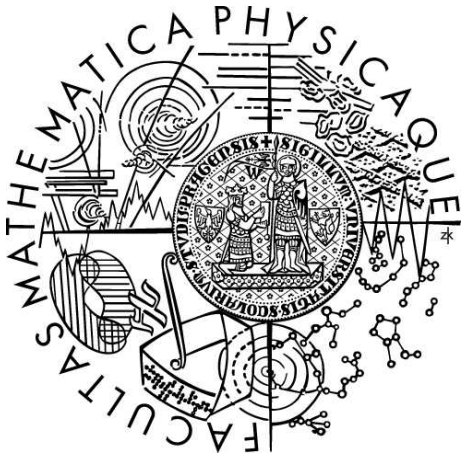
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE



Dita Rensová

Robustní klasifikace a diskriminace

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: RNDr. Jan Kalina, Ph.D.
Studijní program: Matematika
Studijní obor: Pravděpodobnost, matematická statistika a ekonometrie

Praha 2013

Na tomto místě bych ráda poděkovala vedoucímu mé diplomové práce RNDr. Janu Kalinovi, Ph.D. za cenné rady a ochotu, se kterou mi věnoval svůj čas.

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

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

V ..... dne ..... Dita Rensová

Název práce: Robustní klasifikace a diskriminace

Autor: Dita Rensová

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: RNDr. Jan Kalina, Ph.D., Ústav informatiky AV ČR, v.v.i.

e-mail vedoucího: kalina@cs.cas.cz

Abstrakt: V této práci se zabýváme modely klasifikační analýzy a jejich robustními obměnami. Nejprve popíšeme základní lineární a kvadratická klasifikační pravidla a uvedeme postupy, jak odhadnout pravděpodobnost špatné klasifikace. Poté se zaměříme na popis robustních mnohorozměrných odhadů, jejich vlastností a metod používaných pro jejich výpočet. Tyto odhady posléze použijeme k vytvoření robustních verzí klasifikačních pravidel. Dále si popíšeme analýzu hlavních komponent jako metodu pro redukci dimenze dat a budeme se zabývat i její robustifikací. Na závěr předvedeme použití robustní klasifikační analýzy v simulacích a na reálných datech. Ukážeme si také, jak tuto klasfikaci ovlivní použití analýzy hlavních komponent.

Klíčová slova: Mnohorozměrná statistika, lineární klasifikace, kvadratická klasifikace, robustní odhady, analýza hlavních komponent

Title: Robust classification and discrimination

Author: Dita Rensová

Department: Department of Probability and Mathematical Statistics

Supervisor: RNDr. Jan Kalina, Ph.D., Institute of Computer Science of the ASCR, v.v.i.

Supervisor's e-mail address: kalina@cs.cas.cz

Abstract: This thesis is focused on classification methods and their robust alternatives. First, we recall the standard classification rules of linear and quadratic discrimination analysis. We also show some methods for estimating their probability of missclassification. Next we describe some existing robust multivariate estimators, their properties and computational algorithms. These estimators are consequently used to construct robust classification rules. Then, we describe the principal component analysis as a technique for dimension reduction. Again, we study methods for its robustification. Finally, we illustrate the usage of robust classification on both numerical simulations and real data. We also investigate the influence of the principal component analysis on classification results.

Keywords: Multivariate statistics, linear classification, quadratic classification, robust estimations, principal component analysis

Obsah

Úvod

Mnohorozměrná statistická ananlýza nabízí celou řadu různých metod pro analýzu mnohorozměrných dat. Tyto metody jsou však typicky založeny na následujících předpokladech: všechna pozorování jsou nezávislá, pocházejí z určitého rozdělení a známe parametry, které toto rozdělení určují. Metody jsou navrženy tak, aby při splnění těchto předpokladů dávaly v jistém smyslu optimální výsledky. Řada metod je však citlivá na jejich porušení. Mnohdy stačí jediné pozorování, které se nějak vymyká charakteru zbylého souboru dat, aby metoda vedla k zavádějícím, někdy až nesmyslným výsledkům a závěrům. Takové atypické pozorování se nazývá *odlehlé* a může jím být chyba měření, přepis, vadný výrobek nebo třeba výjimečný jedinec v populaci. Statistické metody, které jsou vůči vlivům odlehlých pozorování imunní, zde budeme nazývat *robustní*. Podotkněme ještě, že pojem *robustnosti* je v různých pracích chápán různě. Například Huber a Ronchetti (2009) definují robustnost obecně jako „necitlivost vůči malým odchylkám od předpokladů".

Jednou z úloh, se kterými se můžeme v mnohorozměrné statistice setkat, je úloha klasifikace, či jinak diskriminace. V takovém případě se nacházíme v situaci, kdy pozorujeme několik veličin na objektech, které jsou přirozeně rozděleny do skupin. Naším cílem je pak sestavit na základě těchto pozorování pravidla, s jejichž pomocí od sebe jednotlivé skupiny dobře rozlišíme. Tato pravidla nám navíc umožňují klasifikovat do skupin nová pozorování, o kterých doposud nevíme, do které skupiny patří. Například jedince v nějaké populaci můžeme rozdělit podle toho, jestli trpí nebo netrpí určitou chorobou. U každého jedince pak můžeme změřit několik veličin, u kterých předpokládáme, že jsou chorobou ovlivněny. Na základě změřených hodnot se pak pokusíme sestavit pravidlo, které odliší zdravého jedince od nemocného. Pokud dává dobré výsledky, můžeme toto pravidlo použít pro diagnostikování nemoci u nějakého nového jedince.

Poznamemenjme, že v této práci budeme pojmy klasifikace a diskriminace chápat jako synonyma. Někteří autoři však tyto pojmy rozlišují. Diskriminaci pak chápou spíše jako první část celého procesu. Tedy jako úlohu nalézt funkci, která nějakým způsobem kvantifikuje rozdíly mezi skupinami a kterou je také možné interpretovat. Kalsifikací pak myslí proces zařazování objektů do skupin na základě výsledků, které tato funkce dává.

V kapitole 1 některé metody klasifikační analýzy popíšeme. Uvedeme i postup, jak vyčíslit schopnost pravidla správně klasifikovat pozorování do příslušné skupiny. Ukazuje se však, že tyto metody klasifikace nejsou ve své základní podobě robustní. V kapitole 3 popíšeme, jak tento nedostatek odstranit. K tomu však budeme potřebovat robustní odhad střední hodnoty a varianční matice. Proto nejprve v kapitole 2 popíšeme, jaké vlastnosti se od robustních odhadů obvykle vyžadují. Následně si několik základních odhadů představíme v kapitole 2.1.

Popíšeme některé jejich vlastnosti a také algoritmy, díky kterým je můžeme v praxi spočítat.

Pokud na objektech ve skupinách pozorujeme velké množství veličin, může být sestavování klasifikačních pravidel obtížná úloha. Před samotnou klasifikací proto může být žádoucí snížit dimenzionalitu dat. Oblíbenou technikou, jak toho docílit, je analýza hlavních komponent. Stručně ji popíšeme v kapitole 4.1. Její základní myšlenka spočívá v projekci dat do několika málo směrů, ve kterých se projevuje největší variabilita dat. Tato metoda však také není robustní a byly proto navrženy postupy, jak její robustifikaci realizovat. Ty popíšeme v kapitole 4.2 společně s příslušnými algoritmy pro jejich výpočet.

V kapitole 5 porovnáme na základě simulací chování nerobustních a několika robustních verzí klasifikačních pravidel pro několik modelů. Zaměříme se také na to, jakým způsobem klasifikaci ovlivní předchozí provedení analýzy hlavních komponent, ať už robustní nebo nerobustní. V kapitole 6 pak obdobné srovnání provedeme na reálných datech.

Kapitola 1

Klasifikační analýza

V této kapitole popíšeme klasická klasifikační pravidla. Nejprve je budeme uvažovat v teoretické podobě a následně popíšeme i jejich výběrové protějšky, které jsou známy jako kvadratická, resp. lineární diskriminační pravidla. Budeme přitom vycházet z knihy Rencher (1998), kde je toto téma zpracováno podrobněji a jsou zde uvedeny i potřebné důkazy. Tématu se částečně věnovala i práce Rensová (2008).

Budeme předpokládat, že máme $k, k \geq 2$ disjunktních skupin objektů, které označíme $\pi^1, \pi^2, \ldots, \pi^k$. Horní index i bude vždy značit příslušnost k i-té skupině. Toto značení je sice neobvyklé, ale usnadní nám orientaci v dalším textu. Na každém objektu pozorujeme náhodný vektor $\boldsymbol{X}_p = (X_1, X_2, \ldots, X_p)'$. Pravidla budeme sestavovat na základě hodnot $\boldsymbol{x} \in \mathbb{R}^p$ pozorovaných na objektech, tj. na konkrétních realizacích $\boldsymbol{X}_p$. Dále budeme předpokládat, že rozdělení $\boldsymbol{X}_p$ je v každé skupině $\pi^i, i = 1, 2, \ldots, k$ absolutně spojité s hustotou $f^i(\boldsymbol{x})$. Navíc budeme uvažovat apriorní pravděpodobnosti $p^i, i = 1, 2, \ldots, k$, že pozorování $\boldsymbol{x}$ pochází ze skupiny π^i. Základní klasifikační pravidlo, při kterém je minimalizována pravděpodobnost špatné klasifikace, má následující tvar.

Pravidlo 1 *Pozorování $\boldsymbol{x}$ zařadíme do skupiny π^i, jestliže*

$$p^i f^i(\boldsymbol{x}) > p^j f^j(\boldsymbol{x}) \text{ pro } j = 1, 2, \ldots, k, j \neq i \tag{1.1}$$

tj. jestliže $p^i f^i(\boldsymbol{x}) = \max_j p^j f^j(\boldsymbol{x})$.

