldsymbol{\xi} = (\boldsymbol{\theta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\theta}_c^{\mathrm{T}})$. Tuto derivaci musíme následně vynásobit pravděpodobností ze vztahu (8), jakožto potřebnou váhou, s jakou bude derivace obsažena ve výsledném součtu. Ten probíhá přes všechny složky a přes všechna pozorování.

Vztahy (9) a (10) byly v minulosti zavedeny hned několika matematiky ve snaze vyřešit rovnici (7) a slouží jako základ k iteračnímu výpočtu, který nazýváme EM algoritmus.

Vzhledem k tomu, že na data nahlížíme jako na nekompletní (viz kapitola 1.3), musíme tomu přizpůsobit i věrohodnostní funkci. Připomeňme si, že z_{ij} je realizace složek Z_{ij} náhodného vektoru $\mathbf{Z}_j$. Hodnoty z_{ij} mohou nabývat hodnot

0 či 1 dle toho, zda pozorování $\mathbf{y}_j$ pochází ze složky i. Potom

$$\ln L_c(\mathbf{\Psi}) = \sum_{i=1}^{c}\sum_{j=1}^{n} z_{ij}\ \ln\{\pi_i f_i(\mathbf{y}_j;\boldsymbol{\theta}_i)\} \tag{11}$$

$$= \sum_{i=1}^{c}\sum_{j=1}^{n} z_{ij}\{\ln \pi_i + \ln f_i(\mathbf{y}_j;\boldsymbol{\theta}_i)\}$$

je logaritmická věrohodnostní funkce pro $\mathbf{\Psi}$ za předpokladu kompletních dat.

2.3 EM algoritmus

EM algoritmus (z anglického Expectation-Maximization algorithm) je v praxi velmi používanou metodou odhadování parametrů smíšených distribucí. Jedná se o iterační algoritmus sloužící k nalezení maximálně věrohodných odhadů. Tato metoda je užitečná především proto, že jednodušší metody často nelze efektivně použít. O historii této metody jsme se stručně zmínili v kapitole 2.1.

EM algoritmus je dvoukrokový proces skládající se z E-kroku a M-kroku. Jednotlivé kroky obnášejí:

- **E-krok**: Spočítáme aposteriorní pravděpodobnost, že k-té pozorování náleží do i-té složky, která je dána aktuálními parametry. (Expectation)
- **M-krok**: Aktualizujeme odhady parametrů pomocí maximalizace očekávané logaritmické věrohodnostní funkce. (Maximalization)

Nyní se na oba kroky podíváme blíže. [10]

2.3.1 E-krok

Při použití EM algoritmu na z_{ij} nahlížíme jako na chybějící data a jejich zahrnutí do výpočtu provedeme v E-kroku. Ten pracuje s podmíněnou očekávanou hodnotou logaritmické věrohodnostní funkce pro kompletní data, uvažující napozorovaná data $\mathbf{y}$ a současný odhad pro $\mathbf{\Psi}$. Proto je nutné specifikovat $\widehat{\mathbf{\Psi}}^{(0)}$ jako počáteční řešení (viz kapitola (2.5)).

V prvním kroku iteračního EM algoritmu provedeme v E-kroku výpočet očekávané logaritmické věrohodnostní funkce, $\ln L_c(\mathbf{\Psi})$, za podmínky, že máme k dispozici data $\mathbf{y}$, a použijeme počáteční odhad $\widehat{\mathbf{\Psi}}^{(0)}$, což můžeme zapsat jako

$$Q(\mathbf{\Psi}, \widehat{\mathbf{\Psi}}^{(0)}) = E_{\widehat{\mathbf{\Psi}}^{(0)}}\{\ln L_c(\mathbf{\Psi})|\mathbf{y}\}.$$

Stejně tak pro *(k+1)*-ní iteraci platí

$$Q(\mathbf{\Psi}, \widehat{\mathbf{\Psi}}^{(k)}) = E_{\widehat{\mathbf{\Psi}}^{(k)}}\{\ln L_c(\mathbf{\Psi})|\mathbf{y}\}.$$

Očekávaná hodnota E je indexována vektorem $\widehat{\mathbf{\Psi}}^{(k)}$, aby bylo zřejmé, že místo vektoru $\mathbf{\Psi}$ využívá jeho odhad $\widehat{\mathbf{\Psi}}^{(k)}$ získaný v k-té iteraci. Víme-li, že $\ln L_c(\mathbf{\Psi})$, logaritmická věrohodnostní funkce kompletních dat, je lineární v nepozorovatelných datech z_{ij}, viz vztah (11), potom zřejmě E-krok v (k+1)-ní iteraci vyžaduje výpočet současné očekávané hodnoty náhodné veličiny Z_{ij} podmíněné tím, že máme k dispozici napozorovaná data $\mathbf{y}$. To zapíšeme jako

$$\begin{aligned} E_{\widehat{\mathbf{\Psi}}^{(k)}}(Z_{ij}|\mathbf{y}) &= 1 \cdot P_{\widehat{\mathbf{\Psi}}^{(k)}}\{Z_{ij} = 1|\mathbf{y}\} + 0 \cdot P_{\widehat{\mathbf{\Psi}}^{(k)}}\{Z_{ij} = 0|\mathbf{y}\} \\ &= P_{\widehat{\mathbf{\Psi}}^{(k)}}\{Z_{ij} = 1|\mathbf{y}\} = \tau_i(\mathbf{y}_j, \widehat{\mathbf{\Psi}}^{(k)}), \end{aligned}$$

kde zohledníme vztah (8), takže

$$\tau_i(\mathbf{y}_j, \widehat{\mathbf{\Psi}}^{(k)}) = \frac{\widehat{\pi}_i^{(k)} f_i(\mathbf{y}_j, \widehat{\boldsymbol{\theta}}_i^{(k)})}{f(\mathbf{y}_j, \widehat{\mathbf{\Psi}}^{(k)})} = \frac{\widehat{\pi}_i^{(k)} f_i(\mathbf{y}_j, \widehat{\boldsymbol{\theta}}_i^{(k)})}{\sum_{h=1}^{c} \widehat{\pi}_h^{(k)} f_h(\mathbf{y}_j, \widehat{\boldsymbol{\theta}}_h^{(k)})}$$

$$\text{pro} \qquad i = 1, \ldots, c \quad a \quad j = 1, \ldots, n.$$

S touto znalostí již můžeme $Q(\mathbf{\Psi}, \widehat{\mathbf{\Psi}}^{(k)})$ přepsat pomocí vztahu (11) jako [13]

$$Q(\mathbf{\Psi}, \widehat{\mathbf{\Psi}}^{(k)}) = \sum_{i=1}^{c} \sum_{j=1}^{n} \tau_i(\mathbf{y}_j, \widehat{\mathbf{\Psi}}^{(k)})\{\ln \pi_i + \ln f_i(\mathbf{y}_j; \boldsymbol{\theta}_i)\}.$$

2.3.2 M-krok

V M-kroku v $(k+1)$-ní iteraci provedeme maximalizaci $Q(\boldsymbol{\Psi}, \widehat{\boldsymbol{\Psi}}^{(k)})$ přes všechna $\boldsymbol{\Psi}$ z parametrického prostoru $\boldsymbol{\Omega}$. Toho dosáhneme zderivováním funkce $Q(\boldsymbol{\Psi}, \widehat{\boldsymbol{\Psi}}^{(k)})$, tj.

$$\frac{\partial \sum_{i=1}^{c} \sum_{j=1}^{n} \tau_i(\mathbf{y}_j, \widehat{\boldsymbol{\Psi}}^{(k)}) \ln\{\pi_i f_i(\mathbf{y}_j; \boldsymbol{\theta}_i)\}}{\partial \boldsymbol{\Psi}}.$$

Tato derivace je položena rovna nule a je spočítáno maximum. Tím zaktualizujeme odhad $\boldsymbol{\Psi}$ a získáme $\widehat{\boldsymbol{\Psi}}^{(k+1)}$.

Kdyby z_{ij} byly pozorovatelné hodnoty, bylo by možné maximálně věrohodný odhad váhy π_i psát jako

$$\widehat{\pi}_i = \sum_{j=1}^{n} \frac{z_{ij}}{n}, \qquad i = 1, \ldots, c. \tag{12}$$

Avšak v E-kroku dochází k nahrazení z_{ij} očekávanou hodnotou $\tau_i(\mathbf{y}_j, \widehat{\boldsymbol{\Psi}}^{(k)})$ v logaritmické věrohodnostní funkci kompletních dat, tudíž aktualizovaný odhad π_i vznikne analogicky nahrazením z_{ij} očekávanou hodnotou $\tau_i(\mathbf{y}_j, \widehat{\boldsymbol{\Psi}}^{(k)})$ ve vztahu (12). Ve výsledku dostáváme

$$\widehat{\pi}_i^{(k+1)} = \sum_{j=1}^{n} \frac{\tau_i(\mathbf{y}_j, \widehat{\boldsymbol{\Psi}}^{(k)})}{n}, \qquad i = 1, \ldots, c.$$

E-krok a M-krok se opakují, dokud se rozdíl mezi hodnotami věrohodnostních funkcí v k-té a $(k+1)$-ní iteraci

$$L(\widehat{\boldsymbol{\Psi}}^{(k+1)}) - L(\widehat{\boldsymbol{\Psi}}^{(k)})$$

nezmění pouze o zanedbatelnou hodnotu. Tento rozdíl bude klesat s každou novou iterací, protože algoritmus je konvergentní. Stejně tak bude platit, že hodnota věrohodnostní funkce roste v každém kroku, tedy

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

[13]

2.4 Počet komponent

V některých modelech je počet komponent předem znám, nicméně pokud tomu tak není, je třeba ho odvodit z dat společně s dalšími neznámými parametry. Za tímto účelem bylo navrženo mnoho postupů, v této sekci si pouze stručně představíme hlavní kategorie metod a jejich teoretické pozadí. Skupiny metod jsou následující:

1. Testy hypotéz
2. Informační kritéria
3. Klasifikační kritéria
4. Minimální informační poměrový ukazatel a přidružená kritéria
5. Ostatní kritéria

Intuitivní postup, jak rozhodnout o počtu složek v modelu, je testování hypotézy o tomto počtu. Obvykle testujeme nulovou hypotézu, že model se skládá z K složek, proti alternativě, že počet tříd je $K+1$. K tomu se většinou používá obecně známý test poměrem věrohodností. V rámci tohoto testu je využívána metoda Monte Carlo, která vyžaduje použití bootstrappingu k získání odhadu p-hodnoty. Metoda Monte Carlo je vhodná především proto, že smíšené distribuce zcela nesplňují podmínky regularity.

Přirozeně, jak počet složek v modelu roste, hodnota věrohodnostní funkce roste též. Testy využívající věrohodnostní funkce proto mají tendenci upřednostňovat přehnaně komplikované modely. To lze vykompenzovat zahrnutím nějakého druhu penalizace k logaritmické věrohodnostní funkci. Penalizace poroste se stoupajícím počtem složek, aby byla vybalancována věrohodnost odhadu proti počtu

odhadovaných parametrů. Tento přístup k odhadu správného počtu složek smíšené distribuce řeší informační kritéria, mezi která řadíme například Akaikeho informační kritérium, Takeuchiho informační kritérium či Bayesovské informační kritérium (k dispozici je samozřejmě i řada dalších informačních kritérií, ale uvedené příklady patří mezi ty nejznámější).

Zatímco informační kritéria se snaží zabránit přeurčení modelu, klasifikační kritéria se snaží zajistit, aby složky smíšené distribuce byly dostatečně odděleny, tedy aby se jedna od druhé dostatečně lišily. K ohodnocení dostatečné separace složek se dají využít odhady vah jednotlivých složek. Mezi tato kritéria řadíme například entropické kritérium či klasifikační věrohodnostní kritérium.

Minimální informační poměrové ukazatele jsou inspirovány vlastnostmi matice informačních poměrových ukazatelů. Tato kritéria mohou být vypočítána jako funkce míry konvergence EM algoritmu.

Poslední skupina kritérií obsahuje všechny metody, které nemohou být zařazeny do žádné z výše zmíněných kategorií. Řadí se sem například grafické metody či metody využívající derivace funkce.

Jelikož představení a popis všech možných metod odhadujících počet složek by byl velmi náročný a zcela nad rámec této práce, omezíme se pouze na kritérium, které bude použité v praktické části, což je Bayesovské informační kritérium (BIC), a na test poměrem věrohodností, jakožto obecně známé testovací kritérium. [14]

Test poměrem věrohodnosti s využitím bootstrappingu

Nejznámějším přístupem k rozhodování o počtu složek ve smíšené distribuci je stanovit hypotézu a ověřit její platnost pomocí poměru věrohodností. Mějme nulovou hypotézu H_0 uvažující s_0 složek a proti ní postavme alternativní hypotézu H_1 s s_1 složkami, tedy

$$H_0 : S = s_0, \qquad H_1 : S = s_1, \qquad \text{kde } s_1 > s_0, s_1 = s_0 + 1.$$

Test poměrem věrohodností je při splnění podmínek regularity vhodným k porovnání těchto dvou modelů a zadefinujeme ho následujícím způsobem

$$-2\ln\lambda = -2\ln\left[\frac{L(\widehat{\mathbf{\Psi}}_{s_0})}{L(\widehat{\mathbf{\Psi}}_{s_1})}\right] = 2\big\{\ln L(\widehat{\mathbf{\Psi}}_{s_1}) - \ln L(\widehat{\mathbf{\Psi}}_{s_0})\big\}. \tag{13}$$

Pokud nulovou hypotézu zamítáme, je běžným postupem pokračovat v testování a přidávat pokaždé o jednu komponentu víc, dokud změna ve věrohodnosti nebude tak malá, abychom mohli zamítnout alternativní hypotézu.

Na smíšené distribuce ovšem nemůžeme klasický test poměrem věrohodnosti uplatnit, protože vztah (13) nesplňuje klasické podmínky regularity týkající se asymptotických vlastností maximálně věrohodných odhadů (viz [1]). Hlavním problémem je zde fakt, že za platnosti H_0 pracujeme s neidentifikovatelnou smíšenou distribucí v parametrickém prostoru $\mathbf{\Omega}_{s_1}$. Proto nelze, jako ve většině běžných situací, předpokládat, že má testová statistika při platnosti nulové hypotézy asymptoticky χ^2 rozdělení pravděpodobnosti, jehož stupně volnosti jsou rovny rozdílu mezi počty odhadovaných parametrů při nulové a alternativní hypotéze. Neidentifikovatelnost v tomto případě způsobuje, že za předpokladu nulové hypotézy je nemožné určit asymptotické rozdělení maximálně věrohodného odhadu. Zároveň vede k degeneraci informační matice, jestliže pracujeme s asymptotickým rozdělením logaritmické věrohodnostní funkce vytvořené za platnosti alternativní hypotézy. [13, 14]

Vhodnou metodou, jak se vypořádat s tímto nedostatkem, je možnost bootstrappingu. Postup pro tuto metodu v tomto kontextu je následující. Provedeme odhad parametrů smíšené distribuce za platnosti nulové i alternativní hypotézy, tedy pro model s počtem složek s_0 a s_1. Bootstrappingový vzorek je vygenerován ze smíšené distribuce odhadnuté za nulové hypotézy, jinak řečeno, výběr je pořízen za použití vektoru $\widehat{\mathbf{\Psi}}_{s_0}$, maximálně věrohodného odhadu vektoru $\mathbf{\Psi}_{s_0}$. Analogicky postupujeme pro model odhadnutý za platnosti alternativní hypotézy. Následně je pro každou dvojici bootstrappingových vzorků (jeden uvažující

model o s_0 složkách a druhý pořízený za předpokladu s_1-složkového modelu) spočítána hodnota $-2\ln\lambda$. Proces se B-krát nezávisle opakuje, dokud nezískáme dostatečný počet hodnot $-2\ln\lambda$ k odhadnutí rozdělení pravděpodobností této statistiky. Jakmile odhadneme parametry tohoto rozdělení, můžeme již spočítat p-hodnotu a vyhodnotit celý test poměrem věrohodností. [13]

BIC

Bayesovské informační kritérium je hlavní kritérium založené na bayesovském přístupu, který upřednostňuje model s nejvyšší podmíněnou pravděpodobností. Pokud uvažujeme S možných modelů, podmíněná pravděpodobnost modelu M_s $(s = 1, \ldots, S)$, pokud máme k dispozici vektor pozorování $\mathbf{y}$, bude na základě Bayesovy věty následující:

$$P(M_s|\mathbf{y}) = \frac{P(\mathbf{y}|M_s)P(M_s)}{\sum_{s=1}^{S} P(\mathbf{y}|M_s)P(M_s)} \qquad \text{pro} \quad s = 1, \ldots, S.$$

Podmíněnou pravděpodobností $P(\mathbf{y}|M_s)$ je v tomto kontextu myšlena věrohodnost modelu M_s a pravděpodobnost $P(M_s)$ značí celkovou pravděpodobnost, že model bude vybrán, kterou v tomto případě vnímáme pro všechny modely jako ekvivalentní.

Bayesovské kritérium bylo navržené Gideonem E. Schwarzem v roce 1978, který původně použil aproximaci věrohodnosti modelu a nazval jí Schwarzovým kritériem, jehož podoba je:

$$\ln P(\mathbf{y}|M_s) \approx \ln L(M_s) - \frac{k}{2}\ln(n),$$

za podmínky, že k je počet parametrů a n je počet pozorování. Kass a Raftery v roce 1995 vynásobili pravou stranu mínus dvěma a výsledek označili jako Bayesovské informační kritérium,

$$\mathrm{BIC} = k\ln(n) - 2\ln L(M_s).$$

Jako vhodný model zvolíme ten s nejmenší hodnotou BIC. [14]

2.5 Počáteční řešení

Jelikož je EM algoritmus iterační procedura, na jejím začátku je nutné zadat počáteční odhad vektoru $\boldsymbol{\Psi}$, který označíme $\widehat{\boldsymbol{\Psi}}^{(0)}$. Tento počáteční odhad může společně s ukončovacími kritérii silně ovlivnit výsledné odhady parametrů. Také víme, že konvergence EM algoritmu je poměrně pomalá, přičemž při špatné volbě startovacích odhadů se doba konvergence celé procedury ještě více prodlužuje. Zároveň se může stát, že na okrajích parametrického prostoru je věrohodnostní funkce neomezená a posloupnost odhadů pořízených EM algoritmem $\{\widehat{\boldsymbol{\Psi}}^{(k)}\}$ diverguje, pokud je počáteční odhad zvolen příliš blízko těmto okrajům.

Další závažný problém představuje věrohodnostní rovnice, která má obvykle více než jeden kořen odpovídající lokálnímu maximu. Pokud nemáme k dispozici žádnou napozorovanou hodnotu či konzistentní odhad vektoru $\boldsymbol{\Psi}$, logicky vybíráme ten kořen věrohodnostní rovnice, který odpovídá největšímu lokálnímu maximu. Ani tento postup však nezaručuje, že výsledná posloupnost kořenů bude konzistentní a asymptoticky eficientní.

Při použití EM algoritmu v každém E-kroku provádíme aktualizaci podmíněné pravděpodobnosti $\boldsymbol{\tau}(\mathbf{y}_j, \boldsymbol{\Psi})$ toho, že pozorování náleží do dané složky smíšené distribuce. Pro první E-krok v algoritmu je tedy nutné specifikovat všechny počáteční hodnoty $\tau_j^{(0)}$ pro $\tau(\mathbf{y}_j, \boldsymbol{\Psi})$, kde

$$\boldsymbol{\tau}(\mathbf{y}_j, \boldsymbol{\Psi}) = (\tau_1(\mathbf{y}_j, \boldsymbol{\Psi}), \ldots, \tau_c(\mathbf{y}_j, \boldsymbol{\Psi}))$$

je vektor obsahující všech c podmíněných pravděpodobností udávajících náležitost k jednotlivým složkám pro napozorovaný vektor $\mathbf{y}_j$. [13]

2.5.1 Náhodné počáteční odhady

Jedním ze způsobů, jak určit počáteční odhady $\widehat{\mathbf{z}}^{(0)}$, je náhodné rozřazení dat do c skupin odpovídajících c složkám smíšené distribuce. To můžeme provést tak, že pro každé $\mathbf{y}_j$ vygenerujeme číslo od 1 do c, včetně okrajových hodnot.

Jestliže takto získáme hodnotu h, následně nastavíme i-tou složku vektoru $\widehat{\mathbf{z}}_j^{(0)}$ rovnu jedné pro $i = h$, pro ostatní hodnoty i budou složky rovny nule.

Pokud EM algoritmus pracuje s těmito náhodnými startovacími hodnotami, centrální limitní věta zaručuje, že parametry všech složek budou podobné přinejmenším při velkém počtu pozorování. Jedním ze způsobů, jak tomuto efektu zabránit, je nejprve vybrat menší náhodný vzorek z dat, který náhodně rozdělíme do c skupin. Poté v rámci tohoto vzorku provedeme M-krok, avšak musíme dbát na to, aby byl vzorek pro tyto účely dostatečně velký. Pokud například chceme odhadnout parametry smíšené distribuce sestavené z p-dimenzionálních normálních rozdělení, musíme mít k dispozici nejméně $(p+1)$ pozorování ke každé složce, abychom se vyhnuli singularitě odhadů.

Alternativní cestou, jak určit náhodné startovací hodnoty, je vygenerovat přímo odhady parametrů jednotlivých rozdělení pravděpodobností, která ve smíšené distribuci vystupují. Uvažujme pro jednoduchost směs c normálních komponent se středními hodnotami $\boldsymbol{\mu}_i$ a variančními maticemi $\boldsymbol{\Sigma}_i$. Místo $\widehat{\mathbf{z}}^{(0)}$ budeme tedy generovat $\widehat{\boldsymbol{\mu}}_i^{(0)}$ jako

$$\widehat{\boldsymbol{\mu}}_i^{(0)}, \ldots, \widehat{\boldsymbol{\mu}}_c^{(0)} \sim N(\overline{\mathbf{y}}, \mathbf{V}),$$

kde $\overline{\mathbf{y}}$ je výběrový průměr a $\mathbf{V}$ je výběrová kovarianční matice. Použití této metody zaručuje větší rozdíly mezi počátečními odhady $\widehat{\boldsymbol{\mu}}_i^{(0)}$ pro střední hodnotu $\boldsymbol{\mu}_i$ a odhady pořízenými pomocí náhodného přiřazení dat do jedné z c skupin. Další výhodou je menší výpočetní náročnost.

Varianční matice jednotlivých komponent $\boldsymbol{\Sigma}_i$ určíme jako $\widehat{\boldsymbol{\Sigma}}_i^{(0)} = \mathbf{V}$ a váhy komponent $\widehat{\pi}_i^{(0)}$ pomocí vztahu

$$\widehat{\pi}_i^{(0)} = \frac{1}{c} \qquad \text{pro} \quad 1, \ldots, c.$$

Klíčovým faktorem se ukázal být co nejpřesnější odhad vah složek smíšené distribuce. Například pro jednorozměrné smíšené distribuce Charless Fowlkes navrhl využití Q-Q grafu k zjištění jejich počátečního odhadu. Ostatní parametry mo-

hou být potom jednoduše odhadnuty rozdělením dat do tříd na základě odhadů vah. [14]

2.5.2 Deterministický žíhaný EM algoritmus

Jak již bylo řečeno dříve, EM algoritmus velmi trpí kvůli problému nalezení nejlepšího lokálního maxima. Špatně zvolené počáteční odhady můžou způsobit, že algoritmus bude konvergovat pouze do lokálního, nikoli globálního maxima či do sedlového bodu věrohodnostní funkce. Tomu se dá zabránit například tím, že EM algoritmus opakujeme několikrát s náhodnou volbou startovních hodnot a jako výsledek vezmeme ty odhady, které nejvíce svědčí pro to, že jsme nalezli globální maximum. Kromě tohoto postupu byl navržen také deterministický žíhaný EM algoritmus (DAEM z Deterministic Annealing EM algorithm), který místo maximalizace logaritmické věrohodnostní funkce řeší minimalizaci termodynamické volné energie definovanou principem maximální entropie a analogií se statistickou mechanikou.

Zatímco klasický EM algoritmus maximalizuje v každém M-kroku funkci $Q(\boldsymbol{\Psi}, \widehat{\boldsymbol{\Psi}}^{(k)})$, DAEM se pokouší řešit minimalizaci funkce volné energie, kterou zapíšeme následovně:

$$F_\beta(\boldsymbol{\Psi}) = -\frac{1}{\beta} \ln \sum_{\chi_{ch}} P(\chi_p, \chi_{ch}|\boldsymbol{\Psi})^\beta,$$

kde χ_{ch} jsou nepozorovatelná (tedy chybějící) data a χ_p napozorovaná data. V podstatě se jedná o nákladovou funkci závisící na parametru $\frac{1}{\beta}$, který nazýváme *teplota*.

V kontextu smíšených distribucí se DAEM projeví tím, že dosud používaný odhad podmíněné pravděpodobnosti $\tau_i(\mathbf{y}_j, \widehat{\boldsymbol{\Psi}}^{(k)})$, že pozorování $\mathbf{y}_j$ náleží do i-té skupiny, nahradíme výrazem [13]

$$\frac{\left\{\tau_i(\mathbf{y}_j, \widehat{\boldsymbol{\Psi}}^{(k)})\right\}^\beta}{\sum_{h=1}^{c} \left\{\tau_h(\mathbf{y}_j, \widehat{\boldsymbol{\Psi}}^{(k)})\right\}^\beta} \qquad \text{pro} \quad i = 1, \ldots, c.$$

Doporučuje se začít s hodnotou β blízkou nule, přičemž $0 \leq \beta \leq 1$, a později ji zvyšovat po každé iteraci podle vzorce

$$\beta^{(k+1)} = d\beta^{(k)}.$$

Koeficient d se pohybuje v intervalu $\langle 1, 1.5 \rangle$, dokud $\beta^{(k+1)} = 1$.

Pro β blízké nule jsou efekty $\mathbf{y}_j$ téměř totožné, tudíž počáteční odhady podmíněných pravděpodobností určujících náležitost k jednotlivým složkám smíšené distribuce jsou blízké hodnotě $\frac{1}{d}$. Jak hodnota parametru β roste k jedné, efekty $\mathbf{y}_j$ se postupně lokalizují. Ueda a Nakano (1998) argumentují, že tímto postupem je algoritmus schopen potlačit vliv špatně zvolených počátečních odhadů, které mohou být velmi vzdálené pravým hodnotám. [13, 19]

2.5.3 Stochastický EM algoritmus

Stochastický EM algoritmus (zkráceně SEM z anglického Stochastic EM algorithm) je modifikovanou verzí EM algoritmu. Stejně jako DAEM byl navržen proto, aby dokázal zabránit konvergenci EM algoritmu ke špatnému lokálnímu maximu, což je obvykle způsobenou špatnou volbou počátečních odhadů, na kterou je EM algoritmus velmi citlivý.

Stochastický EM algoritmus vychází z Monte Carlo EM algoritmu, který je obohacen o S-krok, kde jsou P-krát nasimulována chybějící data. V následujícím M-kroku dochází ke klasické maximalizaci věrohodnostní funkce, ve které teď budou zahrnuta nejen napozorovaná data, ale též nasimulované chybějící hodnoty $\boldsymbol{z}$. Stochastický EM algoritmus, který byl poprvé navržen statistiky Celeuxem a Dieboltem v roce 1985, pracuje na podobném principu, avšak místo P simulací je v S-kroku provedena simulace jediná. To provedeme tak, že provedeme náhodný výběr $\widehat{\mathbf{z}}_j^{(k)}$ pro každé $j, j = 1, \ldots, n$, z multinomického rozdělení o c kategoriích s pravděpodobnostmi

$$\boldsymbol{\tau}(\mathbf{y}_j, \widehat{\boldsymbol{\Psi}}^{(k)}) = (\tau_1(\mathbf{y}_j, \widehat{\boldsymbol{\Psi}}^{(k)}), \ldots, \tau_c(\mathbf{y}_j, \widehat{\boldsymbol{\Psi}}^{(k)}))^{\mathrm{T}}. \tag{14}$$

Tímto postupem přiřadíme každé pozorování $\mathbf{y}_j$ přímo do jedné z c složek smíšené distribuce místo toho, abychom jim pouze přiřazovali váhy příslušnosti ke složkám, využívajíc současných odhadů podmíněných pravděpodobností, které jsou dány vektorem (14).

Neoddiskutovatelnou výhodou stochastického EM algoritmu je možnost iteračního procesu vzpamatovat se z počátečních odhadů, jež vedou ke špatnému lokálnímu maximu, a začít konvergovat správným směrem. Na druhou stranu tento algoritmus není nejvhodnější ve chvíli, když se blížíme k vhodnému maximu, proto je výhodnější v pozdějších stádiích iteračního procesu začít opět používat obyčejný algoritmus EM. [13]

3 Směsi lineárních regresních modelů

3.1 Regresní analýza

Jelikož jsme již dostatečně popsali základní principy smíšených distribucí, můžeme se přesunout ke stěžejní částí této práce, kterou představují aplikace smíšených distribucí v regresní analýze.

Jednorozměrná regresní analýza umožňuje jejímu uživateli popsat závislost výstupní veličiny Y na vstupních nezávislých veličinách pomocí matematického vzorce, který popisuje Y jako funkci proměnných $x_1, \ldots, x_k$. Regresí rozumíme závislost mezi střední hodnotou jednorozměrné náhodné veličiny Y a proměnnou $\mathbf{x}$.

V této práci se omezíme na regresní analýzu opřenou o lineární regresní model, což znamená, že funkce závislosti f je sestavená jako lineární kombinace k funkcí

$$\mathrm{E}(Y|\mathbf{x}) = \beta_1 f_1(\mathbf{x}) + \ldots + \beta_k f_k(\mathbf{x}),$$

přičemž jednotlivé funkce $f_1(\mathbf{x}), \ldots, f_k(\mathbf{x})$ nemusí být lineární, avšak závislost funkce $f(\mathbf{x})$ na koeficientech $\beta_1, \ldots, \beta_k$ musí. Tyto koeficienty navíc rozšíříme o absolutní člen β_0, jehož zařazení do rovnice je matematicky výhodné a vyplývá z existence nezařazených a neuvažovaných vlivů.

Zcela lineárním modelem rozumíme

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon = \beta_0 + \sum_{j=1}^{k} \beta_j x_j + \varepsilon.$$

Podívejme se na tento model z více praktického hlediska a představme si, že experiment provedeme při n různých nastaveních $x_1, x_2, \ldots, x_k$. Tuto situaci můžeme přepsat pomocí vektoru $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ a nestochastické matice čísel $\mathbf{X}_{n\times(k+1)}$, kde každý řádek představuje jedno z n nastavení. Označme $k + 1 = p$ a navíc uvažujme plnou hodnost matice $\mathbf{X}$, $h(\mathbf{X}) = p$, a $n \geq p$. Tím je zaručeno, že matice $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ je symetrická regulární matice, ke které existuje inverzní matice a jejíž determinant je větší než nula.

Pokud se $\mathbf{Y}$ řídí lineárním modelem, pak

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

kde $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^{\mathrm{T}}$ je vektor neznámých parametrů a $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{T}}$ je vektor náhodných chyb.

Nejpoužívanější metodou k odhadu regresních parametrů je metoda nejmenších čtverců, která je také nejefektivnější metodou v případě normálně rozdělených dat.

Věta 3. *(Gauss-Markovova) Uvažujme následující podmínky:*

- $\mathrm{E}(\varepsilon_j) = 0 \quad pro \quad j = 1, 2, \ldots, n$
- $\mathrm{var}(\varepsilon_j) = \sigma^2 \quad pro \quad j = 1, 2, \ldots, n$
- $\mathrm{cov}(\varepsilon_j; \varepsilon_i) = 0,\ j \neq i$
- x_i *je nenáhodná konstanta*

Za těchto podmínek je odhad vektoru $\boldsymbol{\beta}$ pořízený metodou nejmenších čtverců nejlepší nestranný lineární odhad. [17]

3.2 Směsi lineárních regresních modelů

Metoda nejmenších čtverců je, jak potvrzuje i metoda maximální věrohodnosti, nejlepší volbou za podmínky, že jsou chyby měření normálně rozděleny. Pokud tomu tak ovšem není a můžeme u vektoru $\mathbf{Y}$ pozorovat podobnost se smíšenou distribucí, můžeme přistoupit k použití směsí regresních modelů. Cílem této metody je věrohodně popsat podmíněnou distribuci $\mathbf{Y}_j|\mathbf{X}_j$.

Směsi regresních modelů byly primárně zkoumány v kontextu ekonometrie a poprvé byly představeny maďarským ekonomem R.E. Quandtem jako "*switching regimes*" (také "*switching regression*"). Systém "*switching regimes*" je často přirovnáván ke strukturálním změnám, avšak strukturální změna v datech předpokládá, že systém je deterministicky závislý na nějaké pozorovatelné proměnné,

zatímco "*switching regimes*" vychází z faktu, že nejsme schopni určit příčinu změn mezi regresemi. Richard De Veaux přizpůsobil EM algoritmus pro směs dvou regresí. Jones, McLachlan a Turner tuto práci dále rozšiřovali, zatímco Hawkins se začal zabývat odhadem významných složek směsi regresních modelů.

Stejně jako v úvodu této kapitoly uvažujeme vektor $\mathbf{y}$, skládající se z n nezávislých pozorování $y_1, \ldots, y_n$, a designovou matici $\mathbf{X}_{n\times p}$. Dále předpokládejme, že každé pozorování $(y_j, \mathbf{x}_j)$ náleží do jedné z c tříd. Za podmínky, že pozorování náleží do i-té třídy, můžeme vztah mezi y_j a $\mathbf{x}_j$ zapsat jako klasický regresní model

$$Y_j = \mathbf{x}_j^{\mathrm{T}}\boldsymbol{\beta}_i + \varepsilon_{ij},$$

kde $\varepsilon_{ij} \sim \mathrm{N}(0, \sigma_i^2)$, $\boldsymbol{\beta}_i$ je p-dimenzionální vektor regresních koeficientů a σ_i^2 je rozptyl náhodných chyb i-té třídy. Všechny tyto podmínky platí, pokud mluvíme o i-té složce. Kompletní směs regresních modelů vypadá následovně:

$$Y_j = \begin{cases} \mathbf{x}_j^{\mathrm{T}}\boldsymbol{\beta}_1 + \varepsilon_{1j} & \text{s pravděpodobností } \pi_1 \\ \mathbf{x}_j^{\mathrm{T}}\boldsymbol{\beta}_2 + \varepsilon_{2j} & \text{s pravděpodobností } \pi_2 \\ \qquad\vdots & \\ \mathbf{x}_j^{\mathrm{T}}\boldsymbol{\beta}_c + \varepsilon_{cj} & \text{s pravděpodobností } \pi_c \end{cases},$$

Dodržením struktury směsí se dostaneme ke vztahu pro podmíněnou hustotu

$$f(y_j|\mathbf{x}_j, \boldsymbol{\Psi}) = \sum_{i=1}^{c} \pi_i \phi(y_j|\mathbf{x}_j^{\mathrm{T}}\boldsymbol{\beta}_i, \sigma_i^2). \tag{15}$$

Značením $\phi(\cdot|\mathbf{x}_j^{\mathrm{T}}\boldsymbol{\beta}_i, \sigma_i^2)$ myslíme hustotu normálního rozdělení se střední hodnotou $\mathbf{x}_j^{\mathrm{T}}\boldsymbol{\beta}_i$ a rozptylem σ_i^2. Při této znalosti můžeme smíšenou distribuci rozepsat jako

$$f(y_j|\mathbf{x}_j, \boldsymbol{\Psi}) = \sum_{i=1}^{c} \pi_i \left(2\pi\sigma_i^2\right)^{-\frac{1}{2}} \exp\left\{\frac{-(y_j - \mathbf{x}_j^{\mathrm{T}}\boldsymbol{\beta}_i)^2}{2\sigma_i^2}\right\}.$$

Vektor $\boldsymbol{\Psi}$ má v tomto případě podobu

$$\boldsymbol{\Psi} = (\pi_1, \ldots, \pi_c, (\boldsymbol{\beta}_1^{\mathrm{T}}, \sigma_1^2), \ldots, (\boldsymbol{\beta}_c^{\mathrm{T}}, \sigma_c^2))^{\mathrm{T}}.$$

Stejně jako u smíšených distribucí se budeme snažit maximalizovat logaritmickou věrohodnostní funkci [7, 2]

$$\ln L(\boldsymbol{\Psi}|\mathbf{x}_1,\ldots,\mathbf{x}_n,y_1,\ldots,y_n)=\sum_{j=1}^{n}\ln\Big(\sum_{i=1}^{c}\pi_i\phi_i(y_j|\mathbf{x}_j^{\mathrm{T}}\boldsymbol{\beta}_i,\sigma_i^2)\Big),$$

neboli

$$\ln L(\boldsymbol{\Psi}|\mathbf{x}_1,\ldots,\mathbf{x}_n,y_1,\ldots,y_n)=\sum_{j=1}^{n}\ln\left(\sum_{i=1}^{c}\pi_i\left(2\pi\sigma_i^2\right)^{-\frac{1}{2}}\exp\left\{\frac{-(y_j-\mathbf{x}_j^{\mathrm{T}}\boldsymbol{\beta}_i)^2}{2\sigma_i^2}\right\}\right).$$

Podobně jako ve směsi distribucí můžeme vyjádřit podmíněnou pravděpodobnost, že j-té pozorování patří do i-té složky. Tato bude obdobná jako ve vztahu (12), avšak hustotu $f_i(\mathbf{y}_j,\boldsymbol{\theta}_i)$ nahradíme hustotou normálního rozdělení pravděpodobností $\phi(y_j|\mathbf{x}_j^{\mathrm{T}}\boldsymbol{\beta}_i,\sigma_i^2)$. Tím se dostáváme ke vztahu

$$\tau_{ij}=\frac{\pi_i\phi(y_j|\mathbf{x}_j^{\mathrm{T}}\boldsymbol{\beta}_i,\sigma_i^2)}{\sum_{h=1}^{c}\pi_h\phi(y_j|\mathbf{x}_j^{\mathrm{T}}\boldsymbol{\beta}_h,\sigma_h^2)},\tag{16}$$

kde τ_{ij} je zkrácený zápis pro $\tau_i(\mathbf{y}_j,\boldsymbol{\Psi})$, který budeme v následujícím textu používat pro úspornější podobu vzorců.