Uvedeme zde podobu pravidla 1 pro důležitý speciální případ, kdy je rozdělení ve skupinách $\pi^i, i = 1, 2, \ldots, k$, p-rozměrné normální s parametry $(\boldsymbol{\mu}^i, \boldsymbol{\Sigma}^i)$. Získáme tak *kvadratické klasifikační pravidlo*. Poznamenejme, že $|\boldsymbol{A}|$ bude nadále značit determinant čtvercové matice $\boldsymbol{A}$.

Pravidlo 2 *Pozorování $\boldsymbol{x}$ zařadíme do skupiny π^i, jestliže*

$$d_Q^i(\boldsymbol{x}) > d_Q^j(\boldsymbol{x}) \text{ pro } j = 1, 2, \ldots, k, j \neq i \tag{1.2}$$

kde $d_Q^i(\boldsymbol{x}) = -\frac{1}{2} \ln |\boldsymbol{\Sigma}^i| - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}^i)'(\boldsymbol{\Sigma}^i)^{-1}(\boldsymbol{x} - \boldsymbol{\mu}^i) + \ln p^i$.

Nerovnost (1.2) získáme lehce dosazením hustoty normálního rozdělení do (1.1) s využitím faktu, že úloha $\max_j p^j f^j(\boldsymbol{x})$ je, co do řešení, ekvivalentní s úlohou $\max_j \ln (p^j f^j(\boldsymbol{x}))$.

Pokud jsou navíc varianční matice ve všech skupinách shodné, $\boldsymbol{\Sigma}^1 = \boldsymbol{\Sigma}^2 = \ldots = \boldsymbol{\Sigma}^k = \boldsymbol{\Sigma}$, pak se pravidlo 1 pro normální rozdělení zjednoduší na *lineární klasifikační pravidlo.*

Pravidlo 3 *Pozorování* $\boldsymbol{x}$ *zařadíme do skupiny* π^i*, jestliže*

$$d_L^i(\boldsymbol{x}) > d_L^j(\boldsymbol{x}) \text{ pro } j = 1, 2, \ldots, k, j \neq i \tag{1.3}$$

kde $d_L^i(\boldsymbol{x}) = \boldsymbol{\mu}^{i\prime}\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}^{i\prime}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}^i + \ln p^i$.

Skutečné hodnoty $\boldsymbol{\mu}^i$, $\boldsymbol{\Sigma}^i$, případně $\boldsymbol{\Sigma}$, typicky neznáme. Běžnou praxí je nahradit je v (1.2) a (1.3) jejich výběrovými protějšky. Mějme pro každé $i, i = 1, 2 \ldots, k$ k dispozici výběr z $\mathsf{N}(\boldsymbol{\mu}^i, \boldsymbol{\Sigma}^i)$, případně z $\mathsf{N}(\boldsymbol{\mu}^i, \boldsymbol{\Sigma})$, o rozsahu n^i, $\boldsymbol{X}^i = (\boldsymbol{x}_1^i, \boldsymbol{x}_2^i, \ldots, \boldsymbol{x}_{n^i}^i)$. Zdůrazněme, že matice $\boldsymbol{X}^i$ je typu $p \times n^i$. Jako odhad střední hodnoty $\boldsymbol{\mu}^i$ vezmeme průměr $\overline{\boldsymbol{x}}^i = \sum_{j=1}^{n^i} \boldsymbol{x}_j^i / n^i$ a jako odhad $\boldsymbol{\Sigma}^i$ použijeme výběrovou varianční matici $\boldsymbol{S}^i = \sum_{j=1}^{n^i} (\boldsymbol{x}_j^i - \overline{\boldsymbol{x}}^i)(\boldsymbol{x}_j^i - \overline{\boldsymbol{x}}^i)'/(n^i - 1)$. V případě shodných variančních matic použijeme jako odhad $\boldsymbol{\Sigma}$ sdruženou výběrovou varianční matici

$$\boldsymbol{S} = \frac{1}{(N-k)} \sum_{i=1}^{k} (n^i - 1)\boldsymbol{S}^i, \tag{1.4}$$

kde $N = \sum_{i=1}^{k} n^i$. Dostaneme tak *výběrové kvadratické* a *výběrové lineární klasifikační pravidlo.*

Pravidlo 4 *Pozorování* $\boldsymbol{x}$ *zařadíme do skupiny* π^i*, jestliže*

$$d_{QV}^i(\boldsymbol{x}) > d_{QV}^j(\boldsymbol{x}) \text{ pro } j = 1, 2, \ldots, k, j \neq i \tag{1.5}$$

kde $d_{QV}^i(\boldsymbol{x}) = -\frac{1}{2}\ln|\boldsymbol{S}^i| - \frac{1}{2}(\boldsymbol{x} - \overline{\boldsymbol{x}}^i)'(\boldsymbol{S}^i)^{-1}(\boldsymbol{x} - \overline{\boldsymbol{x}}^i) + \ln p^i$.

Pravidlo 5 *Pozorování* $\boldsymbol{x}$ *zařadíme do skupiny* π^i*, jestliže*

$$d_{LV}^i(\boldsymbol{x}) > d_{LV}^j(\boldsymbol{x}) \text{ pro } j = 1, 2, \ldots, k, j \neq i, \tag{1.6}$$

kde $d_{LV}^i(\boldsymbol{x}) = \overline{\boldsymbol{x}}^{i\prime}\boldsymbol{S}^{-1}\boldsymbol{x} - \frac{1}{2}\overline{\boldsymbol{x}}^{i\prime}\boldsymbol{S}^{-1}\overline{\boldsymbol{x}}^i + \ln p^i$.

Pokud budeme používat výběrové kvadratické klasifikační pravidlo 4, budeme ho v textu dále značit QDA. Obdobně výběrové lineární klasifikační pravidlo 5 budeme značit LDA.

Dalšími parametry, které obvykle neznáme a musíme je odhadnout, jsou apriorní pravděpodobnosti $p^i, i = 1, 2, \ldots, k$. Pokud víme, že rozsahy výběrů n^i odpovídají reálnému zastoupení objektů ve skupinách, pak místo p^i použijeme v pravidlech relativní četnosti $\hat{p}^i = n^i/N$. Pokud tomu tak není, dosadíme za p^i konstantu $\hat{p}^i = 1/k$. V takovém případě nemá člen $\ln \hat{p}^i$ na klasifikaci vliv, neboť je pro všechny skupiny stejný, a můžeme ho tedy vynechat.

Kvalitu jednotlivých klasifikačních pravidel mezi sebou můžeme porovnávat na základě jejich pravděpodobnosti špatné klasifikace. Tuto pravděpodobnost budeme dále nazývat stručně *chyba klasifikace* a značit ji ER (z anglického *error rate*). Předpokládejme, že již máme sestaveno konkrétní klasifikační pravidlo. Bude nás zajímat chyba klasifikace nového, dosud nezařazeného, objektu na

základě tohoto pravidla. Tuto chybu klasifikace budeme značit AER (z anglického termínu *actual error rate*) a definujeme ji vztahem

$$\mathrm{AER} = \sum_{i=1}^{k} p^i P(\pi^1, \ldots, \pi^{i-1}, \pi^{i+1}, \ldots, \pi^k | \pi^i),$$

kde $P(\pi^1, \ldots, \pi^{i-1}, \pi^{i+1}, \ldots, \pi^k | \pi^i)$ značí podmíněnou pravděpodobnost jevu, že objekt bude zařazen do některé ze skupin $\pi^j, j = 1, \ldots, i-1, i+1, \ldots, k$ patří-li ve skutečnosti do skupiny π^i. Pokud v konkrétní situaci máme k dispozici více klasifikačních pravidel, použijeme to, jehož chyba klasifikace AER je nejmenší.

V praxi se tato veličina opět odhaduje. Nejprve popíšeme odhad, který nazveme *přímý odhad chyby klasifikace* a budeme ho značit ApER (z anglického termínu *apparent error rate*). Ten je založený na jednoduchém principu. Klasifikační pravidlo sestavíme na základě všech N pozorování ze všch skupin. Toto pravidlo posléze aplikujeme opět na všechna pozorování a ApER definujeme jako podíl špatně zařazených pozorování mezi všemi pozorováními, tj.

$$\mathrm{ApER} = \frac{\sum_{i=1}^{k} m^i}{N}, \tag{1.7}$$

kde m^i je počet špatně klasifikovaných pozorování z i-té skupiny. Protože však klasifikujeme stejná pozorování, na základě kterých jsme dané pravidlo sestavili, dává ApER příliš optimistické výsledky.

Realističtější odhad chyby klasifikace získáme pomocí *křížové validace* (nebo také *krosvalidace*). Tento odhad budeme označovat zkratkou CV (z anglického *cross-validation*). Nejpoužívanější verzí křížové validace je metoda *leave-one-out*, která je založena na vynechání právě jednoho vzorku z datového souboru. V tomto případě pro každé pozorování $\boldsymbol{x}_j^i, i = 1, 2, \ldots, k, j = 1, 2, \ldots, n^i$, sestavíme klasifikační pravidlo na základě zbylých $N-1$ pozorování a pozorování $\boldsymbol{x}_j^i$ podle tohoto pravidla klasifikujeme. Odhad CV je pak opět podíl špatně zařazených mezi všemi pozorováními, tj.

$$\mathrm{CV} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n^i} \delta_j^i}{N}, \tag{1.8}$$

kde δ_j^i je 1, pokud bylo pozorování $\boldsymbol{x}_j^i$ špatně klasifikováno, nebo 0 v opačném případě. Tento odhad je však výpočetně náročný, a proto se používá především pro malé soubory dat.