Víme-li, že váhy složek jsou normované, tedy $\sum_{i=1}^{c}\pi_i=1$, můžeme logaritmickou věrohodnostní rovnici rozšířit ve smyslu Lagrangeovy metody. Na maximum logaritmické věrohodnostní funkce v tomto případě nahlížíme jako na extrém vázaný podmínkou

$$g_1(\pi_1,\ldots,\pi_c)=0,$$

kterou můžeme také zapsat jako

$$\pi_1+\pi_2+\ldots+\pi_c-1=\sum_{i=1}^{c}\pi_i-1=0.$$

Potom lze vytvořit Lagrangeovu fukci

$$\ln L_r(\boldsymbol{\Psi}|\mathbf{x}_1,\ldots,\mathbf{x}_n,y_1,\ldots,y_n)\tag{17}$$

$$= \sum_{j=1}^{n} \ln \left(\sum_{i=1}^{c} \pi_i \phi_i(y_j | \mathbf{x}_j^{\mathrm{T}} \boldsymbol{\beta}_i, \sigma_i^2) \right) - \lambda \left(\sum_{i=1}^{c} \pi_i - 1 \right)$$

za použití Lagrangeova multiplikátoru λ. Při snaze maximalizovat tuto rovnici provedeme její parciální derivace dle odhadovaných parametrů a položíme je rovny nule.

$$\frac{\partial L_r}{\partial \pi_i} = \sum_{j=1}^{n} \frac{1}{\sum_{i=1}^{c} \pi_i \phi_i(*)} \phi_i(*) - \lambda = 0 \tag{18}$$

$$\frac{\partial L_r}{\partial \sigma_i^2} = \sum_{j=1}^{n} \frac{1}{\sum_{i=1}^{c} \pi_i \phi_i(*)} \pi_i \frac{\partial \phi_i(*)}{\partial \sigma_i^2} = 0 \tag{19}$$

$$\frac{\partial L_r}{\partial \boldsymbol{\beta}_i} = \sum_{j=1}^{n} \frac{1}{\sum_{i=1}^{c} \pi_i \phi_i(*)} \pi_i \frac{\partial \phi_i(*)}{\partial \boldsymbol{\beta}_i} = 0, \tag{20}$$

kde $\phi_i(*)$ vyjadřuje $\phi_i(y_j | \mathbf{x}_j^{\mathrm{T}} \boldsymbol{\beta}_i, \sigma_i^2)$ a L_r je zkrácený výraz pro rozšířenou věrohodnostní funkci $L_r(\boldsymbol{\Psi} | \mathbf{x}_1, \ldots, \mathbf{x}_n, y_1, \ldots, y_n)$ ze vztahu (17). Abychom mohli odhadnout přidaný parametr λ, vynásobíme obě strany vztahu (18) vahou π_i a sečteme obě strany přes všechna i, takže dostaneme rovnici

$$\sum_{j=1}^{n} \frac{\sum_{i=1}^{c} \pi_i \phi_i(*)}{\sum_{i=1}^{c} \pi_i \phi_i(*)} - \lambda \sum_{i=1}^{c} \pi_i = 0,$$

tedy

$$\widehat{\lambda} = n.$$

Dále se budeme zabývat odhadem parametru π_i. Proto upravíme rovnici (18), vynásobíme ji vahou π_i a dostaneme se ke vztahu

$$\sum_{j=1}^{n} \frac{\pi_i \phi_i(*)}{\sum_{i=1}^{c} \pi_i \phi_i(*)} - \pi_i \lambda = 0,$$

což můžeme vyjádřit jako

$$\sum_{j=1}^{n} \tau_{ij} - \pi_i \, n = 0,$$

tudíž

$$\widehat{\pi}_i = \frac{\sum_{j=1}^{n} \tau_{ij}}{n}.$$

Již zbývá pouze odhadnout σ_i^2 a $\boldsymbol{\beta}_i$. K tomu použijeme odhad τ_{ij} ze vztahu (16) a pomocí něho upravíme rovnice (19) a (20). Nejprve obě strany rovnice (19) vynásobíme výrazem $\phi_i(*)/\phi_i(*)$:

$$\sum_{j=1}^{n} \frac{\pi_i \phi_i(*)}{\sum_{i=1}^{c} \pi_i \phi_i(*)} \frac{\partial \phi_i(*)}{\partial \sigma_i^2} \frac{1}{\phi_i(*)} = 0.$$

Potom tedy

$$\frac{\partial L_r}{\partial \sigma_i^2} = \sum_{j=1}^{n} \tau_{ij} \frac{\partial \ln \phi_i(*)}{\partial \sigma_i^2} = 0$$

a analogicky

$$\frac{\partial L_r}{\partial \boldsymbol{\beta}_i} = \sum_{j=1}^{n} \tau_{ij} \frac{\partial \ln \phi_i(*)}{\partial \boldsymbol{\beta}_i} = 0.$$

Tyto upravené rovnice očividně korespondují se vztahem (10). Není to nic neobvyklého, výše uvedený postup je totiž analogický i pro směs distribucí, přestože odhadujeme odlišné parametry. Tudíž rovnice pro odhad regresních parametrů $\boldsymbol{\beta}_i$ a rozptylu σ_i^2 jsou v podstatě vážené průměry maximálně věrohodných rovnic [3]

$$\frac{\partial \ln \phi_i(*)}{\partial \boldsymbol{\xi}} = 0.$$

3.3 EM algoritmus pro směs regresí

Postup při výpočtu odhadu regresních parametrů pomocí EM algoritmu bude analogický jako u smíšených distribucí. Nejprve provedeme E-krok, který je stejný

pro všechny konečné smíšené modely. Nechť $\widehat{\mathbf{\Psi}}^{(k)}$ je odhad parametrů po k-té iteraci, potom E-krok v *(k+1)*-ní iteraci obsahuje výpočet hodnoty funkce Q, která zastupuje očekávanou hodnotu logaritmické věrohodnostní funkce za předpokladu kompletních dat

$$Q(\mathbf{\Psi}, \widehat{\mathbf{\Psi}}^{(k)}) = \sum_{j=1}^{n} \sum_{i=1}^{c} \tau_{ij}^{(k)} \, \phi_i(y_j | \mathbf{x}_j^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_i^{(k)}, \widehat{\sigma}_i^{2(k)}),$$

kde vztah pro odhad aposteriorní pravděpodobnosti $\tau_{ij}^{(k)}$ odpovídá (16), tedy

$$\tau_{ij}^{(k)} = \frac{\widehat{\pi}_i^{(k)} \phi(y_j | \mathbf{x}_j^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_i^{(k)}, \widehat{\sigma}_i^{2(k)})}{\sum_{h=1}^{c} \widehat{\pi}_h^{(k)} \phi(y_j | \mathbf{x}_j^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_h^{(k)}, \widehat{\sigma}_h^{2(k)})} \tag{21}$$

M-krok slouží k aktualizaci odhadu $\widehat{\mathbf{\Psi}}^{(k+1)}$, který maximalizuje funkci Q přes všechny $\mathbf{\Psi}$. Nejprve spočítáme aktualizaci vah složek

$$\widehat{\pi}_i^{(k+1)} = \frac{\sum_{j=1}^{n} \tau_{ij}^{(k)}}{n} \qquad \text{pro} \quad i = 1, \ldots, c.$$

Potom můžeme přistoupit k odhadu regresních parametrů pro jednotlivé složky pomocí metody maximální věrohodnosti. Pro i-tou složku uvažujeme náhodné chyby $\varepsilon_{ij} \sim N(0, \sigma_i^2)$, $i = 1, 2, \ldots, n$. Dále mějme $(n \times p)$-dimenzionální designovou matici $\mathbf{X}$, diagonální matici $\mathbf{W}_i$ rozměrů $(n \times n)$ s prvky τ_{ij} na diagonále, tj. $\mathbf{W}_i = \mathrm{Diag}(\tau_{i1}, \ldots, \tau_{in})$, a n-dimenzionální vektor $\mathbf{y}$ realizací závislé proměnné. Odhady parametrů lze potom odvodit pomocí maximálně věrohodného odhadu pro váženou regresi s maticí vah $\mathbf{W}_i$. Pro jednoduchost vynechejme označení čísla iterace a uvažujme obecný výpočet.

Funkce hustoty jednotlivých chyb má v tomto případě tvar

$$f(\varepsilon_{ij}) = \frac{1}{\sigma_i \sqrt{2\pi}} \exp\Big(- \frac{1}{2\sigma_i^2} \frac{\varepsilon_{ij}^2}{\tau_{ij}} \Big).$$

Podobně vyjádříme tvar hustoty vektoru náhodných chyb jako sdruženou hustotu $\prod_{i=1}^{n} f(\varepsilon_{ij})$. Věrohodnostní funkce se dá zapsat následovně

$$L(\boldsymbol{\beta}_i, \sigma_i^2) = \prod_{i=1}^{n} f(\varepsilon_{ij}) = \frac{1}{\sigma_i^n}(2\pi)^{\frac{n}{2}} \exp\left(-\frac{1}{2\sigma_i^2} \boldsymbol{\varepsilon_i}^{\mathrm{T}} \mathbf{W}_i \boldsymbol{\varepsilon_i} \right),$$

nebo obdobně

$$L(\boldsymbol{\beta}_i, \sigma_i^2) = \frac{1}{\sigma_i^{2n}(2\pi)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma_i^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_i)^{\mathrm{T}} \mathbf{W}_i (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_i) \right).$$

Po logaritmizaci se dostáváme ke vztahu

$$\ln\, L(\boldsymbol{\beta}_i, \sigma_i^2) = \ln\left(\frac{1}{\sigma_i^n (2\pi)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma_i^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_i)^{\mathrm{T}} \mathbf{W}_i (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_i) \right) \right)$$

$$= \ln\left(\sigma_i^{-n} (2\pi)^{-\frac{n}{2}}\right) - \frac{1}{2\sigma_i^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_i)^{\mathrm{T}} \mathbf{W}_i (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_i) = -\frac{n}{2} \ln(2\pi)$$

$$-n \ln(\sigma_i) - \frac{1}{2\sigma_i^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_i)^{\mathrm{T}} \mathbf{W}_i (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_i).$$

Tuto funkci s proměnnými $\boldsymbol{\beta}_i$ a σ_i^2 se snažíme maximalizovat, a proto ji budeme derivovat nejprve podle vektoru $\boldsymbol{\beta}_i$ a poté dle proměnné σ_i^2. Tyto výrazy položíme rovny nule a řešíme vzniklé rovnice.

$$\frac{\partial \ln\, L(\boldsymbol{\beta}_i, \sigma_i^2)}{\partial \boldsymbol{\beta}_i} = -\frac{1}{2\sigma_i^2} \left(-2\mathbf{X}^{\mathrm{T}} \mathbf{W}_i \mathbf{y} + 2\mathbf{X}^{\mathrm{T}} \mathbf{W}_i \mathbf{X} \boldsymbol{\beta}_i\right) = \mathbf{0}$$

$$\frac{\partial \ln\, L(\boldsymbol{\beta}_i, \sigma_i^2)}{\partial \sigma_i^2} = -\frac{n}{\sigma_i} + \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_i)^{\mathrm{T}} \mathbf{W}_i (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_i)}{\sigma_i^3} = 0$$

Řešením první rovnice dojdeme k odhadu regresních koeficientů

$$\widehat{\boldsymbol{\beta}}_i = (\mathbf{X}^{\mathrm{T}} \mathbf{W}_i \mathbf{X})^{-1} \mathbf{X}^{\mathrm{T}} \mathbf{W}_i \mathbf{y}. \tag{22}$$

Z druhé rovnice potom vyplývá hodnota odhadu parametru σ^2,

$$\widehat{\sigma}_i^2 = \frac{(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_i)^{\mathrm{T}} \mathbf{W}_i (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_i)}{n}, \tag{23}$$

který můžeme také vyjádřit jako

$$\widehat{\sigma}_i^2 = \frac{\sum_{j=1}^{n} \tau_{ij} \, (y_j - \mathbf{x}_j^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_i)^2}{\sum_{j=1}^{n} \tau_{ij}}.$$

Pokud uvažujeme *(k+1)*-ní iteraci, můžeme vztahy (22) a (23) přepsat následujícím způsobem

$$\widehat{\boldsymbol{\beta}}_i^{(k+1)} = (\mathbf{X}^{\mathrm{T}} \mathbf{W}_i^{(k)} \mathbf{X})^{-1} \mathbf{X}^{\mathrm{T}} \mathbf{W}_i^{(k)} \mathbf{y} \qquad \text{pro} \quad i = 1, \ldots, c. \tag{24}$$

a

$$\widehat{\sigma}^{2(k+1)} = \frac{(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_i^{(k+1)})^{\mathrm{T}} \mathbf{W}_i (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_i^{(k+1)})}{n}, \qquad \text{pro} \quad i = 1, \ldots, c, \tag{25}$$

[2, 7, 17]

3.4 Klasifikační EM algoritmus

Pro modelování směsí regresí se dá kromě EM algoritmu použít též dříve zmíněný SEM algoritmus či, jak ukážeme v této kapitole, klasifikační EM algoritmus (CEM z anglického Classification EM algorithm). Podobně jako SEM i klasifikační EM algoritmus obsahuje navíc jeden krok mezi E-krokem a M-krokem. Úlohou C-kroku je přiřadit každé pozorování do jedné z c komponent na základě podmíněných pravděpodobností příslušnosti k jednotlivým složkám. Pozorování bude přiřazeno do složky, jejíž podmíněná pravděpodobnost τ_{ij} je nejvyšší. Takto odhadneme celý vektor $\mathbf{z}$ a získáme logaritmickou věrohodnostní funkci, kterou budeme chtít v M-kroku maximalizovat, v následujícím tvaru:

$$\ln L(\boldsymbol{\Psi}|z_1, \ldots, z_n, \mathbf{x}_1, \ldots, \mathbf{x}_n, y_1, \ldots, y_n) = \sum_{i=1}^{c} \ln \Big(\sum_{\{j|z_j=i\}}^{c} \pi_i \phi_i(y_j|\mathbf{x}_j^{\mathrm{T}} \boldsymbol{\beta}_i, \sigma_i^2) \Big),$$

kde $\{j|z_j = i\}$ značí, že budeme sčítat přes všechna pozorování přiřazená do i-té složky.

E-krok tohoto algoritmu odpovídá EM algoritmu. V C-kroku v $(k+1)$-ní iteraci jsou všechna pozorování $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ rozřazena do skupin vyjádřených vektorem $\mathbf{P}^{(k+1)} = (P_1^{(k+1)}, \ldots, P_c^{(k+1)})$ tak, že

$$P_i^{(k+1)} = \{(\mathbf{x_j}, y_j) : \tau_{ij}^{(k)} = \max_h \tau_{hj}^{(k)}\}.$$

Pokud narazíme na případ, kdy $\tau_{ij}^{(k)} = \tau_{hj}^{(k)}$, kde $i < h$, potom $(\mathbf{x}_j, y_j) \in P_i^{(k+1)}$. Jestliže je nějaká ze složek $\mathbf{P}^{(k+1)}$ prázdná, je vhodné uvážit směs $c-1$ regresních modelů místo c.

V M-kroku je potom provedena aktualizace odhadu parametrů $\widehat{\mathbf{\Psi}}^{(k+1)}$ za použití podmnožin pozorování $P_i^{(k+1)}$. Odhad vah složek je dán explicitně vztahem

$$\widehat{\pi}_i^{(k+1)} = \frac{n_i}{n} \qquad \text{pro} \quad i = 1, \ldots, c,$$

kde n_i značí počet pozorovaní pocházejících z i-té složky. Odhad regresních parametrů po M-kroku bude následující:

$$\widehat{\boldsymbol{\beta}}_i^{(k+1)} = (\mathbf{X}_i^{\mathrm{T}} \mathbf{W}_i^{(k)} \mathbf{X}_i)^{-1} \mathbf{X}^{\mathrm{T}}{}_i \mathbf{W}_i^{(k)} \mathbf{y}_i \qquad \text{pro} \quad i = 1, \ldots, c,$$

přičemž $\mathbf{X}_i$ je $(n_i \times p)$-dimenzionální designová matice pro i-tou složku, $\mathbf{W}_i^{(k)}$ je diagonální matice rozměrů $(n_i \times n_i)$ s prvky $\tau_{ij}^{(k)}$ na diagonále a $\mathbf{y}_i$ je n_i-dimenzionální vektor závislé proměnné pro i-tou složku. Rozptyl náhodných chyb pro složku i odhadneme pomocí vztahu

$$\widehat{\sigma}_i^{2\,(k+1)} = \frac{\sum_{j=1}^{n_i} \tau_{ij}^{(k)} \, (y_j - \mathbf{x}_j^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_i^{(k+1)})^2}{\sum_{j=1}^{n_i} \tau_{ij}^{(k)}} \qquad \text{pro} \quad i = 1, \ldots, c.$$

Kroky E, C a M opakujeme, dokud algoritmus nesplní požadované kritérium konvergence. [7]

4 Modifikace směsí regresních modelů

Již v dřívějších kapitolách bylo zmíněno, že smíšené distribuce jsou řešením pro modelování dat, která pocházejí z více distribucí a tvoří tedy v populaci jisté podskupiny, přičemž důvod této rozdílnosti dat není přímo pozorovatelný či teoreticky podchytitelný. Ve smíšených distribucích lze onu příčinu diference dat označit jako doprovodnou proměnnou. Tuto proměnnou lze zařadit do modelu smíšené distribuce či směsi regresních modelů a dosáhnout tak mnoha výhodných vlastností. V kontextu směsi lineárních regresí se jedná především o predikce závisle proměnné či odhad chybějících dat. Ukazuje se, že možnosti standardního modelu směsi regresí jsou v této oblasti značně omezené, což bude mimo jiné demonstrováno v kapitole (5.3).

Připomeňme si, že klasická směs regresních modelů má tvar

$$f(y_j|\mathbf{x}_j) = \sum_{i=1}^{c} \pi_i \phi(y_j|\mathbf{x}_j^{\mathrm{T}}\boldsymbol{\beta}_i, \sigma_i^2).$$

Pomocí EM algoritmu je možné odhadnout vektor parametrů

$$\boldsymbol{\Psi} = (\pi_1, \ldots, \pi_c, (\boldsymbol{\beta}_1^{\mathrm{T}}, \sigma_1^2), \ldots, (\boldsymbol{\beta}_c^{\mathrm{T}}, \sigma_c^2))^{\mathrm{T}},$$

který nám již umožňuje pracovat s predikcemi či chybějícími hodnotami. Při práci s regresním modelem se často dostáváme do situace, kdy máme k dispozici nové hodnoty vysvětlující proměnné a snažíme se na jejím základě predikovat hodnoty veličiny Y. S odhadnutými parametry standardní směsi regresních modelů je tato situace jednoduchá, avšak výsledek není zcela uspokojující. Nejlepší predikce je v tomto případě dána na základě vztahu (15), tedy

$$\sum_{i=1}^{c} \widehat{\pi}_i(\mathbf{x}_f^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_i), \tag{26}$$

kde $\mathbf{x}_f$ je p-dimenzionální vektor nových pozorování. Takováto predikce je očividně silně nepřesná, protože odhadnutá hodnota bude v podstatě váženým průměrem z hodnot odhadnutých pomocí jednotlivých regresních funkcí, bude se tedy pohybovat někde mezi shluky dat.

Přidáním doprovodné proměnné do modelu se způsoby predikce zlepšují a interpretace modelu je reálnější.

4.1 Směs regresních modelů doprovodné proměnné

Tento model je založen na klasické směsi regresí, ale navíc využívá parametrizace marginálních pravděpodobností π_i, což znamená, že je na tyto pravděpodobnosti nahlíženo jako na funkce doprovodné proměnné. Směs regresních modelů doprovodné proměnné zapíšeme jako

$$f(y_j|\mathbf{x}_j, \boldsymbol{\omega}_j) = \sum_{i=1}^{c} \pi_i(\boldsymbol{\omega}_j, \boldsymbol{\alpha}_i)\phi(y_j|\mathbf{x}_j^{\mathrm{T}}\boldsymbol{\beta}_i, \sigma_i^2),$$

kde již klasicky $\mathbf{x}_j$ označuje j-té pozorování vysvětlující proměnné, y_j pozorování závisle proměnné a $\boldsymbol{\omega}_j$ jsou hodnoty doprovodné proměnné. Již zmíněná parametrizovaná pravděpodobnost $\pi_i(\boldsymbol{\omega}_j, \boldsymbol{\alpha}_i)$ závisí na hodnotách $\boldsymbol{\omega}_j$ a parametrech $\boldsymbol{\alpha}_i$. Vektor $\boldsymbol{\Psi}$ obsahující všechny odhadované parametry se v tomto případě upraví do podoby

$$\boldsymbol{\Psi} = ((\boldsymbol{\alpha}_1^{\mathrm{T}}, \boldsymbol{\beta}_1^{\mathrm{T}}, \sigma_1^2), \ldots, (\boldsymbol{\alpha}_c^{\mathrm{T}}, \boldsymbol{\beta}_c^{\mathrm{T}}, \sigma_c^2))^{\mathrm{T}}.$$

Stejně jako v předchozích kapitolách uvažujeme restrikce marginálních pravděpodobností

$$\sum_{i=1}^{c} \pi_i(\boldsymbol{\omega}_j, \boldsymbol{\alpha}_i) = 1 \quad \text{a} \quad \pi_i(\boldsymbol{\omega}_j, \boldsymbol{\alpha}_i) > 0 \qquad \text{pro} i = 1, \ldots, c.$$

Pro tyto pravděpodobnosti je obvykle uvažován multinomický logitový model, který zapíšeme pomocí vztahu

$$\pi_i(\boldsymbol{\omega}_j, \boldsymbol{\alpha}_i) = \frac{\exp^{\boldsymbol{\omega}_j^{\mathrm{T}}\boldsymbol{\alpha}_i}}{\sum_{h=1}^{c} \exp^{\boldsymbol{\omega}_j^{\mathrm{T}}\boldsymbol{\alpha}_h}} \qquad \text{pro} \qquad i = 1, \ldots, c,$$

přičemž $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_c)$ a parametr $\boldsymbol{\alpha}_1 \equiv \mathbf{0}$. [9]

Podobně jako v případě klasické směsi regresních modelů je k odhadu neznámých parametrů používán EM algoritmus. Máme k dispozici n pozorování

$(y_1, \mathbf{x}_1, \omega_1), \ldots, (y_n, \mathbf{x}_n, \omega_n)$ a skryté proměnné $\mathbf{z}_1, \ldots, \mathbf{z}_n$. EM algoritmus je potom dán ve dvou krocích.

E-krok: Probíhá výpočet podmíněných pravděpodobností τ_{ij}, že j-té pozorování náleží do i-té složky, pomocí vzorce

$$\tau_{ij}^{(k)} = \frac{\pi_i(\boldsymbol{\omega}_j, \widehat{\boldsymbol{\alpha}}_i^{(k)})\phi(y_j|\mathbf{x}_j^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_i^{(k)}, \widehat{\sigma}_i^{2(k)})}{\sum_{h=1}^{c} \pi_h(\boldsymbol{\omega}_j, \widehat{\boldsymbol{\alpha}}_i^{(k)})\phi(y_j|\mathbf{x}_j^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_i^{(k)}, \widehat{\sigma}_i^{2(k)})}$$

za podmínky, že $(\widehat{\boldsymbol{\alpha}}^{(k)}, \widehat{\boldsymbol{\beta}}^{(k)}, \widehat{\sigma}^{2(k)})$ jsou odhady získané v předchozí iteraci.

M-krok: Pomocí pořízených odhadů podmíněných pravděpodobností τ_{ij} lze získat nový odhad $\widehat{\boldsymbol{\Psi}}^{(k+1)}$ vektoru parametrů pomocí maximalizace funkce

$$Q(\boldsymbol{\Psi}, \widehat{\boldsymbol{\Psi}}^{(k)}) = Q_1((\widehat{\boldsymbol{\beta}}^{(k+1)}, \widehat{\sigma}^{2(k+1)}), \widehat{\boldsymbol{\Psi}}^{(k)}) + Q_2(\widehat{\boldsymbol{\alpha}}^{(k+1)}, \widehat{\boldsymbol{\Psi}}^{(k)}),$$

kde

$$Q_1((\widehat{\boldsymbol{\beta}}^{(k+1)}, \widehat{\sigma}^{2(k+1)}), \widehat{\boldsymbol{\Psi}}^{(k)}) = \sum_{j=1}^{n}\sum_{i=1}^{c} \tau_{ij}^{(k)} \ln\left(\phi(y_j|\mathbf{x}_j^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_i^{(k+1)}, \widehat{\sigma}_i^{2(k+1)})\right)$$

a

$$Q_2(\widehat{\boldsymbol{\alpha}}^{(k+1)}, \widehat{\boldsymbol{\Psi}}^{(k)}) = \sum_{j=1}^{n}\sum_{i=1}^{c} \tau_{ij}^{(k)} \ln\left(\pi_i(\boldsymbol{\omega}_j, \widehat{\boldsymbol{\alpha}}_i^{(k+1)})\right).$$

Funkce Q_1 a Q_2 lze také maximalizovat samostatně. Maximalizací Q_1 získáme odhady parametrů $(\widehat{\boldsymbol{\beta}}^{(k+1)}, \widehat{\sigma}^{2(k+1)})$ a podobně maximalizací Q_2 obdržíme aktualizaci odhadů $\widehat{\boldsymbol{\alpha}}^{(k+1)}$. [9][18]

4.2 Saturovaná směs regresních modelů

Dalším způsobem, jak rozšířit standardní směs regresí, je model zvaný saturovaná směs regresí. Jeho výhodou je, že v něm doprovodná proměnná není zahrnutá přímo, ale její vlastnosti promítneme do vysvětlující proměnné $\mathbf{x}$, dojde tedy k jejímu znáhodnění. Sjednocený model v tomto případě vyjádříme jako

$$f(y, \mathbf{x}, \boldsymbol{\Psi}) = \sum_{i=1}^{c} \pi_i \phi(y|\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}, \sigma^2)\, f(\mathbf{x}|\boldsymbol{\gamma}_i),$$

kde $f(\mathbf{x}|\boldsymbol{\gamma}_i)$ je funkce popisující rozdělení pravděpodobností proměnné $\mathbf{x}$ pro i-tou složku. Parametry tohoto rozdělení jsou $\boldsymbol{\gamma}_i$. [18]

Toshiya Hoshikawa ve svém článku ([11]) předpokládá, že $f(\mathbf{x}|\boldsymbol{\gamma}_i)$ je hustota vícerozměrného normálního rozdělení, tedy $\phi(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$. Vektor $\boldsymbol{\mu}_i$ reprezentuje vektor středních hodnot i-té složky a matice $\boldsymbol{\Sigma}_i$ je její varianční matice. Celkový odhadovaný parametr $\boldsymbol{\Psi}$ nelze v tomto případě vyjádřit pomocí vektoru a proto jej vyjádříme množinou ve tvaru

$$\boldsymbol{\Psi} = (\pi_1, \ldots, \pi_{c-1}, \boldsymbol{\beta}_1, \sigma_1^2, \ldots, \boldsymbol{\beta}_c, \sigma_c^2, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i).$$

Klíčový rozdíl mezi klasickým modelem a saturovanou verzí je zřejmý při snaze predikovat hodnoty závisle proměnné Y pomocí $\mathbf{x}_f$, nových hodnot vysvětlující proměnné. Na rozdíl od vztahu (26), saturovaná směs regresí poskytne predikovanou hodnotu Y pomocí hodnot $E(Z_i|\mathbf{x}_f)$, které vyjadřují podmíněnou pravděpodobnost, že nové pozorování náleží do i-té složky za podmínky napozorovaných dat $\mathbf{x}_f$. Tato pravděpodobnost je dána jako

$$\tau_i(\mathbf{x}_f) = E(Z_i|\mathbf{x}_f) = \frac{\widehat{\pi}_i \phi(\mathbf{x}_f|\widehat{\boldsymbol{\mu}}_i, \widehat{\boldsymbol{\Sigma}}_i)}{\sum_{h=1}^{c} \widehat{\pi}_h \phi(\mathbf{x}_f|\widehat{\boldsymbol{\mu}}_h, \widehat{\boldsymbol{\Sigma}}_h)}.$$

Predikci potom vypočítáme pomocí vztahu

$$\sum_{i=1}^{c} \tau_i(\mathbf{x}_f)(\mathbf{x}_f^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_i).$$

Můžeme tedy říct, že čím více jsou distribuce vysvětlujících proměnných mezi skupinami separované, tím lépe si tento model při predikci povede. [11]

Dalším, poněkud praktičtějším přístupem k predikci hodnot závisle proměnné je využití podmíněných pravděpodobností $\tau_i(\mathbf{x}_f)$, pomocí kterých přiřadíme nové pozorování $\mathbf{x}_f$ do i-té složky na základě pravidla

$$\max_{i=1,\ldots,c} \tau_i(\mathbf{x}_f) = \max_{i=1,\ldots,c} \frac{\widehat{\pi}_i \phi(\mathbf{x}_f|\widehat{\boldsymbol{\mu}}_i, \widehat{\boldsymbol{\Sigma}}_i)}{\sum_{h=1}^{c} \widehat{\pi}_h \phi(\mathbf{x}_f|\widehat{\boldsymbol{\mu}}_h, \widehat{\boldsymbol{\Sigma}}_h)}.$$

Predikci závisle proměnné potom spočítáme jako

$$\widehat{y}_f = \mathbf{x}_f^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_i.$$

5 Praktická část

Firma Seco GROUP a.s. je dominantní výrobce žacích traktorů v České republice a významný dodavatel této techniky zejména do zemí Evropské unie. Patří k vedoucím slévárnám v České republice v oblasti menších přesnějších odlitků z tvárné litiny zejména pro evropské výrobce osobních a nákladních automobilů, traktorů a stavebních strojů. Tato společnost je také dominantní výrobce vložených válců do dieselových motorů v České republice a významný dodavatel renomovaných výrobců motorů zejména v Evropské unii.

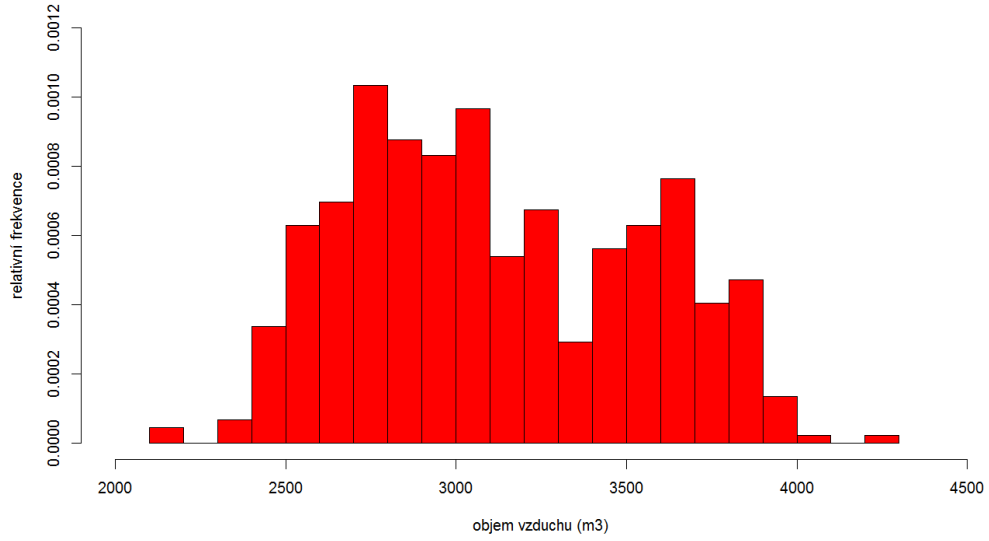
Celá továrna využívá systému trubek sloužícího k rozvodu stlačeného vzduchu. Tento stlačený vzduch je generován čtyřmi kompresory, jež tvoří hlavní zdroj odběru energie pro celou výrobní halu. Pro společnost by proto bylo žádoucí znát vztah mezi dodanou energií do kompresorů a objemem vydaného stlačeného vzduchu. Tři z těchto kompresorů jsou ovládány velmi jednoduše - kdykoliv je objem stlačeného vzduchu nedostatečný, kompresory jsou postupně zapínány, naopak při přebytku vzduchu opět vypínány. Poslední z kompresorů lze jako jediný z celé skupiny regulovat, tudíž je naším hlavním zájmem popsat funkční závislost objemu vydaného vzduchu na dodané energii právě do tohoto kompresoru.

Ideální by samozřejmě bylo modelovat vztah zahrnující všechny čtyři kompresory, avšak tento model by bylo složité teoreticky podložit v předchozích kapitolách a praktická část by se svou délkou také vymykala možnostem této práce. Dalo by se tedy říci, že následující příklad není přesným řešením daného problému, který firma Seco GROUP a.s. zadala, ale je vhodným názorným příkladem k námi položenému teoretickému základu a z praktického hlediska by jeho řešení mohlo přispět k vylepšení regulačního systému, na jehož základě kompresor doposud fungoval. Hodnoty dodané energie do kompresoru a celkového objemu vzduchu jsou k nalezení v příloze 1.

V následujících dvou kapitolách se budeme postupně zabývat modelováním směsi normálních rozdělení pro objem vzduchu a modelováním směsi regresních modelů pro závislost objemu vzduchu na energii dodané do kompresoru. [12]

5.1 Směs dvou normálních distribucí

Na datech popisujících objem komprimovaného vzduchu můžeme přehledně demonstrovat postup při modelování jednorozměrné smíšené distribuce, skládající se ze dvou normálních rozdělení, v programu R. Pokud vykreslíme histogram pro tato data (viz obrázek 5.1), je z něj zřejmé, že hodnoty proměnné nepocházejí z normálního rozdělení.



Obr. 5.1: Histogram pro objem vzduchu.

Avšak podle tvaru histogramu lze zjevně usuzovat na směs dvou normálních hustot. Jejich parametry budeme chtít odhadnout pomocí EM algoritmu, který program R nabízí v rámci funkce `normalmixEM` z balíčku `mixtools`. Tato funkce je určena speciálně pro případ, kdy modelujeme směs normálních distribucí. Jakožto iterační procedura má tato funkce mnohá nastavení související s počátečními odhady, počtem iterací či ukončovacím kritériem. Nejdůležitější parametry si zde stručně představíme:

`x` vektor rozměru $n \times 1$ obsahující jednorozměrná zkoumaná data, matice rozměru $n \times p$ pro p-rozměrná zkoumaná data

`lambda` počáteční hodnoty odhadu vah složek smíšené distribuce. Při zadání pouze jedné hodnoty bude tato váha použita pro všech `k` složek, posléze budou hodnoty normalizovány, aby jejich součet dával číslo 1. Pokud není lambda specifikována, budou její hodnoty náhodně vygenerovány s použitím rovnoměrného rozdělení pravděpodobností a následně opět normalizovány.