Dalším častým postupem, jak odhadnout AER, je rozdělit výběr náhodně na dvě části - tréninkovou a testovací. Na základě tréninkové části sestavíme pravidlo, které následně aplikujeme na pozorování v testovací části. Označme pro $i = 1, 2, \ldots, k$ pozorování ze skupiny π^i vybraná do tréninkové části jako podvýběr $\boldsymbol{X}_{tr}^i$ a jejich počet n_{tr}^i. Obdobně pozorování zařazená do testovací části označíme $\boldsymbol{X}_{ts}^i$ a jejich počet n_{ts}^i. Pak máme tréninkový výběr $\boldsymbol{X}_{tr} = (\boldsymbol{X}_{tr}^1 \boldsymbol{X}_{tr}^2, \ldots, \boldsymbol{X}_{tr}^k)$ s $N_{tr} = \sum_{i=1}^{k} n_{tr}^i$ pozorováními a testovací výběr $\boldsymbol{X}_{ts} = (\boldsymbol{X}_{ts}^1 \boldsymbol{X}_{ts}^2, \ldots, \boldsymbol{X}_{ts}^k)$ s $N_{ts} = \sum_{i=1}^{k} n_{ts}^i$ pozorováními. AER pak odhadneme jako podíl špatně zařazených pozorování z testovací části. Označme tento odhad AER_{ts}. Máme tedy

$$AER_{ts} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_{ts}^i} \delta_j^i}{N_{ts}}, \tag{1.9}$$

kde δ_j^i je opět rovno 1, pokud bylo pozorování $\boldsymbol{x}_j^i$ špatně klasifikováno, a 0 jinak. AER_{ts} se používá zejména v případech, kdy zkoumáme rozsáhlý soubor dat.

Kapitola 2

Mnohorozměrné odhady

Klasickou úlohou mnohorozměrné statistiky je odhad střední hodnoty a varianční matice, které nám dávají informaci o poloze dat v prostoru a jejich rozptýlenosti. Tyto odhady jsou pak často používány jako vstupní parametry v mnoha statistických metodách, jako například LDA a QDA. Proto na ně klademe celou řadu požadavků. Běžně vyžadujeme, aby byl odhad varianční matice pozitivně definitní matice, nebo aby odhady komutovaly s lineární transformací aplikovanou na data. V případě, že data obsahují odlehlá pozorování, bude mezi tyto požadavky patřit i robustnost odhadů.

V průběhu let byla navržena celá řada robustních odhadů a metod. Výlet do historie tohoto hledání nabízí Stigler (2010). Nejprve byla věnována pozoronost odhadům parametru posunutí a lineárnímu regresnímu modelu. Menší pozornost se doposud věnovala odhadům střední hodnoty a varianční matice.

Běžnou praxí je odhadnout střední hodnotu pomocí výběrového průměru a jako odhad varianční matice použít výběrovou varianční matici. Tyto odhady jsou však velmi citlivé na odlehlá pozorování, a nejsou tedy robustní. Tento fakt můžeme ilustrovat na jednoduchém jednorozměrném případě. Uvažujme hodnoty $1, 2, \ldots, 10$. Pak je jejich výběrový průměr 5,5 a výběrový rozptyl 9,2. Pokud však místo hodnoty 10 vezmeme 100, změní se hodnoty průměru a rozptylu na 14,5 a 909,2. Tak například omylem posunutá desetinná čárka může způsobit, že odhadnutý střed dat leží zcela mimo tato data. Navíc zásadně ovlivní i odhad rozptylu. Tento krátký příklad může sloužit jako motivace pro hledání robustní alternativy pro výběrový průměr a výběrovou varianční matici. Několik takových alternativ popíšeme v kapitole 2.1. Nejprve se však zaměříme na značení a popíšeme některé vlastnosti, které budeme od odhadů vyžadovat. Budeme přitom často vycházet z kapitoly 6 knihy Maronna, Martin a Yohai (2006), případně z knihy Jurečková (2001).

Označme $\mathcal{P}_p$ množinu všech rozdělení na $\mathbb{R}^p$ a $\mathrm{SPD}(p)$ množinu všech symetrických pozitivně definitních matic typu $p \times p$. Jednotkovou matici budeme značit $\boldsymbol{I}$, n-rozměrný vektor skládající se ze samých jedniček označíme $\mathbf{1}_n$ a euklidovskou normu vektoru $\boldsymbol{x} \in \mathbb{R}^p$ označíme $\|\boldsymbol{x}\|$.

Uvažujme náhodný výběr $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ z p-rozměrného rozdělení $P \in \mathcal{P}_p$ se střední hodnotou $\boldsymbol{\mu}$ a varianční maticí $\boldsymbol{\Sigma}$. Toto rozdělení budeme dále značit $P_{\boldsymbol{\mu},\boldsymbol{\Sigma}}$. Symbol $\boldsymbol{X}$ budeme chápat jednak jako výběr, jednak jako matici typu $p \times n$. Označme $\boldsymbol{T}(\boldsymbol{X}) \in \mathbb{R}^p$ odhad střední hodnoty $\boldsymbol{\mu}$ a $\boldsymbol{C}(\boldsymbol{X}) \in \mathrm{SPD}(p)$ odhad varianční matice $\boldsymbol{\Sigma}$. Vzhledem k tomu, že v dalším textu budeme často

odhadovat $\boldsymbol{T}(\boldsymbol{X})$ a $\boldsymbol{C}(\boldsymbol{X})$ společně, budeme je v takových situacích uvažovat jako jediný simultánní odhad $(\boldsymbol{T}(\boldsymbol{X}), \boldsymbol{C}(\boldsymbol{X}))$. Pro jednoduchost budeme někdy, pokud toto značení nebude zavádějící, místo $\boldsymbol{T}(\boldsymbol{X})$ a $\boldsymbol{C}(\boldsymbol{X})$ psát pouze $\boldsymbol{T}$ a $\boldsymbol{C}$. Pokud budeme chtít zdůraznit, že odhad je založený na výběru o rozsahu n, vyznačíme tuto skutečnost indexem n u příslušného odhadu.

Některé žádoucí vlastnosti odhadů jsou často zaručeny pouze za předpokladu, že jsou body výběru $\boldsymbol{X}$ „dobře rozptýleny“. Co tímto požadavkem myslíme, objasňuje následující definice.

Definice 2.1 *Řekneme, že výběr $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ je v* obecné pozici*, jestliže v každém $(p-1)$-rozměrném podprostoru $\mathbb{R}^p$ neleží více než p bodů výběru.*

Informaci o globálních robustních vlastnostech odhadu můžeme získat pomocí tzv. *bodu selhání*. Ten popisuje citlivost odhadu na odlehlá pozorování.

Definice 2.2 *Nechť $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ je náhodný výběr z $P_{\boldsymbol{\mu},\boldsymbol{\Sigma}} \in \mathcal{P}_p$. Nechť $\boldsymbol{X}^*$ je náhodný výběr, který vznikne z $\boldsymbol{X}$ nahrazením m náhodně vybraných bodů z $\boldsymbol{X}$ libovolnými hodnotami z $\mathbb{R}^p$. Potom pro odhad $\boldsymbol{T}$ střední hodnoty $\boldsymbol{\mu}$ definujeme*

- bod selhání odhadu $\boldsymbol{T}$ při výběru $\boldsymbol{X}$

$$\epsilon_n^*(\boldsymbol{T}; \boldsymbol{X}) = \min\{\frac{m}{n}, \beta(m; \boldsymbol{T}, \boldsymbol{X}) = \infty\}, \tag{2.1}$$

  kde $\beta(m; \boldsymbol{T}, \boldsymbol{X}) = \sup_{\boldsymbol{X}^*} \|\boldsymbol{T}(\boldsymbol{X}^*) - \boldsymbol{T}(\boldsymbol{X})\|$,

- bod selhání odhadu $\boldsymbol{T}$

$$\epsilon^*(\boldsymbol{T}) = \lim_{n\to\infty} \epsilon_n^*(\boldsymbol{T}; \boldsymbol{X}).$$

Pro odhad $\boldsymbol{C} \in \mathrm{SPD}(p)$ varianční matice $\boldsymbol{\Sigma}$ definujeme

- bod selhání odhadu $\boldsymbol{C}$ při výběru $\boldsymbol{X}$

$$\epsilon_n^*(\boldsymbol{C}; \boldsymbol{X}) = \min\{\frac{m}{n}, \sup_{\boldsymbol{X}^*} D(\boldsymbol{C}(\boldsymbol{X}), \boldsymbol{C}(\boldsymbol{X}^*)) = \infty\},$$

  kde pro $\boldsymbol{A}, \boldsymbol{B} \in \mathrm{SPD}(p)$ je

$$D(\boldsymbol{A}, \boldsymbol{B}) = \max\{|\lambda_1(\boldsymbol{A}) - \lambda_1(\boldsymbol{B})|, |\lambda_p(\boldsymbol{A})^{-1} - \lambda_p(\boldsymbol{B})^{-1}|\},$$

  $\lambda_1(\boldsymbol{A}) \geq \ldots \geq \lambda_p(\boldsymbol{A})$ jsou vlastní čísla matice $\boldsymbol{A}$ a $\lambda_1(\boldsymbol{B}) \geq \ldots \geq \lambda_p(\boldsymbol{B})$ jsou vlastní čísla matice $\boldsymbol{B}$,