`mu` počáteční hodnoty pro odhad středních hodnot složek. Jestliže je zadán nenulový vektor, algoritmus tyto hodnoty vyhodnotí jako počáteční hodnoty a zároveň automaticky nastaví parametr `k` na hodnotu délky tohoto vektoru. Pokud však zadáme pouze jednu hodnotu, ta bude platit jako počáteční odhad pro všechny složky a zároveň bude parametr `arbmean` nastaven na FALSE. Nezadáme-li hodnotu parametru sami, bude počáteční odhad náhodně vygenerován z normálního rozdělení, jehož střední hodnoty jsou dány rozdělením dat do skupin.

`sigma` startovací hodnoty pro odhad směrodatné odchylky. Jestliže je zadán nenulový vektor, algoritmus tyto hodnoty vyhodnotí jako počáteční hodnoty a zároveň automaticky nastaví parametr `k` na hodnotu délky tohoto vektoru. Pokud však zadáme pouze jednu hodnotu, ta bude platit jako počáteční odhad směrodatných odchylek pro všechny složky a zároveň bude parametr `arbvar` nastaven na FALSE. Nezadáme-li hodnotu parametru sami, bude počáteční odhad roven převrácené hodnotě odmocniny z vektoru hodnot náhodně vygenerovaných z exponencionálního rozdělení, kde střední hodnoty jsou definovány opět pomocí přerozdělení dat do skupin.

`k` počet komponent smíšené distribuce. Zřetel je na tuto hodnotu brán pouze, pokud nejsou zadané vektory počátečních odhadů středních hodnot, vah složek či směrodatných odchylek. Jinak řečeno je primárně hodnocena délka těchto vektorů, a teprve pokud nejsou tyto startovací hodnoty zadány nebo jsou v podobě skaláru, bude počet složek určen parametrem `k`.

`epsilon` ukončovací kritérium. Algoritmus je ukončen, jakmile je rozdíl v hodnotách logaritmických věrohodnostních funkcí v po sobě jdoucích iteracích menší než daná hodnota parametru.

`maxit` maximální počet iterací.

`arbmean` Hodnota tohoto parametru je implicitně nastavena na TRUE, což znamená, že algoritmus nepředpokládá stejné `mu` pro všechny složky. Hodnota tohoto parametru je ignorována, pokud jsou zadané vektory počátečních odhadů středních hodnot, vah složek či směrodatných odchylek.

`armvar` Hodnota tohoto parametru je implicitně nastavena na TRUE, což znamená, že algoritmus nepředpokládá stejné `sigma` pro všechny složky. Hodnota tohoto parametru je ignorována, pokud jsou zadané vektory počátečních odhadů středních hodnot, vah složek či směrodatných odchylek.

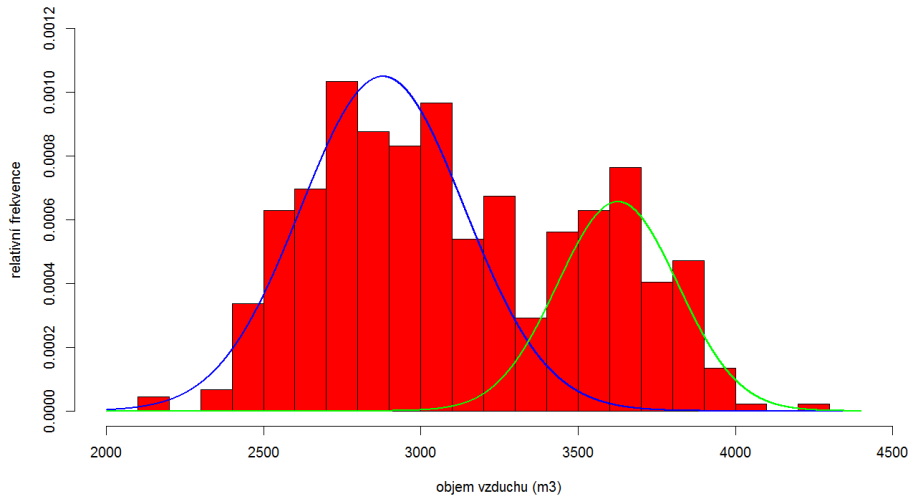
Pro názornost si představíme několik způsobů zadání funkce pro naše data. Nejprve vyzkoušejme proceduru `normalmixEM` za předpokladu, že zadáme pouze odhadovaný počet složek, který určíme z histogramu. Počáteční odhady necháme vygenerovat, maximální počet iterací a hodnotu epsilon necháme na implicitním nastavení, tedy `maxit`=1000 a `epsilon`=10^{-8}. Jelikož nepředpokládáme rovnost středních hodnot či rozptylů, oba parametry `arbmean` a `armvar` ponecháme na přednastavené hodnotě TRUE.

```
> library(mixtools)
> mix=normalmixEM(y, lambda = NULL, mu = NULL, sigma = NULL, k=2)
number of iterations= 124
> summary(mix)
summary of normalmixEM object:
            comp 1      comp 2
lambda    0.685125    0.314875
mu     2879.692103 3626.775127
sigma   260.394462  191.325710
loglik at estimate:  -3286.568
```

Program R poskytuje odhad proporcí složek pod názvem `lambda`, odhad středních hodnot `mu` a směrodatných odchylek `sigma`. Mezi výstupy můžeme nalézt také logaritmickou věrohodnost odhadu a počet provedených iterací algoritmu. Počet iterací se samozřejmě bude lišit v závislosti na náhodném generování startovacích hodnot. Tyto základní výstupy nám umožňují vykreslit odhadnuté hustoty normálních rozdělení, aby bylo možné je porovnat s dříve uvedeným histogramem napozorovaných hodnot (viz obrázek 5.2) Kód potřebný k vytvoření tohoto grafického výstupu je následující:

```
hist(y, 30, col="red",ylim=c(0,0.0012),xlim=c(2000,4500),main="Histogram",
 xlab="objem vzduchu (m3)", ylab="relativní frekvence",prob=TRUE)
xfit<-seq(2000,4400,length=40000)
yfit<-mix$lambda[1]*dnorm(xfit,mean=mix$mu[1],sd=mix$sigma[1])
lines(xfit, yfit, col="blue", lwd=2)

xfit<-seq(2000,4400,length=40000)
```

```
yfit<- mix$lambda[2]*dnorm(xfit,mean=mix$mu[2],sd=mix$sigma[2])
lines(xfit, yfit, col="green", lwd=2)
```



Obr. 5.2: Histogram pro objem vzduchu společně s odhadnutými hustotami normálních rozdělení pravděpodobností.

Podrobnosti o celém iteračním algoritmu získáme jednoduše pomocí vypsání všech poskytovaných výstupů

```
> mix[c("lambda", "mu", "sigma","all.loglik","restarts","posterior")]
```

Tento příkaz nám umožní pracovat nejen se základními hodnotami, ale také zobrazí hodnoty logaritmických věrohodnostních funkcí `$all.loglik` pro každou iteraci, počet restartů EM algoritmu `$restarts` (nastává v případě špatně zvolených startovacích hodnot) a posteriorní pravděpodobnosti příslušnosti k jednotlivým složkám pro každé pozorování `$posterior`. Ukázku jmenovaných výstupů můžeme vidět zde

```
> mix[c("lambda", "mu", "sigma","all.loglik","restarts","posterior")]
$lambda
[1] 0.6851244 0.3148756

$mu
[1] 2879.692 3626.775

$sigma
```

```
[1] 260.3943 191.3259

$all.loglik
  [1] -3444.444 -3308.050 -3302.103 -3299.186 -3297.535 -3296.617 -3296.061
  [8] -3295.657 -3295.304 -3294.963 -3294.621 -3294.273 -3293.919 -3293.560

                        ⋮

 [92] -3286.568 -3286.568 -3286.568 -3286.568 -3286.568 -3286.568 -3286.568
 [99] -3286.568 -3286.568 -3286.568

$restarts
[1] 0

$posterior
              comp.1       comp.2
  [1,] 0.9998905534 1.094466e-04
  [2,] 0.9999976449 2.355097e-06
  [3,] 0.9999874950 1.250500e-05
  [4,] 0.9999999150 8.496126e-08
  [5,] 0.9999994588 5.411789e-07
  [6,] 0.0477560826 9.522439e-01
  [7,] 0.0209495260 9.790505e-01
  [8,] 0.0020922212 9.979078e-01
  [9,] 0.0080444135 9.919556e-01
 [10,] 0.0014095558 9.985904e-01

                        ⋮

[444,] 0.9998386500 1.613500e-04
[445,] 0.9999818870 1.811300e-05
```

Analogicky můžeme postupovat v případě, kdy udáme přibližné odhady pro váhy a střední hodnoty. Tyto hodnoty jsou čitelné z histogramu na obrázku 5.1, kdy zvolíme $\mu_1 = 2900$, $\mu_2 = 3700$ a $\pi_1 = 0.6$, $\pi_2 = 0.4$. To urychlí konvergenci EM algoritmu, takže zatímco v předchozím případě byl počet iterací obvykle vyšší než 100, nyní se pohybujeme mezi 60 až 100 iteracemi.

```
> mix2=normalmixEM(y,lambda = c(0.6,0.4), mu = c(2900,3700), sigma = NULL, k=2)
number of iterations= 80
> summary(mix2)
summary of normalmixEM object:
            comp 1      comp 2
lambda    0.685125    0.314875
```

```
mu     2879.692314 3626.775423
sigma   260.394596  191.325533
loglik at estimate:  -3286.568
```

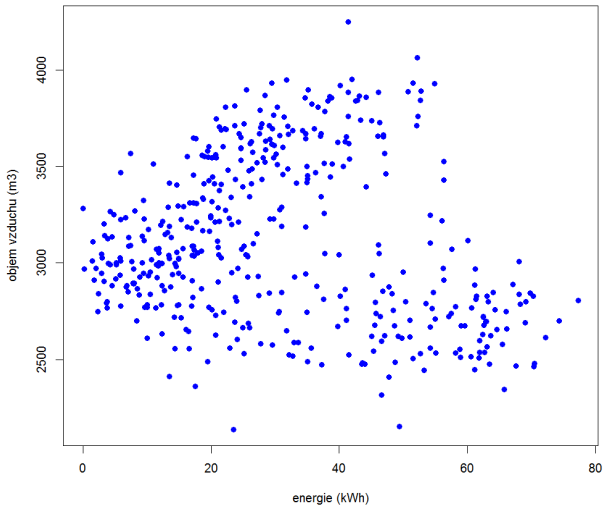
Získané odhady jsou očividně stejně přesné jako odhady získané předchozím algoritmem (nepatrné rozdíly vznikají pro každý algoritmus i při stejně nastavených parametrech kvůli zahrnuté náhodnosti v iteračním procesu). Jistý rozdíl v odhadnutých hodnotách lze pozorovat při omezení počtu iterací, což lze pozorovat v první části níže uvedeného kódu, či při snížení hodnoty `epsilon`.

```
> mix3=normalmixEM(y, lambda = NULL, mu = NULL, sigma = NULL, k=2, maxit=40)
WARNING! NOT CONVERGENT!
number of iterations= 40
> summary(mix3)
summary of normalmixEM object:
            comp 1      comp 2
lambda    0.572274    0.427726
mu     3374.989647 2759.050123
sigma   342.801305  202.226122
loglik at estimate:  -3295.986

> mix4=normalmixEM(y, lambda = NULL, mu = NULL, sigma = NULL, k=2, epsilon=0.1)
number of iterations= 36
> summary(mix4)
summary of normalmixEM object:
            comp 1      comp 2
lambda    0.654774    0.345226
mu     2859.373989 3595.639613
sigma   248.116991  210.676178
loglik at estimate:  -3287.029
```

V obou těchto případech jsou odhady velmi nepřesné a v případě omezení počtu iterací i samotný program R upozorňuje uživatele hláškou `WARNING! NOT CONVERGENT!`, že algoritmus nekonverguje. Jelikož EM algoritmus je pro tyto jednoduché směsi distribucí v programu R záležitostí několika vteřin i při velkém počtu iterací, omezování počtu iterací či ukončovacího kritéria se nedoporučuje. [6]

5.2 Směs lineárních regresních modelů

Nyní se pokusíme odhadnout parametry regresní funkce vyjadřující závislost mezi energií dodanou do kompresoru a objemem stlačeného vzduchu. Pro použití obyčejné lineární regrese je však potřeba, aby data splňovala jisté požadavky. Jeden z nich je normální rozdělení chyb měření, což v našem případě, jak brzy ukážeme, není splněno. Navíc ze samotného zobrazení závislosti dvou zkoumaných veličin je zřejmé, že data se pravděpodobně skládají ze dvou různých skupin, jak lze pozorovat na obrázku 5.3.

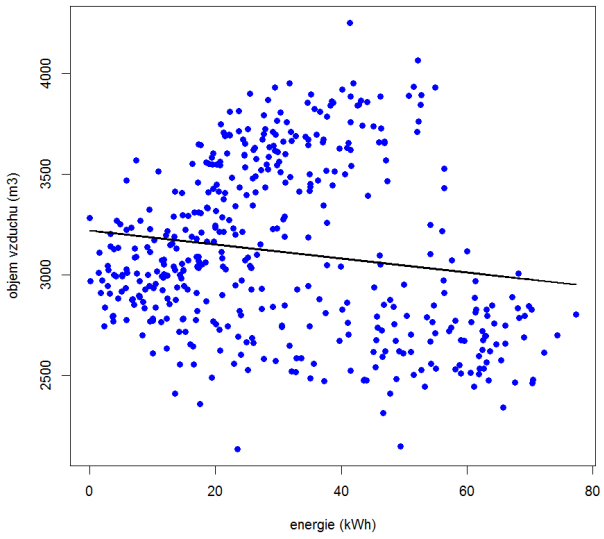


Obr. 5.3: Zobrazení závislosti objemu vzduchu na dodané energii.

Při pokusu o použití klasické lineární regrese $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ dostaneme následující výsledky:

```
> linreg=lm(y~x)
> summary(linreg)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-1001.5  -320.5  -103.8   372.9  1171.9

Coefficients:
```

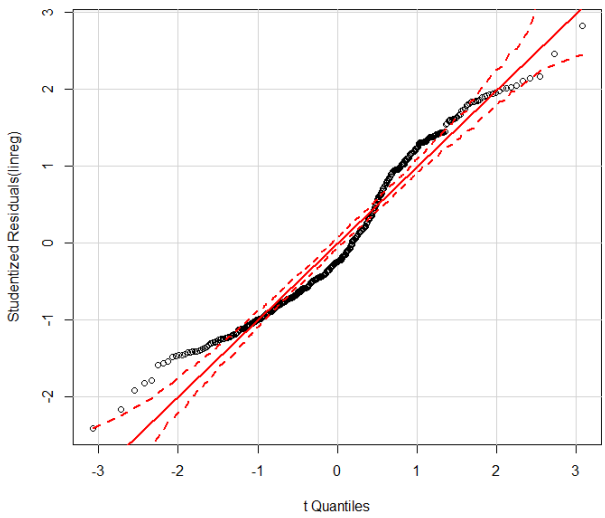
```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3220.896     38.374  83.935  < 2e-16 ***
x             -3.465      1.074  -3.226  0.00135 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 418.4 on 443 degrees of freedom
Multiple R-squared: 0.02295,    Adjusted R-squared: 0.02074
F-statistic: 10.41 on 1 and 443 DF,  p-value: 0.001349
```



Obr. 5.4: Proložení dat regresní přímkou.

Již podle indexu determinace a obrázku 5.4, který demonstruje proložení dat regresní přímkou s parametry získanými klasickou lineární regresí, je zřejmé, že model nebude příliš spolehlivý. Zároveň je možné analyzovat předpoklad normality např. pomocí Q-Q grafu (viz obrázek 5.5) nebo testů normality. Q-Q graf lze v programu R vykreslit pomocí kódu

```
> library(car)
> qqPlot(linreg, main = "QQ Plot")
```



Obr. 5.5: Q-Q graf porovnávající teoretické kvantily normálního rozdělení s kvantily náhodných chyb.

Jelikož se hodnoty od středové přímky, reprezentující podobnost s normálním rozdělením, silně odchylují (zcela mimo meze vykreslených intervalů), lze usuzovat, že náhodné chyby nejsou normálně rozděleny. Tuto hypotézu potvrzuje také Shapiro-Wilkův test normality

```
> shapiro.test(residuals(linreg))

        Shapiro-Wilk normality test

data:  residuals(linreg)
W = 0.9585, p-value = 6.993e-10
```

Na základě p-hodnoty můžeme zamítnout nulovou hypotézu, že data pochází z normálního rozdělení.

Podle obrázku 5.3 lze navíc soudit, že se v datech vyskytují jisté podskupiny, a jelikož jsme nebyli schopni identifikovat latentní proměnnou a na jejím základě data rozdělit, je pouze logické pokusit se závislost těchto dvou veličin modelovat pomocí směsi lineárních regresí. Pro tuto proceduru program R používá v jednodušších případech funkci `regmix` z balíčku `fpc`, u které si stejně jako v předchozí kapitole představíme základní parametry, jež budeme využívat. Tato procedura umožňuje výpočet maximálně věrohodného odhadu regresních parametrů, rozptylu a vah složek pomocí EM algoritmu. Počet složek je určen pomocí Bayesovského informačního kritéria.

`indep` číselný vektor či matice hodnot nezávisle proměnné.

`dep` číselný vektor hodnot závisle proměnné.

`ir` pozitivní číslo udávající počet iterací, které budou provedeny pro každý zadaný počet složek. Implicitně je tento argument nastaven na hodnotu 1.

`nclust` vektor obsahující pozitivní čísla, která reprezentují počty složek, pro které algoritmus bude spuštěn.

`icrit` ukončovací kritérium (rozdíl hodnot logaritmických věrohodnostních funkcí ve dvou po sobě jdoucích iteracích). Implicitní nastavení je hodnota 10^{-5}.

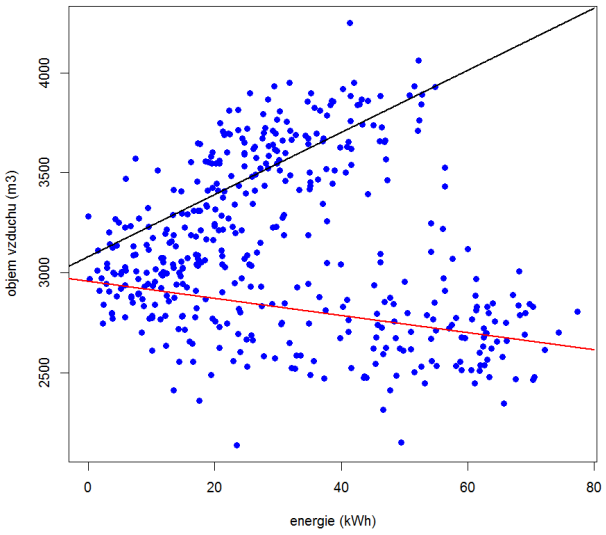
`minsig` pozitivní číslo reprezentující minimální hodnotu pro odhady rozptylů - pokud mohou jít odhady rozptylů k nule, věrohodnost se stává neomezenou. Předdefinovaná hodnota je 10^{-6}.

`m` matice pozitivních čísel, jejíž počet sloupců musí být roven počtu složek a počet řádků musí být totožný s počtem pozorování. Součet každého řádku přes všechny sloupce musí být roven jedné. Jedná se o počáteční odhady pravděpodobností, že pozorování přísluší k dané složce. Startovací hodnoty jsou vygenerovány pomocí funkce `randcmatrix`.

V našem případě zvolíme velmi obecné nastavení. Zvýšíme pouze počet iterací z jedné na dvě, abychom zaručili větší přesnost. Dále spustíme algoritmus pro jednu, dvě či tři složky, čímž nastíníme možnost postupu v případě, že si počtem složek nejsme jistí. Program R nám jako výsledek poskytne hned několik užitečných výstupů. Hodnota parametru `clnopt` udává ideální počet složek, `loglik` je logaritmická maximální věrohodnost pro tento počet, `bic` obsahuje všechny logaritmické věrohodnosti, `coef` uvádí odhady regresních parametrů, `var` potom odhady rozptylů a `eps` odhady vah složek. Matice `z` obsahuje váhy příslušnosti bodů k jednotlivým shlukům a `g` je vektor zobrazující rozdělení napozorovaných bodů do jednotlivých složek. [5]

```
> library(fpc)
> fit=regmix(indep=x, dep=y, ir=2, nclust=1:3)
Iteration  1  for  1  clusters.
Iteration  2  for  1  clusters.
Iteration  1  for  2  clusters.
Iteration  2  for  2  clusters.
Iteration  1  for  3  clusters.
```

```
Iteration  2  for  3  clusters.
> fit
$clnopt
[1] 2

$loglik
[1] -3215.19

$bic
[1] -6651.628 -6473.066 -6493.605

$coef
            [,1]       [,2]
[1,] 2956.415545 3080.46436
[2,]   -4.280718   15.53675

$var
[1] 43329.15 39225.57

$eps
[1] 0.553335 0.446665

$z
               [,1]         [,2]
  [1,] 1.000000e+00 4.481728e-09
  [2,] 9.951437e-01 4.856335e-03
  [3,] 9.966573e-01 3.342714e-03
  [4,] 9.999989e-01 1.077640e-06
  [5,] 1.000000e+00 8.986967e-09
  [6,] 1.181662e-03 9.988183e-01
  [7,] 3.652361e-04 9.996348e-01
  [8,] 3.213027e-06 9.999968e-01
  [9,] 2.318477e-05 9.999768e-01
 [10,] 2.627882e-07 9.999997e-01

                     ⋮

[445,] 9.970331e-01 2.966924e-03

$g
  [1] 1 1 1 1 1 2 2 2 2 2 2 2 1 1 1 1 1 2 2 2 2 2 2 1 1 1 1 1 2 2 1 1 1 1 1 1 1

                                          ⋮

[408] 1 1 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 1 1 1 1
[445] 1
```



Obr.5.6: Data proložená regresními funkcemi.

Na obrázku 5.6 lze vidět odhadnuté regresní přímky pro první složku $\widehat{y}_{1i} = 2956.4155 - 4.2807x_i$ a druhou složku $\widehat{y}_{2i} = 3080.4644 + 15.5368x_i$.

Samotná dokumentace programu R však uvádí, že k odhadu parametrů směsi regresí jsou vhodnější funkce a metody obsažené v balíčku `flexmix`. Vhodnou funkcí pro naše účely je funkce `initFlexmix`. Tato funkce umožňuje několikrát opakovat EM algoritmus pro každý zadaný počet komponent, načež pro každý počet vybere model s maximální věrohodností. Základní parametry této procedury, které v této práci využijeme, jsou

`formule` symbolický zápis modelu.

`data` možnost zvolení datasetu, z kterého čerpáme hodnoty proměnných.

`concomitant` objekt třídy FLXP. Udává, jaký model pro doprovodnou proměnnou budeme uvažovat. Implicitně je zadán model `FLXPconstant`, který pracuje s konstantními marginálními pravděpodobnostmi. Další volbou je `FLXPmultinom(formula= 1)` reprezentující multinomický logitový model.

`control` objekt třídy `FLXcontrol`. Udává seznam příkazů kontrolujících průběh procedury. Mezi tyto příkazy patří například způsob klasifikace dat (zde lze zařadit S-krok a C-krok), maximální počet iterací či ukončovací kritérium.

`k` vektor obsahující pozitivní čísla, která reprezentují počty složek, pro které algoritmus bude spuštěn.

`nrep` pro každý počet složek k bude algoritmus proveden právě `nrep`-krát a ponechává si pouze odhadnutý model s maximální věrohodností.

Pokusíme se aplikovat funkci `initFlexmix` na naše data a analogicky jako v předchozím případě zadáme `k=(1:3)` a necháme algoritmus určit ideální počet složek. Navíc však přidáme počet opakování algoritmu pro každý počet z `k` pomocí parametru `nrep`. Znovu spuštění algoritmu několikrát za sebou zvýší pravděpodobnost, že alespoň v jednom případě EM algoritmus dokonverguje do globálního maxima věrohodnostní funkce.

```
> flex <- initFlexmix(y ~ x, nrep = 5, k = c(1,2,3), data = dataset)
1 : * * * * *
2 : * * * * *
3 : * * * * *
* * *
> flex

Call:
initFlexmix(y ~ x, data = dataset, k = c(1, 2, 3), nrep = 5)

  iter converged k k0    logLik      AIC      BIC      ICL
1    2      TRUE 1  1 -3316.669 6639.339 6651.633 6651.633
2   19      TRUE 2  2 -3215.197 6444.393 6473.080 6573.779
3   36      TRUE 3  3 -3213.278 6448.556 6493.635 6704.818
```

Podle výstupu, který nám funkce primárně poskytla, je zřejmé, že ideální počet složek je `k=2`, jeho věrohodnost sice není maximální, avšak to je pochopitelné vzhledem k faktu, že věrohodnost bude pro rostoucí počet složek vždy stoupat. Směrodatné jsou v tomto případě hodnoty Akaikeho a Byesovského informačního kritéria, jež dosahují minima právě pro směs dvou regresí. Je tedy naším zájmem právě tento model prozkoumat blíže, proto příkazem `getModel()` zvolíme pro zobrazení pouze model s dvěma složkami. Nejprve nám program R zobrazí tabulku s proporcemi jednotlivých složek (sloupec `prior`) a počtem pozorování do složek přiřazených (sloupec `size`). Dále v tabulce nalezneme sloupec `post>0` udávající pro každou složku počet pozorování, pro která je podmíněná pravděpodobnost, že dané pozorování náleží do i-té složky, větší než δ (implicitně $\delta = 10^{-4}$). V posledním sloupci lze nalézt poměr počtu pozorování ze `size` a pozorování ze sloupce `post>0`. Pomocí funkce `refit()` lze získat další podrobnosti

o modelu.

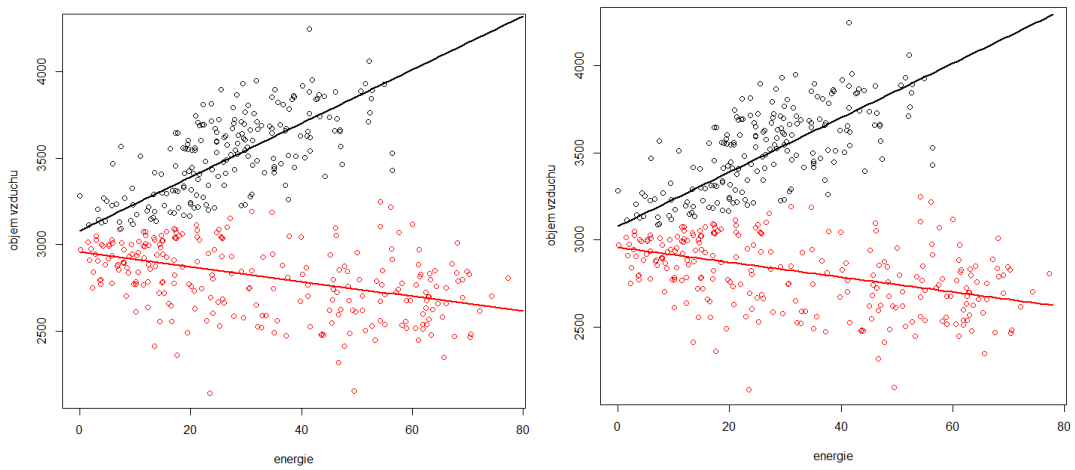
```
> fit=getModel(flex, which=2)
> summary(fit)

Call:
initFlexmix(y ~ x, data = dataset, k = 2, nrep = 5)

       prior size post>0 ratio
Comp.1 0.448  194    335 0.579
Comp.2 0.552  251    408 0.615

'log Lik.' -3215.197 (df=7)
AIC: 6444.393   BIC: 6473.08

> summary(refit(fit))
$Comp.1
            Estimate Std. Error z value  Pr(>|z|)
(Intercept) 3077.232     46.883  65.636 < 2.2e-16 ***
x             15.624      1.417  11.027 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

$Comp.2
              Estimate Std. Error z value  Pr(>|z|)
(Intercept) 2955.92604   31.49177 93.8635 < 2.2e-16 ***
x             -4.27281    0.69434 -6.1537 7.568e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Procedura `initFlexmix` kromě základních informací o složkách a proporcích nabízí také detaily obou regresních modelů, což může být značnou výhodou. Pokud navíc chceme informace o rozdělení jednotlivých pozorování do složek na základě podmíněných pravděpodobností, můžeme toho lehce dosáhnout pomocí příkazu `clusters()`, který poskytne číslo složky, do které je pozorování $1, \ldots, n$ přiřazeno. Funkce `posterior()` potom zobrazí pro každé pozorování pravděpodobnost příslušnosti k jedné či druhé složce.

```
> clusters(fit)
  [1] 2 2 2 2 2 1 1 1 1 1 1 2 2 2 2 2 1 1 1 1 1 2 2 2 2 2 1 1 2 2 2 2 2 2 2

                                        ⋮

[408] 2 2 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 2 2 2 2
[445] 2
```

```
> posterior(fit)
                [,1]         [,2]
  [1,] 5.204984e-09 1.000000e+00
  [2,] 5.323314e-03 9.946767e-01
  [3,] 3.652405e-03 9.963476e-01
  [4,] 1.272171e-06 9.999987e-01
  [5,] 1.078326e-08 1.000000e+00
  [6,] 9.987836e-01 1.216438e-03
  [7,] 9.996218e-01 3.782174e-04
  [8,] 9.999966e-01 3.404020e-06
  [9,] 9.999756e-01 2.437861e-05
 [10,] 9.999997e-01 2.818729e-07

                  ⋮

[445,] 3.240439e-03 9.967596e-01
```

Pokud si situaci vyjádříme graficky, můžeme pozorovat rozdělení dat do shluků a odhad regresní funkce pro každý z těchto shluků. Takto můžeme porovnat výsledky metody `regmix` a `initFlexmix`, jak lze pozorovat na obrázku 5.7. Již z pouhého pohledu lze soudit, že shluky jsou téměř totožné za použití jakékoli funkce, totéž platí pro odhady parametrů.



Obr. 5.7: Porovnání grafů pro rozdělení dat a odhad regresních funkcí pomocí funkce `regmix` (vlevo) a `initFlexmix` (vpravo).

Dále můžeme přistoupit k metodám predikování hodnot závislé proměnné, které pořídíme pro nová data $\mathbf{x} = (0, 8.7, 17.3, 26, 34.7, 43.3, 52, 60.7, 69.3, 78)$.

```
> X <- seq(0, 78, length.out = 10)
> predict(fit, newdata = data.frame(x = X))
```

Pomocí výše uvedených příkazů program R vypočítá predikci vysvětlované proměnné pro každou složku. Program tedy vezme hodnotu x_i a spočítá pro ni předpověď $\widehat{y}_{1i}$ podle první regresní funkce a předpověď $\widehat{y}_{2i}$ podle funkce druhé. Tyto predikce jsou zanesené v tabulce 5.1. Dále skrze váhy komponent víme, s jakou pravděpodobností bude dané pozorování náležet do jedné či druhé složky. Program R se tedy odchyluje od běžného pojetí predikce v klasické směsi regresí, jak jsme ho představili ve vztahu (26), avšak výsledek je podobně nejasný, protože nemůžeme nijak rozhodnout, do které složky pozorování zařadit - můžeme pouze odhadovat na základě vah složek. [4]

x	$\widehat{y}_1$	$\widehat{y}_2$
0	3077,2	2955,831
8,666667	3212,536	2918,795
17,33333	3347,871	2881,76
26	3483,207	2844,725
34,66667	3618,542	2807,69
43,33333	3753,878	2770,655
52	3889,214	2733,62
60,66667	4024,549	2696,584
69,33333	4159,885	2659,549
78	4295,22	2622,514

Tabulka 5.1: Nová pozorování x a jim příslušící předpověď proměnné Y.

Dalším přístupem k odhadu parametrů regresních modelů pro jednotlivé složky by mohla být metoda založená na analýze závisle proměnné. Jak bylo demonstrováno v kapitole (5.1), po spuštění EM algoritmu pro vysvětlovanou proměnnou

bylo jedním z výstupů také rozdělení dat do jednotlivých složek. Nabízí se tedy otázka, zda místo aplikace EM algoritmu na směs regresí nestačí rozdělit data na skupiny podle těchto klasifikačních údajů a poté provést klasickou lineární regresi pro každý ze shluků pozorování. Pravdou však je, že takto pořízené odhady by byly velmi nepřesné, protože by nevycházely z informace o náhodných chybách či podmíněné pravděpodobnosti, ale pouze z analýzy závisle proměnné. Je tedy zjevné, že tak by došlo ke ztrátě informace. Zároveň by v odhadu parametrů jednoho regresního modelu chyběla informace poskytnutá druhým shlukem dat. Ze vztahu (24) je zřejmé, že EM algoritmus pro směs regresních modelů tuto informaci při odhadu parametrů využívá, je to zřejmé z toho, že odhady pro i-tou složku pracují v podstatě se všemi pozorováními, pouze některá z nich penalizují skrze matici vah. Ačkoliv se tedy tento postup jeví intuitivním, v praxi by způsobil značnou nepřesnost odhadů.

5.3 Směs regresních modelů doprovodné proměnné

Teoretický základ k této části byl položen v kapitole 4.1. Tato teorie pracuje s doprovodnou proměnnou $\boldsymbol{\omega}$, která poskytuje vnější informaci, jež by mohla pomoci s budoucí prací s odhadnutým modelem. Někdy se jedná o samostatnou veličinu, avšak v našem případě je možné jako doprovodnou veličinu použít vysvětlující proměnnou. Budeme tedy uvažovat model

$$f(y_j|\mathbf{x}_j) = \sum_{i=1}^{2} \pi_i(\mathbf{x}_j, \boldsymbol{\alpha}_i)\phi(y_j|\mathbf{x}_j^{\mathrm{T}}\boldsymbol{\beta}_i, \sigma_i^2).$$

Pro marginální pravděpodobnosti $\pi_i(\mathbf{x}_j, \boldsymbol{\alpha}_i)$ použijeme multinomický logitový model, který zapíšeme pomocí vztahu

$$\pi_i(\mathbf{x}_j, \boldsymbol{\alpha}_i) = \frac{\exp^{\mathbf{x}_j^{\mathrm{T}}\boldsymbol{\alpha}_i}}{\sum_{h=1}^{2} \exp^{\mathbf{x}_j^{\mathrm{T}}\boldsymbol{\alpha}_h}} \qquad \text{pro} \qquad i = 1, \ldots, c. \tag{27}$$

Použitím tohoto modelu opět dostaneme dvě predikce hodnoty Y_i, tedy $\widehat{y}_{1i}$ odpovídající regresnímu modelu první složky a předpověď $\widehat{y}_{2i}$ odpovídající regresnímu modelu druhé složky. Rozdíl přichází v podobě marginálních pravděpodobností, které udávají, s jakou pravděpodobností bude pozorování náležet do první či druhé složky - a to pouze na základě hodnoty x_i. Jako výhodnější potom zvolíme

předpověď z té složky, pro kterou je marginální pravděpodobnost příslušnosti daného pozorování vyšší.

Odhady parametrů výše uvedené směsi regresních modelů doprovodné proměnné je možné získat opět pomocí funkce `initFlexmix`, která má možnost pracovat mimo jiné i se směsí regresních modelů doprovodné proměnné, která byla představena v kapitole 4.1. Model bude odpovídat teoretickému základu této metody a kód pro R software bude následující:

```
> fit2 <- initFlexmix(y ~ x, nrep = 5, k = 2, data = dataset,
+                     concomitant = FLXPmultinom(~x))
```

Pro stručnější zápis se omezíme již pouze na model s dvěma komponentami, naopak přidáme příkaz `concomitant = FLXPmultinom( x)`, kterým programu R sdělíme, že chceme použít směs regresních modelů doprovodné proměnné, kterou je v tomto případě vysvětlující proměnná podléhající multinomickému logitovému modelu. Nejprve zavoláme funkci `initFlexmix` a necháme si zobrazit základní informace o počtu iterací a velikosti shluků. Pomocí `summary()` lze obdržet další informace o směsi a jejích složkách, avšak pokud chceme blíže prozkoumat regresní modely pro obě složky, je potřeba stejně jako v předešlém příkladě využít funkci `refit()`.