- bod selhání odhadu $\boldsymbol{C}$

$$\epsilon^*(\boldsymbol{C}) = \lim_{n\to\infty} \epsilon_n^*(\boldsymbol{C}; \boldsymbol{X}).$$

Pro simultánní odhad $(\boldsymbol{T}, \boldsymbol{C})$ zavedeme bod selhání $(\boldsymbol{T}, \boldsymbol{C})$ při výběru $\boldsymbol{X}$ *jako*

$$\epsilon_n^*(\boldsymbol{T}, \boldsymbol{C}; \boldsymbol{X}) = \min\{\epsilon_n^*(\boldsymbol{T}; \boldsymbol{X}), \epsilon_n^*(\boldsymbol{C}; \boldsymbol{X})\}$$

a bod selhání $(\boldsymbol{T}, \boldsymbol{C})$ *jako*

$$\epsilon^*(\boldsymbol{T}, \boldsymbol{C}) = \min\{\epsilon^*(\boldsymbol{T}), \epsilon^*(\boldsymbol{C})\}.$$

Bod selhání odhadu nám vlastně říká, jakým nejmenším zlomkem musíme původní data „kontaminovat“, aby odhad „selhal“. Pro odhad střední hodnoty $\boldsymbol{T}$ to znamená, že se jeho nová hodnota dostane libovolně daleko od jeho původní hodnoty pro nekontaminovaná data. Odhad varianční matice $\boldsymbol{C}$ selže, pokud. „exploduje“ nebo „imploduje“, tj. pokud hodnoty v matici budou nabývat libovolně velkých hodnot (a s nimi i největší vlastní číslo matice), nebo pokud bude matice $\boldsymbol{C}$ singulární (a nejmenší vlastní číslo bude nulové). K tomu, aby selhal simultánní odhad $(\boldsymbol{T}, \boldsymbol{C})$, stačí, aby selhal jeden z odhadů $\boldsymbol{T}$, $\boldsymbol{C}$. Čím vyšší je bod selhání odhadu, tím je odhad odolnější vůči odlehlým pozorováním, a je tedy robustnější. Z tohoto hlediska lze za nejlepší považovat odhady, které dosahují maximálního bodu selhání.

K popisu lokálních robustních vlastnstí se využívá koncept tzv. *influenční funkce* (viz např. Hampel, 1974). K jejímu vyjádření potřebujeme sestrojit statistický funkcionál odpovídající našemu odhadu. Předpokládejme tedy, že náhodný výběr $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ pochází z rozdělení $P_{\boldsymbol{\theta}} \in \mathcal{P}_p$, kde $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ je odhadovaný parametr. Pro odhad střední hodnoty bude $\boldsymbol{\Theta} = \mathbb{R}^p$ a pro odhad varianční matice $\boldsymbol{\Theta} = \mathrm{SPD}(p)$, což lze chápat jako podmnožinu $\mathbb{R}^{p(p+1)/2}$. Označme P_n empirické rozdělení náhodného výběru $\boldsymbol{X}$. Funkcionál $\boldsymbol{T} : \mathcal{P}_p \to \boldsymbol{\Theta}$ hledáme v takovém tvaru, aby platilo $\boldsymbol{T}(\boldsymbol{X}) = \boldsymbol{T}(P_n)$. Například pro výběrový průměr jako odhad střední hodnoty $\boldsymbol{\mu}$ je odpovídajícím funkcionálem $T(P_{\boldsymbol{\mu}}) = \mathsf{E}_{P_{\boldsymbol{\mu}}}(\boldsymbol{X})$.

Definice 2.3 *Mějme náhodné rozdělení $P \in \mathcal{P}_p$. Označme $\Delta_{\boldsymbol{x}}$ degenerované rozdělení soustředěné v bodě $\boldsymbol{x}$. Pro $\boldsymbol{x} \in \mathbb{R}^p$ definujeme* influenční funkci odhadu $\boldsymbol{T}$ v P

$$IF(\boldsymbol{x}; \boldsymbol{T}, P) = \lim_{\epsilon \searrow 0} \frac{T((1-\epsilon)P + \epsilon\Delta_{\boldsymbol{x}})}{\epsilon},$$

pokud limita na pravé straně existuje.

Influenční funkce měří limitní vliv, jaký má na odhad malá kontaminace umístěná do jediného bodu $\boldsymbol{x}$. Je tedy žádoucí, aby influenční funkce námi zvoleného odhadu byla omezená. Od influenční funkce odvozujeme další robustní charakteristiky. V tomto případě se budeme držet terminologie, kterou používá Jurečková (2001). Nejhorší možný vliv kontaminace na odhad měří *globální citlivost* (v anglické literatuře značena jako GES z *gross-error sensitivity*)

$$\gamma^*(\boldsymbol{T}, P) = \sup_{\boldsymbol{x}} \|IF(\boldsymbol{x}; \boldsymbol{T}, P)\|,$$

kde se supremum hledá přes všechna $\boldsymbol{x}$, ve kterých je influenční funkce definovaná. Obdobně lze zavést *lokální citlivost* (často značená LSS z *local shift sensitivity*)

$$\lambda^*(\boldsymbol{T}, P) = \sup_{\boldsymbol{x} \neq \boldsymbol{y}} \frac{\|IF(\boldsymbol{x}; \boldsymbol{T}, P) - IF(\boldsymbol{y}; \boldsymbol{T}, P)\|}{\|\boldsymbol{x} - \boldsymbol{y}\|},$$

která měří, jaký vliv má na odhad posun pozorování $\boldsymbol{x}$ do bodu $\boldsymbol{y}$. Tematika influenčních funkcí je důležitou součástí robustní statistiky a je jí věnována například kniha Hampel a kol. (1986).

V situaci, kdy už máme pro soubor dat $\boldsymbol{X}$ sestrojené odhady $\boldsymbol{T}(\boldsymbol{X})$ a $\boldsymbol{C}(\boldsymbol{X})$, můžeme data v prostoru posunout ve směru $\boldsymbol{b} \in \mathbb{R}^p$. V takovém případě intuitivně očekáváme, že odhad střední hodnoty se také posune ve směru $\boldsymbol{b}$ (bude

ekvivariantní vůči posunutí) a odhad varianční matice se nezmění (bude *invariantní* vůči posunutí). Jakkoliv je tento požadavek přirozený, není jeho splnění samozřejmostí. Obdobně je tomu i s ekvivariancí odhadů vůči rotaci nebo změně měřítka, kdy očekáváme, že se odhad $\boldsymbol{C}$ odpovídajícím způsobem „deformuje“. Tyto vlastnosti odhadů nyní definujeme přesně.

Definice 2.4 *Řekneme, že odhad* $\boldsymbol{T}(\boldsymbol{X})$ *je*

- translačně ekvivariantní, *jestliže*

$$\boldsymbol{T}(\boldsymbol{X} + \boldsymbol{b}\mathbf{1}_n') = \boldsymbol{T}(\boldsymbol{X}) + \boldsymbol{b},$$

  pro libovolný vektor $\boldsymbol{b} \in \mathbb{R}^p$,

- ortogonálně ekvivariantní, *jestliže*

$$\boldsymbol{T}(\boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}\mathbf{1}_n') = \boldsymbol{A}\boldsymbol{T}(\boldsymbol{X}) + \boldsymbol{b},$$

  pro libovolný vektor $\boldsymbol{b} \in \mathbb{R}^p$ *a libovolnou ortogonální matici* $\boldsymbol{A}$ *typu* $p \times p$,

- afinně ekvivariantní, *jestliže*

$$\boldsymbol{T}(\boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}\mathbf{1}_n') = \boldsymbol{A}\boldsymbol{T}(\boldsymbol{X}) + \boldsymbol{b},$$

  pro libovolný vektor $\boldsymbol{b} \in \mathbb{R}^p$ *a libovolnou regulární matici* $\boldsymbol{A}$ *typu* $p \times p$.

Řekneme, že odhad $\boldsymbol{C}(\boldsymbol{X})$ *je*

- translačně invariantní, *jestliže*

$$\boldsymbol{C}(\boldsymbol{X} + \boldsymbol{b}\mathbf{1}_n') = \boldsymbol{C}(\boldsymbol{X}),$$

  pro libovolný vektor $\boldsymbol{b} \in \mathbb{R}^p$,

- ortogonálně ekvivariantní, *jestliže*

$$\boldsymbol{C}(\boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}\mathbf{1}_n') = \boldsymbol{A}\boldsymbol{C}(\boldsymbol{X})\boldsymbol{A}',$$

  pro libovolný vektor $\boldsymbol{b} \in \mathbb{R}^p$ *a libovolnou ortogonální matici* $\boldsymbol{A}$ *typu* $p \times p$,

- afinně ekvivariantní, *jestliže*

$$\boldsymbol{C}(\boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}\mathbf{1}_n') = \boldsymbol{A}\boldsymbol{C}(\boldsymbol{X})\boldsymbol{A}',$$

  pro libovolný vektor $\boldsymbol{b} \in \mathbb{R}^p$ *a libovolnou regulární matici* $\boldsymbol{A}$ *typu* $p \times p$.

Řekneme, že odhad $(\boldsymbol{T}(\boldsymbol{X}), \boldsymbol{C}(\boldsymbol{X}))$ *je*

- translačně ekvivariantní, *jestliže je odhad* $\boldsymbol{T}(\boldsymbol{X})$ *translačně ekvivariantní a odhad* $\boldsymbol{C}(\boldsymbol{X})$ *translačně invariantní,*

- ortogonálně ekvivariantní, *jestliže jsou oba odhady* $\boldsymbol{T}(\boldsymbol{X})$ *a* $\boldsymbol{C}(\boldsymbol{X})$ *ortogonálně ekvivariantní,*

- afinně ekvivariantní, *jestliže jsou oba odhady* $\boldsymbol{T}(\boldsymbol{X})$ *a* $\boldsymbol{C}(\boldsymbol{X})$ *afinně ekvivariantní.*

Afinně ekvivariantním odhadům je věnována značná pozornost. Davies (1987) určil horní hranici pro bod selhání afinně ekvivariantního odhadu $\boldsymbol{C}(\boldsymbol{X})$. Tuto horní hranici tedy nepřekročí ani bod selhání pro společný odhad $(\boldsymbol{T}(\boldsymbol{X}), \boldsymbol{C}(\boldsymbol{X}))$.

Věta 1 *Nechť náhodný výběr* $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ *pochází z rozdělení* $P_{\boldsymbol{\mu},\boldsymbol{\Sigma}}$, $n \geq p+1$ *a výběr* $\boldsymbol{X}$ *je v obecné pozici. Předpokládejme, že* $\boldsymbol{C}(\boldsymbol{X})$ *je afinně ekvivariantní odhad* $\boldsymbol{\Sigma}$. *Pak pro jeho bod selhání platí*

$$\epsilon_n^*(\boldsymbol{C}; \boldsymbol{X}) \leq \frac{\lfloor (n-p+1)/2 \rfloor}{n}, \tag{2.2}$$

kde $\lfloor k \rfloor$ *značí dolní celou část* k, *tj. největší* $z \in \mathbb{Z}$ *takové, že* $k \geq z$.

Důkaz. Viz Davies (1987). □

Bodem selhání afinně ekvivariantních odhadů se pak detailně zabývali Lopuhaä a Rousseeuw (1991). Mimo jiné odvodili, že bod selhání afinně ekvivariantních odhadů je invariantní vůči afinní transformaci.

Věta 2 *Nechť náhodný výběr* $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ *pochází z rozdělení* $P_{\boldsymbol{\mu},\boldsymbol{\Sigma}}$ *a* $(\boldsymbol{T}(\boldsymbol{X}), \boldsymbol{C}(\boldsymbol{X}))$ *je afinně ekvivariantní odhad* $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. *Pak platí*

$$\epsilon_n^*(\boldsymbol{T}; \boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}\boldsymbol{1}_n') = \epsilon_n^*(\boldsymbol{T}; \boldsymbol{X}), \tag{2.3}$$
$$\epsilon_n^*(\boldsymbol{C}; \boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}\boldsymbol{1}_n') = \epsilon_n^*(\boldsymbol{C}; \boldsymbol{X}), \tag{2.4}$$

pro libovolný vektor $\boldsymbol{b} \in \mathbb{R}^p$ *a libovolnou regulární matici* $\boldsymbol{A}$ *typu* $p \times p$.

K důkazu věty 2 budeme potřebovat následující lemma.

Lemma 1 *Nechť* $\boldsymbol{A}$ *je symetrická matice typu* $p \times p$ *a označme její vlastní čísla* $\lambda_1(\boldsymbol{A}) \geq \lambda_2(\boldsymbol{A}) \geq \cdots \geq \lambda_p(\boldsymbol{A})$. *Pak platí*

$$\lambda_1(\boldsymbol{A}) = \sup_{\boldsymbol{x} \neq \boldsymbol{0}} \frac{\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}'\boldsymbol{x}}, \quad \lambda_p(\boldsymbol{A}) = \inf_{\boldsymbol{x} \neq \boldsymbol{0}} \frac{\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}'\boldsymbol{x}}.$$

Důkaz. Viz Rao (1978, kap. 1f.2). □

Důkaz věty 2. Označme $\boldsymbol{X}^*$ výběr, který získáme nahrazením m náhodných bodů z $\boldsymbol{X}$ libovolnými hodnotami z $\mathbb{R}^p$. Pro $\boldsymbol{b} \in \mathbb{R}^p$ a regulární matici $\boldsymbol{A}$ typu $p \times p$ označme $\boldsymbol{Z} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}\boldsymbol{1}_n'$ a $\boldsymbol{Z}^* = \boldsymbol{A}\boldsymbol{X}^* + \boldsymbol{b}\boldsymbol{1}_n'$.

Nejprve dokážeme (2.3). Z afinní ekvivariance odhadu $\boldsymbol{T}$ dostáváme

$$\|\boldsymbol{T}(\boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}\boldsymbol{1}_n') - \boldsymbol{T}(\boldsymbol{A}\boldsymbol{X}^* + \boldsymbol{b}\boldsymbol{1}_n')\| = \|\boldsymbol{A}(\boldsymbol{T}(\boldsymbol{X}) - \boldsymbol{T}(\boldsymbol{X}^*))\|. \tag{2.5}$$

S využitím lemmatu 1 dostáváme pro matici $\boldsymbol{A}'\boldsymbol{A}$

$$\lambda_p(\boldsymbol{A}'\boldsymbol{A}) \leq \frac{\boldsymbol{x}'\boldsymbol{A}'\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}'\boldsymbol{x}} \leq \lambda_1(\boldsymbol{A}'\boldsymbol{A}), \tag{2.6}$$

pro $\boldsymbol{x} \neq \boldsymbol{0}$. Dosadíme $\boldsymbol{T}(\boldsymbol{X}) - \boldsymbol{T}(\boldsymbol{X}^*)$ za $\boldsymbol{x}$ a vzhledem k (2.5) dostáváme

$$\lambda_p(\boldsymbol{A}'\boldsymbol{A}) \leq \frac{\|\boldsymbol{T}(\boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}\boldsymbol{1}_n') - \boldsymbol{T}(\boldsymbol{A}\boldsymbol{X}^* + \boldsymbol{b}\boldsymbol{1}_n')\|^2}{\|\boldsymbol{T}(\boldsymbol{X}) - \boldsymbol{T}(\boldsymbol{X}^*)\|^2} \leq \lambda_1(\boldsymbol{A}'\boldsymbol{A}).$$

Odtud plyne, že $\sup_{\boldsymbol{X}^*} \|\boldsymbol{T}(\boldsymbol{X}) - \boldsymbol{T}(\boldsymbol{X}^*)\|$ je konečné (resp. nekonečné) právě tehdy, když je konečné (resp. nekonečné) $\sup_{\boldsymbol{Z}^*} \|\boldsymbol{T}(\boldsymbol{Z}) - \boldsymbol{T}(\boldsymbol{Z}^*)\|$. Dle definice bodu selhání (2.1) jsme tím dokázali (2.3).

Nyní dokážeme (2.4). S využitím afinní ekvivariance $\boldsymbol{C}$ a lemmatu 1 pro $\boldsymbol{C}(\boldsymbol{Z})$ dostáváme

$$\begin{aligned}\lambda_1(\boldsymbol{C}(\boldsymbol{Z})) &= \lambda_1(\boldsymbol{C}(\boldsymbol{AX}+\boldsymbol{b})) = \lambda_1(\boldsymbol{A}\boldsymbol{C}(\boldsymbol{X})\boldsymbol{A}') = \sup_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\boldsymbol{x}'\boldsymbol{A}\boldsymbol{C}(\boldsymbol{X})\boldsymbol{A}'\boldsymbol{x}}{\boldsymbol{x}'\boldsymbol{x}} \\ &= \sup_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\boldsymbol{x}'\boldsymbol{A}\boldsymbol{C}(\boldsymbol{X})\boldsymbol{A}'\boldsymbol{x}}{\boldsymbol{x}'\boldsymbol{A}\boldsymbol{A}'\boldsymbol{x}} \frac{\boldsymbol{x}'\boldsymbol{A}\boldsymbol{A}'\boldsymbol{x}}{\boldsymbol{x}'\boldsymbol{x}}.\end{aligned}$$

Odtud, z lemmatu 1 pro $\lambda_1(\boldsymbol{C}(\boldsymbol{X}))$ a z analogie (2.6) pro matici $\boldsymbol{A}\boldsymbol{A}'$ plyne

$$\lambda_1(\boldsymbol{C}(\boldsymbol{X}))\lambda_p(\boldsymbol{A}\boldsymbol{A}') \leq \lambda_1(\boldsymbol{C}(\boldsymbol{Z})) \leq \lambda_1(\boldsymbol{C}(\boldsymbol{X}))\lambda_1(\boldsymbol{A}\boldsymbol{A}'). \tag{2.7}$$

Tentýž výsledek platí také pro kontaminovaný výběr, tedy

$$\lambda_1(\boldsymbol{C}(\boldsymbol{X}^*))\lambda_p(\boldsymbol{A}\boldsymbol{A}') \leq \lambda_1(\boldsymbol{C}(\boldsymbol{Z}^*)) \leq \lambda_1(\boldsymbol{C}(\boldsymbol{X}^*))\lambda_1(\boldsymbol{A}\boldsymbol{A}'). \tag{2.8}$$

Z první nerovnosti v (2.7) a z druhé nerovnosti v (2.8) plyne po malé úpravě

$$\lambda_1(\boldsymbol{C}(\boldsymbol{Z})) - \lambda_1(\boldsymbol{C}(\boldsymbol{Z}^*)) \geq \lambda_1(\boldsymbol{A}\boldsymbol{A}')\left(\lambda_1(\boldsymbol{C}(\boldsymbol{X})) - \lambda_1(\boldsymbol{C}(\boldsymbol{X}^*))\right) - \alpha,$$

kde $\alpha = \lambda_1(\boldsymbol{C}(\boldsymbol{X}))\left(\lambda_1(\boldsymbol{A}\boldsymbol{A}') - \lambda_p(\boldsymbol{A}\boldsymbol{A}')\right)$. Analogicky z druhé nerovnosti v (2.7) a z první nerovnosti v (2.8) plyne

$$\lambda_1(\boldsymbol{C}(\boldsymbol{Z})) - \lambda_1(\boldsymbol{C}(\boldsymbol{Z}^*)) \leq \lambda_p(\boldsymbol{A}\boldsymbol{A}')\left(\lambda_1(\boldsymbol{C}(\boldsymbol{X})) - \lambda_1(\boldsymbol{C}(\boldsymbol{X}^*))\right) + \alpha.$$

Odtud vidíme, že $\sup_{\boldsymbol{Z}^*} |\lambda_1(\boldsymbol{C}(\boldsymbol{Z})) - \lambda_1(\boldsymbol{C}(\boldsymbol{Z}^*))|$ je konečné právě tehdy, když je konečné $\sup_{\boldsymbol{X}^*} |\lambda_1(\boldsymbol{C}(\boldsymbol{X})) - \lambda_1(\boldsymbol{C}(\boldsymbol{X}^*))|$.