```
> fit2

Call:
initFlexmix(y ~ x, data = dataset, concomitant = FLXPmultinom(~x),
    k = 2, nrep = 5)

Cluster sizes:
  1   2
295 150

convergence after 25 iterations
> summary(fit2)

Call:
initFlexmix(y ~ x, data = dataset, concomitant = FLXPmultinom(~x),
    k = 2, nrep = 5)

       prior size post>0 ratio
Comp.1 0.651  295    353 0.836
Comp.2 0.349  150    343 0.437

'log Lik.' -3180.064 (df=8)
```

```
AIC: 6376.127   BIC: 6408.912
```

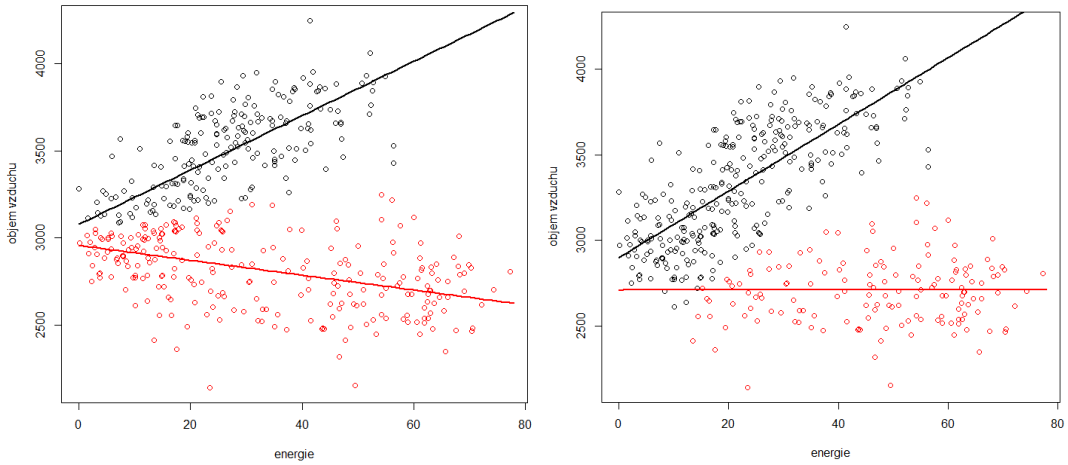
```
> rfit2 <- refit(fit2)
> summary(rfit2)
$Comp.1
             Estimate Std. Error z value  Pr(>|z|)
(Intercept) 2898.0120    30.5450  94.877 < 2.2e-16 ***
x             19.4791     1.1005  17.700 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

$Comp.2
              Estimate Std. Error z value Pr(>|z|)
(Intercept) 2.7081e+03 6.1943e+01  43.719   <2e-16 ***
x           1.5225e-02 1.1752e+00   0.013   0.9897
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Můžeme si též nechat zobrazit odhady parametrů pro doprovodnou proměnnou, když argumentem `which` specifikujeme, že se zajímáme právě o tento model.

```
> summary(rfit2, which = "concomitant")
$Comp.2
             Estimate Std. Error z value  Pr(>|z|)
(Intercept) -3.684427   0.557854 -6.6046 3.985e-11 ***
x            0.091593   0.012918  7.0901 1.340e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Odhady regresních parametrů jsou v tomto případě očividně odlišné, stejně tak se liší rozdělení pozorování do složek smíšených regresí, jak lze vidět na obrázku 5.8. Směs modelů doprovodné proměnné porovnáváme s klasickou směsí regresí, čímž myslíme směs bez doprovodné proměnné - pro naše potřeby zvolíme model vypočítaný pomocí funkce `initFlexmix`.

Nespornou výhodou tohoto postupu je možnost zařadit pozorování do jedné ze složek na základě vah složek, které je tato funkce schopna spočítat pro každé nové pozorování, a to pouze na základě hodnot proměnné x. Nechť nové hodnoty proměnné x jsou totožné s novými hodnotami v předchozí kapitole, potom pravděpodobnosti příslušnosti k jednotlivým složkám pro každé pozorování vypočítáme pomocí příkazu `prior`:



Obr. 5.8: Porovnání grafů pro rozdělení dat a odhad regresních funkcí pomocí klasické směsi regresních modelů (vlevo) a směsi regresních modelů doprovodné proměnné (vpravo).

```
> prior(fit2, newdata = data.frame(x = X))
             1          2
1   0.97550357 0.02449643
2   0.94737366 0.05262634
3   0.89056519 0.10943481
4   0.78626849 0.21373151
5   0.62448483 0.37551517
6   0.42914966 0.57085034
7   0.25364402 0.74635598
8   0.13316953 0.86683047
9   0.06493860 0.93506140
10  0.03043898 0.96956102
```

Tyto hodnoty jsou vypočteny na základě vztahu (27), kde implicitně $\boldsymbol{\alpha}_1 = \mathbf{0}$ a $\boldsymbol{\alpha}_2 = (-3.6844, 0.0916)^{\mathrm{T}}$ je maximálně věrohodný odhad vektoru $\boldsymbol{\alpha}_2$. Dále můžeme predikovat hodnoty Y v každé složce

```
> predict(fit2, newdata = data.frame(x = X))

$Comp.1
        [,1]
1   2898.012
2   3066.831
3   3235.650
```

```
4  3404.468
5  3573.287
6  3742.106
7  3910.925
8  4079.744
9  4248.562
10 4417.381

$Comp.2
       [,1]
1  2708.081
2  2708.213
3  2708.345
4  2708.477
5  2708.609
6  2708.741
7  2708.873
8  2709.005
9  2709.136
10 2709.268
```

a pro každé nové pozorování vybrat složku, pro kterou svědčí vyšší pravděpodobnost příslušnosti. [4]

5.4 Klasifikační a stochastický EM algoritmus

Posledními metodami, které bude tato práce demonstrovat, jsou klasifikační a stochastický EM algoritmus. Oba algoritmy lze spustit pomocí jednoduché modifikace funkce `initFlexmix`, kde přidáme parametr `control` a pomocí něj zařadíme do algoritmu C-krok či S-krok. Zároveň je pro tento účel nutné zařadit do funkce argument `model`, kterým zadefinujeme parametrickou třídu distribucí, ze které bude algoritmus vycházet. V našem případě předpokládáme normální rozdělení složek, tudíž zvolíme `family="gaussian"`. Protože základní informace o algoritmu jsou opět pouze elementární, necháme si detailní výstupy vypsat pomocí `summary()`. Ani tato funkce však neposkytne informace o regresních modelech a v případě klasifikačního a stochastického algoritmu nelze využít funkce `refit()`, proto si o parametry regresních funkcí řekneme přímo příkazem `parameters()`.

```
> fitCE <- initFlexmix(y ~ x, nrep = 5, k = 2, data = dataset,
+          model = FLXMRglm(~x,family="gaussian") ,control=list(classify="CEM"))
2 : * * * * *
> fitCE
```

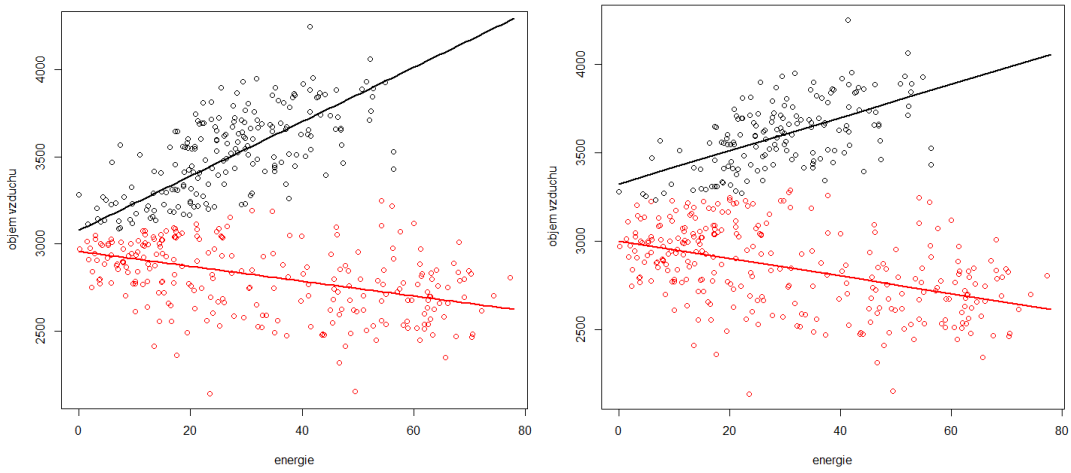
```
Call:
initFlexmix(y ~ x, data = dataset, model = FLXMRglm(~x, family = "gaussian"),
    k = 2, control = list(classify = "CEM"), nrep = 5)

Cluster sizes:
  1   2
157 288

convergence after 7 iterations
> summary(fitCE)

Call:
initFlexmix(y ~ x, data = dataset, model = FLXMRglm(~x, family = "gaussian"),
    k = 2, control = list(classify = "CEM"), nrep = 5)

       prior size post>0 ratio
Comp.1 0.359  157    295 0.532
Comp.2 0.641  288    411 0.701

'log Lik.' -3226.624 (df=7)
AIC: 6467.247   BIC: 6495.934

> parameters(fitCE)
                      Comp.1      Comp.2
coef.(Intercept) 3323.641116 3001.271901
coef.x              9.396995   -4.943646
sigma             158.902161  206.894326
```

Odhadnuté parametry se v tomto případě opět poněkud liší od odhadů pořízených předchozími metodami. Pro názornost graficky porovnejme regresní funkce získané touto metodou a klasickou směsí regresí v obrázku 5.9.

Analogicky tomu bude při použití stochastického EM algoritmu, taktéž zadaná funkce a výstupy budou v podobném tvaru.

```
> fitSE <- initFlexmix(y ~ x, nrep = 5, k = 2, data = dataset,
+                    model = FLXMRglm(~x,family="gaussian") ,control=list(classify="SEM"))
2 : * * * * *
> fitSE

Call:
initFlexmix(y ~ x, data = dataset, model = FLXMRglm(~x, family = "gaussian"),
    k = 2, control = list(classify = "SEM"), nrep = 5)

Cluster sizes:
```

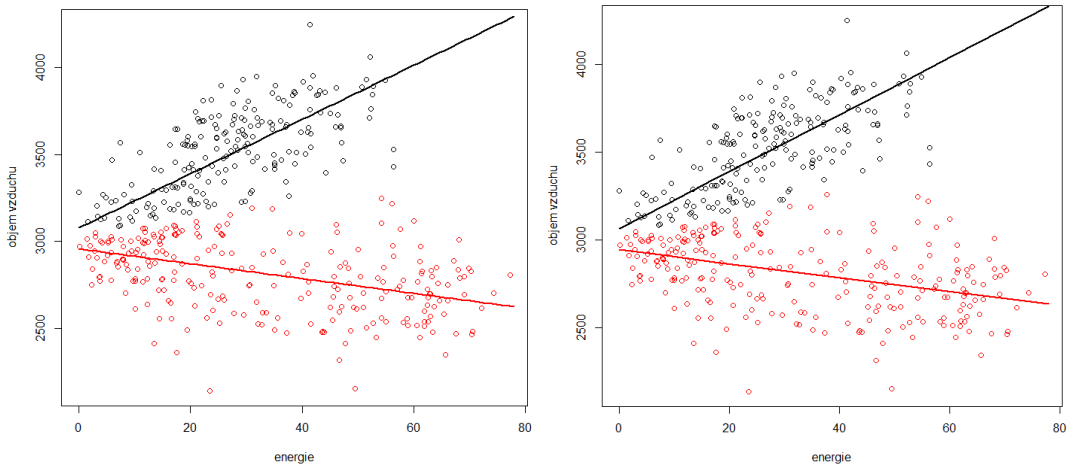
```
  1   2
193 252

convergence after 34 iterations
> summary(fitSE)

Call:
initFlexmix(y ~ x, data = dataset, model = FLXMRglm(~x, family = "gaussian"),
    k = 2, control = list(classify = "SEM"), nrep = 5)

       prior size post>0 ratio
Comp.1 0.449  193    331 0.583
Comp.2 0.551  252    411 0.613

'log Lik.' -3215.636 (df=7)
AIC: 6445.272   BIC: 6473.958

> parameters(fitSE)
                     Comp.1      Comp.2
coef.(Intercept) 3064.76278 2943.449170
coef.x             16.22244   -3.914172
sigma             193.37561  212.403852
```



Obr. 5.9: Porovnání grafů pro rozdělení dat a odhad regresních funkcí získaných pomocí klasického EM algoritmu (vlevo) a klasifikačního EM algoritmu (vpravo).

Výsledné odhady opět porovnáme se standardní směsí regresí, kterou považujeme za jakýsi základ v problematice odhadování parametrů směsi regresí. Oba

grafy jsou k nalezení v obrázku 5.10 a lze z nich vyčíst, že tentokrát jsou odhady pořízené pomocí SEM velmi blízké těm původním. [4] [9]



Obr. 5.10: Porovnání grafů pro rozdělení dat a odhad regresních funkcí získaných pomocí klasického EM algoritmu (vlevo) a stochastického EM algoritmu (vpravo).

5.5 Porovnání jednotlivých modelů

Na závěr je vhodné pro větší přehlednost uvést všechny modely společně s jejich odhady parametrů a charakteristikami hodnotícími jejich kvalitu. Tabulka 5.2 obsahuje každý stěžejní demonstrovaný model a pro něj odhady regresních parametrů pro obě složky, směrodatné odchylky těchto parametrů, odhad parametru σ a statistiky hodnotící celý model, tedy Bayesovské informační kritérium a logaritmickou věrohodnost modelu.

V tabulce 5.2 nalezneme pět základních metod, rozdělení na složky a jim příslušící odhady. Charakteristiky sloužící k porovnání celého modelu, kterými jsou v našem případě věrohodnost modelu a Bayesovské informační kritérium, lze využít v případě, že k odhadu parametrů používáme maximalizaci věrohodnostní funkce stejného tvaru. To můžeme prohlásit o klasické směsi regresí, kde jsme použili buď funkci `regmix` či `initFlexmix`, ale též o klasifikačním či stochastickém EM algoritmu, kde maximalizujeme stále tu samou věrohodnostní funkci,

pouze měníme postup maximalizace přidáním C- a S-kroku. Podle těchto statistik jsou nejkvalitnější modely odhadnuté pomocí standardního EM algoritmu a funkcí `regmix` a `initFlexmix`. Z tabulky lze vyčíst, že jejich věrohodnost je shodná a BIC se liší jen minimálně, to odpovídá skutečnosti, že i jejich parametry jsou velmi blízké. Těmto hodnotám se též blíží odhady pořízené pomocí SEM algoritmu, jehož věrohodnost je jen nepatrně menší než pro klasickou regresi a Bayesovské informační kritérium je rovněž nižší pouze o pár desetin. Algoritmus CEM v tomto ohledu poskytuje nejméně výstižné odhady.

	sl.	váha	β_0	β_1	σ	$\sqrt{\text{var}(\beta_0)}$	$\sqrt{\text{var}(\beta_1)}$	log. věrohodnost	BIC
klas. směs (`flexmix`)	1	0.448	3077.23	15.62	199.09	46.88	1.42	-3215.2	6473.08
	2	0.552	2955.93	-4.27	208.73	31.49	0.69		
klas. směs (`regmix`)	1	0.447	3080.47	15.54	198.05	-	-	-3215.2	6473.07
	2	0.553	2956.42	-4.28	208.16	-	-		
doprovodná proměnná	1	0.651	2898.01	19.48	230.87	30.55	1.1	-3180.06	6408.91
	2	0.349	2708.1	0.015	194.15	61.94	1.18		
CEM	1	0.359	3323.64	9.4	158.90	-	-	-3226.62	6495.93
	2	0.641	3001.27	-4.94	206.89	-	-		
SEM	1	0.449	3064.76	16.22	193.38	-	-	-3215.64	6473.96
	2	0.551	2943.45	-3.91	212.40	-	-		

Tabulka 5.2: Přehled charakteristik jednotlivých modelů.

Dále lze porovnávat směrodatnou odchylku regresních parametrů, kterou bereme jako ukazatele přesnosti odhadnutých parametrů. Údaje o směrodatných odchylkách poskytuje funkce `flexmix` pouze pro případ klasické směsi a směsi regresních modelů doprovodné proměnné. Jelikož jsme však klasickou směs regresí dříve vyhodnotili jako nejkvalitnější podle hodnot věrohodnosti a BIC, můžeme ji brát jako základní metodu k porovnávání. Podle směrodatných odchylek je potom patrné, že modely nelze takto srovnat, když pro první složku je výhodnější model s doprovodnou proměnnou a pro druhou naopak klasická směs.

Dále lze k porovnání použitých metod využít odhad sigma, tedy odhad směrodatné odchylky náhodných chyb. Podle těchto údajů lze soudit, že nejvíce chybový

je model směsi regresních modelů doprovodné proměnné a překvapivě přesný se naopak jeví model pořízený pomocí CEM algoritmu.

Přes všechny tyto poznatky je ovšem dobré si připomenout smysl použití jednotlivých procedur. Například model využívající doprovodnou proměnnou se může jevit jako značně nepřesný, avšak je třeba si uvědomit, že důvod k použití tohoto modelu tkví především ve zlepšení možnosti predikce. Podobně je třeba uvažovat také o stochastickém a klasifikačním EM algoritmu. Účelem těchto metod je dosáhnout lepší konvergence algoritmu, takže i přes to, že v tomto případě se tyto metody nemusí jevit nejlépe, je třeba uvážit jejich nespornou výhodu v případě, kdy máme k dispozici problémová data a máme obavy o konvergenci algoritmu. Zároveň tyto algoritmy konvergenci iterační procedury značně zrychlují. Grün a Leisch ([9]) navrhují postup, při kterém jsou nejprve použity algoritmy SEM a CEM, z jejichž odhadů vybereme ten nejkvalitnější a ten následně použijeme jako počáteční řešení pro EM algoritmus. Tento postup by měl zaručit rychlou konvergenci algoritmu k vhodnému lokálnímu maximu.

Závěr

Závěrem lze říci, že problematika smíšených distribucí je velmi obsáhlá a tato práce obsahuje opravdu jen její zlomek. Celá práce byla soustředěna nejprve na úvod do tématu smíšených distribucí a jejich vlastností, poté byly představeny charakteristické příklady takových směsí. Dále byl důraz kladen na EM algoritmus a jeho modifikace související s rychlostí a kvalitou jeho konvergence.

Po tomto nezbytném úvodu bylo možné přistoupit ke směsi regresních modelů, pro kterou jsme si předvedli princip fungování EM algoritmu, jež je spojením klasického EM algoritmu a základních vztahů z regresní analýzy. Pro lepší možnost predikcí jsme navázali se směsí regresních modelů doprovodné proměnné a saturovanou směsí regresí.

Závěrečný úsek práce obsahuje aplikaci dříve zmíněných metod na reálných datech poskytnutých firmou SECO Group a.s., nejprve byl demonstrován výpočet odhadů parametrů směsi normálních hustot a následně jsme postoupili ke směsi regresních modelů - s i bez doprovodné proměnné. Hlavním přínosem této sekce je prezentace funkcí, pomocí kterých se v programu R lze s podobnými daty vyrovnat.

Celá praktická část byla zakončena porovnáním užitých metod a přehledným výčtem získaných parametrů a charakteristik odhadovaných modelů. V součtu tedy práce poskytla ucelený pohled na problematiku smíšených distribucí a směsí regresí skládajících se z normálně rozdělených složek.

Příloha 1

Tabulka obsahující hodnoty závisle a nezávisle proměnné. Energie dodaná do kompresoru měřená v kilowatthodinách představuje vysvětlující proměnnou, metry krychlové stlačeného vzduchu potom reprezentují vysvětlovanou proměnnou. Obě veličiny jsou měřeny v hodinových intervalech.

pozor.	energie (kWh)	vzduch (m^3)	pozor.	energie (kWh)	vzduch (m^3)	pozor.	energie (kWh)	vzduch (m^3)
1	63,08	2830,12	21	45,08	3738,36	41	42,6	3839,59
2	16,04	2656,59	22	36,28	3468,09	42	52,76	3891,67
3	20,68	2729,97	23	21,52	3407,13	43	46,16	3885,27
4	32,72	2518,63	24	26,6	3100,13	44	52,64	3844,31
5	46,6	2594,29	25	25,12	3088,29	45	44,16	3859,55
6	47,08	3569,32	26	15,64	3073,72	46	46,92	3663,27
7	34,72	3644,93	27	2,84	2946,33	47	32,76	3687,95
8	35,16	3897,33	28	9,44	2946,14	48	34,72	3187,5
9	43,32	3740,13	29	9,56	3115,79	49	10,68	3016,65
10	41,96	3951,71	30	20,08	3318,2	50	13,16	3157,37
11	47,28	3464,2	31	10,2	2934,01	51	14,88	3023,57
12	20,16	3546,73	32	24,04	2604,03	52	10,48	2839,89
13	24,2	2971,81	33	54,68	2849,23	53	3,36	3142,65
14	27,36	2931,67	34	3,76	2798,59	54	22,8	3230,7
15	5,76	2936,48	35	2,44	2840,21	55	17,04	2909,8
16	12,2	2783,53	36	16,92	2776,41	56	14,36	2556,25
17	61,32	2813,32	37	12,28	2634,74	57	50	2954,03
18	37,04	3657,97	38	45,4	2543,71	58	7,92	2891,4
19	22,32	3693,5	39	70,36	2463,72	59	15,28	2716,87
20	37,72	3785,6	40	67,16	2889,96	60	10,04	2772,63

pozor.	energie (kWh)	vzduch (m^3)	pozor.	energie (kWh)	vzduch (m^3)	pozor.	energie (kWh)	vzduch (m^3)
61	3,8	2769,13	86	62,4	2630,48	111	51,52	2503,92
62	5,84	2776,23	87	55,08	2534,88	112	62,84	2697,47
63	13,52	2412,74	88	48,4	2755,92	113	40,64	3500,96
64	54,52	2765,91	89	41,6	3540,86	114	27,2	3521,08
65	50,76	3888,25	90	28,36	3867,9	115	20,76	3545,98
66	43,16	3866,71	91	31,96	3708,83	116	23,2	3200,33
67	38,52	3861,48	92	31,36	3757,18	117	7,4	3569,8
68	42,84	3844	93	26,28	3629,39	118	25,88	3479,11
69	40,2	3920,81	94	28,48	3586,53	119	17,68	3308,99
70	41,52	3620,67	95	16,68	3311,28	120	6,76	2877
71	27,52	3671,87	96	20,6	3213,6	121	40,88	2864,07
72	29,76	3230,39	97	21,6	3027,37	122	7,96	2895,13
73	12,16	3195,77	98	14,56	2983,76	123	37,72	3048,51
74	23,8	3434,31	99	7,56	3007,74	124	32,92	2588,05
75	18,68	3167,43	100	2,04	2973,2	125	31,16	3600,56
76	7,44	3091,58	101	1,56	3111,17	126	46,36	3727,01
77	30,96	3189,41	102	13,32	3289,29	127	52,08	3710,86
78	4,84	3252,64	103	7,92	2969,38	128	52,32	3761,95
79	12,36	2884,61	104	17,52	2360,11	129	56,36	3527,58
80	45,08	2620,25	105	54,2	2670,15	130	38,92	3513,71
81	61,28	2968,88	106	23,16	2949,12	131	37,68	3258,49
82	47,72	2875,5	107	23,64	2693,05	132	0	3281,81
83	61,88	2598,77	108	18,52	2868,43	133	5,84	3225,11
84	59	2676,3	109	16,44	2646,5	134	11,48	3072,18
85	59,56	2673,82	110	39,76	2672,95	135	12,72	2857,95

pozor.	energie (kWh)	vzduch (m^3)	pozor.	energie (kWh)	vzduch (m^3)	pozor.	energie (kWh)	vzduch (m^3)
136	3,88	3126,57	161	21,16	3043,75	186	9,64	2770,18
137	7,12	3132,04	162	19,96	3244,72	187	31,76	2648,87
138	3	3026,47	163	7,04	2852,5	188	33	2927,46
139	4,52	2883,82	164	56,08	3219,01	189	1,72	2910,88
140	10,48	2954,92	165	46,12	3095,74	190	10,04	2783,49
141	3,92	2999,18	166	14,76	2780,18	191	17,32	3069,67
142	8,4	2701,85	167	61,16	2885,61	192	14,96	2945,61
143	7,84	2897,25	168	69,12	2798,62	193	0,16	2970,18
144	20,72	2626,18	169	69,8	2844,53	194	10,04	2610,35
145	77,32	2806,01	170	69,04	2691,33	195	44	2476,48
146	27,72	2582,63	171	51,04	2703,67	196	44,2	3394,08
147	51,04	2618,16	172	46,88	3654,74	197	18,96	3553,5
148	46,04	3656,82	173	26,4	3489,6	198	20,84	3747,54
149	30,32	3807,25	174	21,88	3602,7	199	34,28	3685,89
150	32,08	3665,99	175	24,68	3651,24	200	29,6	3696,08
151	29,48	3617,55	176	22,28	3809,71	201	26,08	3619,56
152	21,52	3690,85	177	28,4	3523,17	202	17,24	3457,19
153	20,24	3447,41	178	14,68	3406,12	203	18,8	3330,26
154	13,52	3414,71	179	18,68	3336,52	204	15,28	3225,38
155	13,52	3024,39	180	24,24	3214,37	205	1,4	3012,08
156	9,2	3140,07	181	16,96	3181,44	206	39,92	3043,55
157	13,4	3041,43	182	13,64	2875,28	207	46,8	2854,19
158	24,72	3071,68	183	5,2	2993,24	208	60,04	3117,35
159	54,2	3103,77	184	4,52	3134,87	209	21,24	3215,79
160	27,16	3151,87	185	9,56	3229,05	210	25,72	3034,96

pozor.	energie (kWh)	vzduch (m^3)	pozor.	energie (kWh)	vzduch (m^3)	pozor.	energie (kWh)	vzduch (m^3)
211	49,24	2621,06	236	30,72	3662,44	261	16,28	3553,45
212	40,2	2829,37	237	17,6	3644,57	262	19,68	3426,11
213	68,12	3007,66	238	35	3416,65	263	29,16	3227,79
214	57,08	2722,5	239	8,04	3271,06	264	7,12	3086,6
215	64	2848,51	240	4,28	3268,4	265	4,2	2992,88
216	70,28	2828,62	241	11,8	3054,08	266	3,68	2794,6
217	68,24	2787,68	242	34,8	2944,08	267	23,8	2821,13
218	63,44	2477,92	243	54,2	3247,22	268	14,64	3054,63
219	41,16	2703,54	244	9,44	3323,45	269	6,84	2881,6
220	33,32	3415,25	245	11,48	2993,56	270	22,96	2559,81
221	23,68	3713,92	246	34,92	2746,93	271	53,56	2789,62
222	26,52	3575,41	247	58,12	2774,6	272	37,52	2811,38
223	24,64	3597,69	248	63,24	2798,74	273	25,16	2530,13
224	22,16	3694,96	249	62,48	2721,39	274	72,2	2613,53
225	26,32	3411,68	250	64,32	2758,45	275	70,48	2479,86
226	21,24	3377,33	251	60,56	2514,19	276	63,04	2565,92
227	24	2802,82	252	48,72	2484,85	277	67,6	2468,2
228	48,28	2842,55	253	46,64	2315,05	278	62,56	2677,34
229	56,2	2973,82	254	32,12	2523,59	279	41,12	3654,43
230	9,24	3000,24	255	39,76	3627,43	280	24,68	3534,56
231	17,12	2823,44	256	27,92	3723,01	281	23,72	3814,8
232	34,64	3854,94	257	24,36	3670,69	282	41,44	3884,85
233	35,72	3823,61	258	22,6	3483,33	283	38,76	3855,42
234	34,72	3671,36	259	18,56	3559,05	284	36,64	3809,87
235	17,28	3647,86	260	24,64	3593,77	285	27,88	3434,3

pozor.	energie (kWh)	vzduch (m^3)	pozor.	energie (kWh)	vzduch (m^3)	pozor.	energie (kWh)	vzduch (m^3)
286	21,12	2901,52	311	17,28	3310,78	336	8,84	2929,55
287	25,6	3039,62	312	6,64	3235,38	337	45,68	2795,41
288	14,92	3020,74	313	16,24	3188,63	338	58,16	2533,5
289	13,8	3132,72	314	21,12	3083,14	339	19,6	3601,88
290	15,6	2928,13	315	5,92	3010,72	340	19,56	3548,65
291	11,84	3002,91	316	19,76	3164,89	341	27,72	3701,5
292	25,44	3044,04	317	14,92	2784,06	342	27,64	3792,27
293	8,76	2835,05	318	33,6	2587,22	343	34,96	3500,13
294	48,6	2673,94	319	45,2	2937,99	344	19,44	3581,86
295	57,6	3071,45	320	11,56	2923,5	345	25,96	3344,24
296	27,4	2833,4	321	2,32	2747,39	346	13,48	3189,82
297	66,12	2660,32	322	30,6	2749,36	347	21	3113,03
298	62,48	2721,47	323	35,08	2489,53	348	22,2	3273,71
299	58,84	2554,04	324	65,44	2578,62	349	11,72	2986,95
300	49,8	2611,76	325	65,72	2344,65	350	2,92	3046,15
301	53,24	2445,81	326	54,24	2560,19	351	19,96	3234,24
302	46,4	2724,54	327	35,08	3452,78	352	3,24	2906,88
303	29,2	3640,78	328	31,92	3487,37	353	14,28	2718,76
304	25,12	3722,56	329	20,72	3560,54	354	11,84	3075,12
305	29,44	3931,68	330	54,88	3929,37	355	18,44	3063,14
306	52,16	4063,58	331	38,24	3840,8	356	20,08	2758,11
307	41,4	4249,3	332	56,36	3431,25	357	25,68	2929,55
308	51,56	3933,88	333	18,84	3410,15	358	24,96	2666,32
309	36,08	3695,41	334	17,76	3212,61	359	41,08	2764,42
310	30,16	3563,76	335	17,2	3039,72	360	35,6	2560,18

pozor.	energie (kWh)	vzduch (m^3)	pozor.	energie (kWh)	vzduch (m^3)	pozor.	energie (kWh)	vzduch (m^3)
361	25,76	2686,42	386	40,96	3629,95	411	24,92	3396,58
362	29,8	3545,6	387	29,84	3608,61	412	31,16	3460,45
363	37,16	3671,56	388	28,52	3632,8	413	28,08	3547,39
364	31,76	3950,08	389	30,32	3511,92	414	29,12	3711,75
365	41,44	3759,53	390	25,52	3898,89	415	30,72	3276,56
366	29,8	3766,15	391	35,04	3436,53	416	14,8	3297,57
367	37,64	3518	392	10,96	3512,79	417	17,88	3052,87
368	15,68	3293,2	393	23,12	3340,85	418	5,24	3002,54
369	12,8	3149,5	394	10,24	3174,57	419	17,32	3036,69
370	5,8	3027,58	395	17,08	3041,54	420	29,08	2843,21
371	3,28	3202,62	396	14,36	2999,31	421	57,48	2738,6
372	17,04	3089,57	397	13,92	2939,66	422	68,04	2836,96
373	11,4	2766,48	398	5,12	2918,98	423	61,44	2829,33
374	56,36	2911,79	399	12,36	2997,26	424	63,68	2622,4
375	46,2	3051,06	400	17,28	3088,94	425	49,44	2151,96
376	30,52	2742,79	401	16,56	2555,28	426	23,48	2138,07
377	61,12	2447,52	402	36,52	2879,21	427	37,24	2473,05
378	45,76	2739,17	403	8,52	2868,14	428	47,76	2410,76
379	66,04	2748,25	404	19,44	2490,22	429	47,04	2623,24
380	64,6	2656,57	405	74,32	2700,86	430	43,64	2482,54
381	63	2698,26	406	26,04	2664,15	431	52,68	2530,11
382	61,84	2509,3	407	58,96	2513,35	432	41,52	2524,12
383	54,96	2711,32	408	43,52	2476,22	433	29,52	2574,2
384	61,96	2537,26	409	62,64	2536,05	434	37,12	3343,91
385	50,4	2798,83	410	38,64	3446,19	435	12,4	3217

pozor.	energie (kWh)	vzduch (m^3)
436	5,88	3469,19
437	20,52	3412,59
438	21,24	3706,82
439	30,96	3290,15
440	21,04	3285,3
441	19,64	2770,58
442	45,52	2678,11
443	60,68	2767,05
444	31,04	2848,68
445	21,96	2746,67

Příloha 2

Přiložené CD obsahuje soubory se vstupními daty, kódy z programu R potřebné k praktické části a kódy použité k vykreslení grafů v kapitole 1.4.

Literatura

[1] ANDĚL, Jiří. Základy matematické statistiky. Praha : Matfyzpress, 2005. ISBN 80-86732-40-1.

[2] BENAGLIA, T., CHAUVEAU, D., HUNTER, D.R. a YOUNG, D.S.. Mixtools: An R Package for Analyzing Finite Mixture Models. Journal of Statistical Software. 2009, roč. 32, č. 6. www.jstatsoft.org [online 27.2.2014]

[3] DESARBO, W.S. a CRON, W.L.. A Maximum Likelihood Methodology for Clusterwise Linear Regression. Journal of Classification. 1988, č. 5. Dostupné z: deepblue.lib.umich.edu [online 3.3.2014]

[4] Dokumentace balíčku 'flexmix'. Dostupné z: cran.r-project.org [online 3.3.2014]

[5] Dokumentace balíčku 'fpc'. Dostupné z: cran.r-project.org [online 3.3.2014]

[6] Dokumentace balíčku 'mixtools'. Dostupné z: cran.r-project.org [online 3.3.2014]

[7] FARIA, Susana, SOROMENHO, Gilda. Fitting mixtures of linear regressions. Journal of Statistical Computation and Simulation (Impact Factor: 0.63). 03/2010; 80:201-225. DOI:10.1080/00949650802590261. Dostupné z: http://www.researchgate.net [online 10. 2. 2014]

[8] FRÜHWIRTH-SCHNATTER, Sylvia. Finite Mixture and Markov Switching Models. New York: Springer, 2006. ISBN 0-387-32909-9.

[9] GRÜN, B. a LEISCH. F.. FlexMix Version 2: Finite Mixtures with Concomitant Variables and Varying and Constant Parameters. Journal of Statistical Software. 2008, roč. 28, č. 4. Dostupné z: cran.at.r-project.org [online 3.3.2014]

[10] HANZLÍČEK, Tomáš. Odhady parametrů smíšených distribučních funkcí náhodného vektoru. Olomouc, 2010. Diplomová. Univerzita Palackého v Olomouci. Vedoucí práce Prof. RNDr. Ing. Lubomír Kubáček, DrSc., Dr.h.c.

[11] HOSHIKAWA, Toshiya. Mixture regression for observational data, with application to functional regression models. 2013. Dostupné z: arxiv.org [online 10.3.2014]

[12] internetové stránky firmy Seco Group a.s. www.secogroup.cz [online 3.3.2014]

[13] MCLACHLAN, Geoffrey, PEEL, David. Finite mixture models. New York: John Wiley & Sons, 2000. ISBN 0-471-00626-2.

[14] OLIVEIRA-BROCHADO, A. and MARTINS, F.V., (2005), Assessing the Number of Components in Mixture Models: a Review, FEP Working Papers, Universidade do Porto, Faculdade de Economia do Porto, Dostupné z: econpapers.repec.org [online 25. 2. 2014]

[15] SAS, Marek. Konvexní množiny. Brno, 2011. Bakalářská. Masarykova Univerzita. Vedoucí práce prof. RNDr. Ondrej Došlý, DrSc.

[16] TITTERINGTON, D.M, MAKOV, E.M., SMITH, A.F.M. Statistical analysis of finite mixture distributions. Chichester: Wiley, 1985. ISBN 0-471-90763-4

[17] VAŇKÁTOVÁ, Kristýna. Metody odhadů regresních parametrů. Olomouc, 2012. Bakalářská. Univerzita Palackého v Olomouci. Vedoucí práce doc. RNDr. Eva Fišerová, Ph.D.

[18] WEDEL, M. Concomitant variables in finite mixture models. Statistica Neerlandica. 2002, roč. 56, č. 3. Dostupné z: onlinelibrary.wiley.com [online 3.3.2014]

[19] WHILEY, M. a TITTERINGTON, D. M. , Applying the deterministic annealing expectation maximisation algorithm to naive Bayesian networks, Univ. Glasgow Tech. Report, 2002. Dostupné z: citeseerx.ist.psu.edu [online 25. 2. 2014]