Analogicky postupujeme i pro $\sup_{\boldsymbol{Z}^*} |1/\lambda_p(\boldsymbol{C}(\boldsymbol{Z})) - 1/\lambda_p(\boldsymbol{C}(\boldsymbol{Z}^*))|$. Nejmenší vlastní číslo $\boldsymbol{C}(\boldsymbol{Z})$ vyjádříme jako

$$\lambda_p(\boldsymbol{C}(\boldsymbol{Z})) = \inf_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\boldsymbol{x}'\boldsymbol{A}\boldsymbol{C}(\boldsymbol{X})\boldsymbol{A}'\boldsymbol{x}}{\boldsymbol{x}'\boldsymbol{A}\boldsymbol{A}'\boldsymbol{x}} \frac{\boldsymbol{x}'\boldsymbol{A}\boldsymbol{A}'\boldsymbol{x}}{\boldsymbol{x}'\boldsymbol{x}}$$

a z lemmatu 1 pro $\lambda_p(\boldsymbol{C}(\boldsymbol{X}))$ a z analogie (2.6) pro matici $\boldsymbol{A}\boldsymbol{A}'$ dostáváme

$$\begin{aligned}\lambda_p(\boldsymbol{C}(\boldsymbol{X}))\lambda_p(\boldsymbol{A}\boldsymbol{A}') \leq \lambda_p(\boldsymbol{C}(\boldsymbol{Z})) \leq \lambda_p(\boldsymbol{C}(\boldsymbol{X}))\lambda_1(\boldsymbol{A}\boldsymbol{A}'), \\ \lambda_p(\boldsymbol{C}(\boldsymbol{X}^*))\lambda_p(\boldsymbol{A}\boldsymbol{A}') \leq \lambda_p(\boldsymbol{C}(\boldsymbol{Z}^*)) \leq \lambda_p(\boldsymbol{C}(\boldsymbol{X}^*))\lambda_1(\boldsymbol{A}\boldsymbol{A}').\end{aligned}$$

Z těchto nerovností pak odvodíme

$$\begin{aligned}\frac{1}{\lambda_p(\boldsymbol{C}(\boldsymbol{Z}))} - \frac{1}{\lambda_p(\boldsymbol{C}(\boldsymbol{Z}^*))} &\geq \frac{1}{\lambda_p(\boldsymbol{A}\boldsymbol{A}')}\left(\frac{1}{\lambda_p(\boldsymbol{C}(\boldsymbol{X}))} - \frac{1}{\lambda_p(\boldsymbol{C}(\boldsymbol{X}^*))}\right) + \beta, \\ \frac{1}{\lambda_p(\boldsymbol{C}(\boldsymbol{Z}))} - \frac{1}{\lambda_p(\boldsymbol{C}(\boldsymbol{Z}^*))} &\leq \frac{1}{\lambda_1(\boldsymbol{A}\boldsymbol{A}')}\left(\frac{1}{\lambda_p(\boldsymbol{C}(\boldsymbol{X}))} - \frac{1}{\lambda_p(\boldsymbol{C}(\boldsymbol{X}^*))}\right) - \beta,\end{aligned}$$

kde $\beta = \left(1/\lambda_1(\boldsymbol{A}\boldsymbol{A}') - 1/\lambda_p(\boldsymbol{A}\boldsymbol{A}')\right)/\lambda_p(\boldsymbol{C}(\boldsymbol{X}))$.

Odtud vidíme, že $\sup_{\boldsymbol{Z}^*} |1/\lambda_p(\boldsymbol{C}(\boldsymbol{Z})) - 1/\lambda_p(\boldsymbol{C}(\boldsymbol{Z}^*))|$ je konečné právě tehdy, když je konečné $\sup_{\boldsymbol{X}^*} |1/\lambda_p(\boldsymbol{C}(\boldsymbol{X})) - 1/\lambda_p(\boldsymbol{C}(\boldsymbol{X}^*))|$.

Tím je dokončen důkaz (2.4), a tedy i celé věty. □

Řada modelů ve statistice vychází z konceptu normálního rozdělení. To se však týká i robustních statistických metod, ktaré jsou typicky navrženy pro data z takového spojitého rozdělení, které v určitém smyslu leží v okolí normálního rozdělení. Nyní zavedeme obecnější třídu rozdělení, která v sobě jako speciální případ zahrnuje právě i mnohorozměrné normální rozdělení.

Definice 2.5 *Řekneme, že p-rozměrný náhodný vektor* $\boldsymbol{X}_p$ *má* sféricky symetrické (sférické) rozdělení*, jestliže má absolutně spojité rozdělení s hustotou ve tvaru*

$$f(\boldsymbol{x}) = g(\|\boldsymbol{x}\|). \tag{2.9}$$

Řekneme, že p-rozměrný náhodný vektor $\boldsymbol{X}$ *má* elipticky symetrické (eliptické) rozdělení*, jestliže má absolutně spojité rozdělení s hustotou ve tvaru*

$$f_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{x}) = |\boldsymbol{\Sigma}|^{-1/2} g((\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})), \tag{2.10}$$

kde $\boldsymbol{\mu} \in \mathbb{R}^p$ *a* $\boldsymbol{\Sigma} \in \mathrm{SPD}(p)$.

Typickým příkladem eliptického rozdělení je p-rozměrné normální $\mathsf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, pro které má funkce g v (2.10) vyjádření $g(y) = (2\pi)^{-p/2}\exp(-y/2)$.

Pro sférická a eliptická rozdělení platí řada zajímavých vlastnstí. Například pro náhondý vektor $\boldsymbol{X}$ se sférickým rozdělením můžeme snadno odvodit jeho střední hodnotu a varianční matici.

Věta 3 *Nechť náhodný vektor* $\boldsymbol{X}_p = (X_1, X_2, \ldots, X_p)'$ *má sférické rozdělení s hustotou (2.9). Pak za předpokladu, že existuje jeho konečná střední hodnotu a varianční matici platí*

$$\mathsf{E}\,\boldsymbol{X}_p = \boldsymbol{0}, \tag{2.11}$$

$$\mathsf{var}(\boldsymbol{X}_p) = c\boldsymbol{I}, \tag{2.12}$$

kde $c > 0$ *je konstanta.*

Důkaz. Pro libovolnou ortogonální matici $\boldsymbol{O}$ typu $p \times p$ platí

$$g(\|\boldsymbol{O}\boldsymbol{x}\|) = g(\sqrt{\boldsymbol{x}'\boldsymbol{O}'\boldsymbol{O}\boldsymbol{x}}) = g(\|\boldsymbol{x}\|).$$

Tedy náhodné vektory $\boldsymbol{X}_p$ a $\boldsymbol{O}\boldsymbol{X}_p$ mají stejné rozdělení. Konkrétně pro $\boldsymbol{O} = -\boldsymbol{I}$ tak dostáváme, že vektory $\boldsymbol{X}_p$ a $-\boldsymbol{X}_p$ mají stejné rozdělení. Platí tedy

$$\mathsf{E}\,\boldsymbol{X}_p = \mathsf{E}(-\boldsymbol{X}_p) = -\,\mathsf{E}\,\boldsymbol{X}_p,$$

a tudíž $\mathsf{E}\,\boldsymbol{X}_p = \boldsymbol{0}$, čímž jsme dokázali (2.11).

Abychom dokázali (2.12), ukážeme nejprve, že složky $\boldsymbol{X}_p$ mají stejný rozptyl. Uvažujme libovolnou permutaci souřadnic $(\pi(1), \pi(2), \ldots, \pi(p))$. Za $\boldsymbol{O}$ pak vezmeme permutační matici, která má na místech $(i, \pi(i)), i = 1, 2 \ldots, p$ jedničky, jinak nuly. Pak rozdělení náhodných vektorů $\boldsymbol{X}_p$ a $\boldsymbol{X}_{p(\pi)} = (X_{\pi(1)}, X_{\pi(2)}, \ldots, X_{\pi(p)})$ jsou stejná. Tedy i pro jejich varianční matice platí $\mathsf{var}(\boldsymbol{X}_p) = \mathsf{var}(\boldsymbol{X}_{p(\pi)})$. Odtud plyne, že $\mathsf{var}(X_1) = \mathsf{var}(X_2) = \cdots = \mathsf{var}(X_p) = c$.

K dokončení důkazu zbývá ukázat, že pro $i, j = 1, 2, \ldots, p, i \neq j$, platí $\mathsf{cov}(X_i, X_j) = 0$. K tomu stačí pro každé $i = 1, 2, \ldots, p$ uvažovat ortogonální matici $\boldsymbol{O}_i$, která má na místě (i, i) hodnotu -1, na místech $(j, j), j \neq i$ hodnotu 1 a jinde 0. Pak mají vektory $\boldsymbol{X}_p$ a $\boldsymbol{O}_i\boldsymbol{X}_p$ stejná rozdělení a platí tedy $\mathsf{cov}(X_i, X_j) = \mathsf{cov}(-X_i, X_j) = -\mathsf{cov}(X_i, X_j) = 0, i, j = 1, 2, \ldots, p, i \neq j$. Tím jsme dokázali (2.12). □

Sférická a eliptická rozdělení spolu úzce souvisejí. Eliptické rozdělení s parametry $\boldsymbol{\mu} = \boldsymbol{0}$ a $\boldsymbol{\Sigma} = c\boldsymbol{I}$ je sférické. Na druhou stranu, pokud má náhodný vektor $\boldsymbol{X}_p$ eliptické rozdělení s hustotou (2.10), můžeme jej vyjádřit jako

$$\boldsymbol{X}_p = \boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{Y}_p, \tag{2.13}$$

kde $\boldsymbol{A}$ je regulární matice typu $p \times p$, $\boldsymbol{A}\boldsymbol{A}' = \boldsymbol{\Sigma}$ a $\boldsymbol{Y}_p$ je náhodný vektor se sférickým rozdělením. Tato paramatrizace není jednoznačná, neboť na místě $\boldsymbol{A}$ můžeme použít $\boldsymbol{O}\boldsymbol{A}$ pro libovolnou $\boldsymbol{O}$ ortogonální. Ničemu to ale nevadí. Na základě této parametrizace a věty 3 pak můžeme získat vyjádření střední hodnoty a varianční matice eliptického rozdělení. Pro střední hodnotu platí

$$\mathsf{E}\,\boldsymbol{X}_p = \boldsymbol{\mu} + \boldsymbol{A}\,\mathsf{E}\,\boldsymbol{Y}_p = \boldsymbol{\mu}$$

a pro varianční matici

$$\mathsf{var}(\boldsymbol{X}_p) = \boldsymbol{A}\,\mathsf{var}(\boldsymbol{Y}_p)\boldsymbol{A}' = c\boldsymbol{A}\boldsymbol{I}\boldsymbol{A}' = c\boldsymbol{\Sigma}. \tag{2.14}$$

Matici $\boldsymbol{\Sigma}$ budeme nazývat *disperzní* (v anglické literatuře bývá označována jako *dispersion matrix* nebo také *scatter matrix*). Vzhledem k (2.14) se v modelech s eliptickým rozdělením často místo varianční matice odhaduje přímo disperzní matice $\boldsymbol{\Sigma}$.

Pokud má vektor $\boldsymbol{X}_p$ eliptické rozdělení s parametry $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, pak pro libovolnou afinní transformaci $\boldsymbol{Z}_p = \boldsymbol{b} + \boldsymbol{B}\boldsymbol{X}_p, \boldsymbol{b} \in \mathbb{R}^p$, $\boldsymbol{B}$ regulární, je rozdělení vektoru $\boldsymbol{Z}_p$ taktéž eliptické s parametry $(\boldsymbol{b}+\boldsymbol{B}\boldsymbol{\mu}, \boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}')$. Při odhadování střední hodnoty a disperzní matice eliptického rozdělení je tedy přirozeným požadavkem ekvivariance těchto odhadů. Při splnění tohoto požadavku můžeme pro tyto odhady získat určitou představu o jejich limitním hodnotách.

Uvažujme tedy náhodný výběr $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$, který pochází z eliptického rozdělení $\mathcal{P}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}$ s hustotou (2.10). Pro $(\boldsymbol{T}_n(\boldsymbol{X}), \boldsymbol{C}_n(\boldsymbol{X}))$ afinně ekvivariantní odhad $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ označme pro $n \to \infty$ jeho limitní hodnotu $(\boldsymbol{T}_\infty(\boldsymbol{X}), \boldsymbol{C}_\infty(\boldsymbol{X}))$. Maronna a kol. (2006) ukázali, že platí $\boldsymbol{T}_\infty(\boldsymbol{X}) = \boldsymbol{\mu}$ a $\boldsymbol{C}_\infty(\boldsymbol{X}) = c\boldsymbol{\Sigma}$, kde c je konstanta.

Dobrý robustní odhad by zároveň měl vykazovat dobré výsledky i v případě, že výběr $\boldsymbol{X}$ neobsahuje odlehlá pozorování. Určitou představu o tomto chování můžeme získat např. z *relativní asymptotické eficience*. Tento postup používají Maronna a kol. (2006). Uvažujme výběr $\boldsymbol{X}$ z rozdělení $P_{\boldsymbol{\theta}}$ a označme $\hat{\boldsymbol{\theta}}_n$ maximálně věrohodný odhad parametru $\boldsymbol{\theta} \in \mathbb{R}^p$ s asymptotickou varianční maticí $\boldsymbol{V}_0$. Dále označme $\tilde{\boldsymbol{\theta}}_n$ námi zkoumaný odhad $\boldsymbol{\theta}$ a předpokládejme, že jeho asymptotické rozdělení je normální s varianční maticí $\boldsymbol{V}$. Pak můžeme definovat relativní asymptotickou eficienci jako

$$\mathrm{ef}(\tilde{\boldsymbol{\theta}}_n) = \min_{\boldsymbol{c}\neq\boldsymbol{0}} \frac{\boldsymbol{c}'\boldsymbol{V}_0\boldsymbol{c}}{\boldsymbol{c}'\boldsymbol{V}\boldsymbol{c}}.$$

Pro takto definovanou eficienci platí $\mathrm{ef}(\tilde{\boldsymbol{\theta}}_n) = \lambda_1(\boldsymbol{V}^{-1}\boldsymbol{V}_0)$, kde $\lambda_1(\boldsymbol{A})$ značí největší vlastní číslo $\boldsymbol{A}$. Pokud platí $\boldsymbol{V} = a\boldsymbol{V}_0$ pro nějakou konstantu a, pak $\mathrm{ef}(\tilde{\boldsymbol{\theta}}_n) = 1/a$.

V případě, že výběr $\boldsymbol{X}$ pochází z normálního rozdělení $\mathsf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, je maximálně věrohodným odhadem parametru $\boldsymbol{\theta} = \boldsymbol{\mu}$ výběrový průměr $\hat{\boldsymbol{\theta}}_n = \bar{\boldsymbol{x}}$, jehož varianční matice je $\boldsymbol{V}_0 = \boldsymbol{\Sigma}$. Maronna a kol. (2006) uvádějí, že pro $\tilde{\boldsymbol{\theta}}_n = \boldsymbol{T}(\boldsymbol{X})$ afinně ekvivariantní odhad $\boldsymbol{\mu}$ je jeho asymptotická varianční matice $\boldsymbol{V} = v\boldsymbol{\Sigma}$, kde v je konstanta. Jeho relativní asymptotická eficience je tedy $1/v$ a závisí pouze na typu odhadu, nikoliv na parametrech ($\boldsymbol{\mu}$,$\boldsymbol{\Sigma}$).

2.1 Robustní mnohorozměrné odhady

Nyní popíšeme několik robustních odhadů střední hodnoty a varinační matice, kterým je v literatuře věnována velká pozornost a které se používají v praxi. V kapitole 3 je pak použijeme k robustifikaci klasifikačních metod z kapitoly 1. Protože budeme v této kapitole často používat Mahalanobisovu vzdálenost, uvedeme nyní její definici.

Definice 2.6 *Pro* $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p$ *a* $\boldsymbol{A} \in \mathrm{SPD}(p)$ *definujeme* Mahalanobisovu vzdálenost $\boldsymbol{x}$ od $\boldsymbol{y}$ měřenou v metrice indukované maticí $\boldsymbol{A}$ *jako*

$$d(\boldsymbol{x}; \boldsymbol{y}, \boldsymbol{A}) = \sqrt{(\boldsymbol{x} - \boldsymbol{y})' \boldsymbol{A}^{-1} (\boldsymbol{x} - \boldsymbol{y})}. \tag{2.15}$$

Poznamenejme, že budeme Mahalanobisovu vzdálenost (2.15) často označovat stručně jen vzdálenost.

2.1.1 MVE-odhad

Jako první popíšeme odhad, jehož autorem je Rousseeuw (1985). Tento odhad je založený na nalezení elipsoidu s nejmenším objemem mezi všemi elipsoidy, které pokrývají požadovanou část dat. Budeme ho značit MVE-odhad podle anglického termínu *minimum volume ellipsoid estimator*.

Definice 2.7 *Nechť* $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ *je náhodný výběr z* $P_{\boldsymbol{\mu},\boldsymbol{\Sigma}} \in \mathcal{P}_p$, $n \geq p + 1$. *Pak definujeme* MVE-odhad $(\boldsymbol{T}_{MVE}, \boldsymbol{C}_{MVE}) \in (\mathbb{R}^p, \mathrm{SPD}(p))$ *jako řešní úlohy minimalizace* $|\boldsymbol{C}|$ *vzhledem k podmínce*

$$\#\{i : (\boldsymbol{x}_i - \boldsymbol{T})' \boldsymbol{C}^{-1} (\boldsymbol{x}_i - \boldsymbol{T}) \leq c^2\} \geq h,$$

kde $h = \lfloor (n + p + 1)/2 \rfloor$ *a* $\#A$ *značí počet prvků množiny* A.

Konstantu c lze volit tak, aby $\boldsymbol{C}_{\mathrm{MVE}}$ byl konzistentní odhah varianční matice předpokládaného rozdělení. Pokud výběr $\boldsymbol{X}$ pochází z eliptického rozdělení s hustotou (2.10), pak je přirozenou volbou ta hodnota c, pro kterou

$$P\{(\boldsymbol{X} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{X} - \boldsymbol{\mu}) \leq c^2\} = \int\limits_{\|\boldsymbol{x}\| \leq c} g(\|\boldsymbol{x}\|) \, \mathrm{d}\, \boldsymbol{x} = \frac{1}{2}.$$

Speciálně pochází-li výběr z normálního rozdělení, pak $c^2 = \chi^2_p(0,5)$.

MVE-odhad je afinně ekvivariantní. Lopuhaä a Rousseeuw (1991) odvodili bod selhání MVE-odhadu v případě, že je výběr $\boldsymbol{X}$ v obecné pozici. Ukázali, že pak dosahuje maximální hodnoty pro afinně ekvivariantní odhady (2.2). Platí tedy $\epsilon_n^*(\boldsymbol{T}_{\mathrm{MVE}}, \boldsymbol{C}_{\mathrm{MVE}}) = \lfloor (n-p+1)/2 \rfloor / n$ a bod selhání MVE-odhadu je tedy $\epsilon^*(\boldsymbol{T}_{\mathrm{MVE}}, \boldsymbol{C}_{\mathrm{MVE}}) = \frac{1}{2}$.

Asymptotickými vlastnostmi MVE-odhadu se zabýval Davies (1992). Odvodil tvar asymptotického rozdělení $(\boldsymbol{T}_{\mathrm{MVE}}, \boldsymbol{C}_{\mathrm{MVE}})$ a ukázal, že MVE-odhad konverguje k tomuto rozdělení v řádu $n^{-1/3}$ (detaily viz Davies, 1992). Díky této pomalé konvergenci je MVE-odhad značně neefícientní. Jednou možností, jak zvýšit eficienci odhadu, je využít informaci, kterou máme o kontaminaci konkrétního výběru $\boldsymbol{X}$ odlehlými pozorováními. Pokud víme, že kontaminace není větší než $\alpha, 0 < \alpha \leq 1/2$, můžeme v definici 2.7 použít $h = \lfloor n(1-\alpha) \rfloor + 1$. Bod selhání MVE-odhadu se pak sníží na α.

Lopuhaä a Rousseeuw (1991) uvádějí druhý postup, který může vést ke zlepšení eficience a přitom zachovává bod selhání odhadu. Uvažujme obecný odhad $(\boldsymbol{T}(\boldsymbol{X}), \boldsymbol{C}(\boldsymbol{X}))$ a váhovou funkci $w : [0, \infty) \to [0, \infty)$, která je nerostoucí, omezená, kladná pro $y \in [0, c]$ a nulová pro $y \in [c_1, \infty)$ pro nějaké $c_1 > c$. Pak můžeme definovat převážený odhad

$$\boldsymbol{T}_1(\boldsymbol{X}) = \frac{\sum_{i=1}^n w(d_i)\boldsymbol{x}_i}{\sum_{i=1}^n w(d_i)} \tag{2.16}$$

$$\boldsymbol{C}_1(\boldsymbol{X}) = \frac{\sum_{i=1}^n w(d_i)(\boldsymbol{x}_i - \boldsymbol{T}_1(\boldsymbol{X}))(\boldsymbol{x}_i - \boldsymbol{T}_1(\boldsymbol{X}))'}{\sum_{i=1}^n w(d_i)}, \tag{2.17}$$

kde $d_i = d(\boldsymbol{x}_i; \boldsymbol{T}(\boldsymbol{X}), \boldsymbol{C}(\boldsymbol{X}))$. Častou volbou funkce $w(y)$ je $\mathrm{I}\{y \in [0, c_1]\}$, kde $\mathrm{I}\{A\}$ značí indikátor jevu A.

V praxi není často možné MVE-odhad spočítat přesně, a proto se pro výpočet jeho přibližné hodnoty používá následující postup. Náhodně vybereme množinu $p+1$ bodů z $\boldsymbol{X}$, označme ji K, a spočteme jejich výběrový průměr $\overline{\boldsymbol{x}}_K$ a výběrovou varianční matici $\boldsymbol{S}_K$. Elipsoid určený $\overline{\boldsymbol{x}}_K$ a $\boldsymbol{S}_K$ pak zvětšíme nebo zmenšíme tak, aby obsahoval právě h bodů z $\boldsymbol{X}$. Tento postup opakujeme N-krát. Jako odhad $(\boldsymbol{T}_{\mathrm{MVE}}, \boldsymbol{C}_{\mathrm{MVE}})$ pak vezmeme parametry toho elipsoidu, který má mezi těmito N elipsoidy nejmenší objem. Pokud jsou rozsah výběru n a dimenze p přijatelně malé, můžeme pracovat se všemi podvýběry, tj. $N = \binom{n}{p+1}$. V opačném případě vybereme N podvýběrů náhodně.

2.1.2 MCD-odhad

Rousseeuw (1985) navrhl ještě jiný odhad, který má výhodnější asymptotické vlastnosti, a proto se používá častěji než MVE-odhad. Budeme ho značit MCD-odhad podle anglického termínu *minimum covariance determinant estimator*. Jak anglický název napovídá, je založený na nalezení varianční matice s minimálním determinantem.

Definice 2.8 *Nechť náhodný výběr* $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ *pochází z rozdělení* $P_{\boldsymbol{\mu},\boldsymbol{\Sigma}} \in \mathcal{P}_p$, $n \geq p+1$. *Z výběru* $\boldsymbol{X}$ *vybereme takovou podmnožinu pozorování* $\boldsymbol{X}_h = (\boldsymbol{x}_{i_1}, \boldsymbol{x}_{i_2}, \ldots, \boldsymbol{x}_{i_h}), h = \lfloor (n+p+1)/2 \rfloor$, *pro kterou má výběrová varianční matice* $\boldsymbol{S}_h$ *výběru* $\boldsymbol{X}_h$ *minimální determinant. Potom definujeme* MCD-odhad

$(\boldsymbol{T}_{MCD}(\boldsymbol{X}), \boldsymbol{C}_{MCD}(\boldsymbol{X}))$ *jako*

$$\boldsymbol{T}_{MCD}(\boldsymbol{X}) = \frac{1}{h}\sum_{j=1}^{h} \boldsymbol{x}_{i_j} \tag{2.18}$$

$$\boldsymbol{C}_{MCD}(\boldsymbol{X}) = c_k \frac{1}{h-1}\sum_{j=1}^{h} (\boldsymbol{x}_{i_j} - \boldsymbol{T}_{MCD}(\boldsymbol{X}))(\boldsymbol{x}_i - \boldsymbol{T}_{MCD}(\boldsymbol{X}))'. \tag{2.19}$$

Konstanta c_k je, podobně jako pro MVE-odhad, volena tak, aby byl odhad (2.19) konzistentní. Pison, Van Aelst a Willems (2002) ukázali, že velikost odhadu $\boldsymbol{C}_{\mathrm{MCD}}(\boldsymbol{X})$ je v případě malého rozsahu výběru n podhodnocena - za odlehlá je označno více pozorování než by mělo být. Doporučují proto odhad (2.19) vynásobit ještě korekčním faktorem c_{mv}. Konstrukci c_{mv} navrhli tak, aby se jeho hodnota blížila jedné pro n jdoucí do nekonečna (detaily viz Pison a kol., 2002).

MCD-odhad je afinně ekvivariantní a jeho bod selhání dosahuje maximální hodnoty pro afinně ekvivariantní odhady

$$\epsilon_n^*(\boldsymbol{T}_{\mathrm{MCD}}, \boldsymbol{C}_{\mathrm{MCD}}; \boldsymbol{X}) = \lfloor (n-p+1)/2 \rfloor / n,$$

a tedy $\epsilon^*(\boldsymbol{T}_{\mathrm{MCD}}, \boldsymbol{C}_{\mathrm{MCD}}) = \frac{1}{2}$.

Butler, Davies a Jhun (1993) se zabývali asymptotickými vlastnostmi MCD-odhadu. Mimo jiné ukázali, že pokud výběr $\boldsymbol{X}$ pochází z eliptického rozdělení, konverguje $\sqrt{n}(\boldsymbol{T}_{\mathrm{MCD}n}(\boldsymbol{X}) - \boldsymbol{\mu})$ k p-rozměrnému normálnímu rozdělení s nulovou střední hodnotou.

Pro zvýšení eficience MCD-odhadu je možné použít stejné postupy jako v kapitole 2.1.1 pro MVE-odhad. Tedy změnu konstanty h v definici 2.8 na základě znalosti kontaminace výběru nebo jednokrokové převážení (2.16), (2.17).

V praxi je pro velký rozsah výběru n nemožné v reálném čase spočítat MCD-odhad podle definice. Z tohoto důvodu se MCD-odhad příliš nepoužíval, dokud Rousseeuw a Van Driessen (1999) nepřišli s rychlým algoritmem pro jeho výpočet. Tento algoritmus nazvali FAST-MCD a nyní ho popíšeme.

1. Náhodně zvolíme počáteční množinu $H_1 \subset \{1, \ldots, n\}, |H_1| = h$.
2. C-krok
   (a) Pro množinu $H_i \subset \{1, \ldots, n\}, |H_i| = h$ spočteme odhady střední hodnoty a varianční matice $\boldsymbol{T}_i = \frac{1}{h}\sum_{j \in H_i} \boldsymbol{x}_j$ a
   $$\boldsymbol{C}_i = \frac{1}{h}\sum_{j \in H_i} (\boldsymbol{x}_j - \boldsymbol{T}_i)(\boldsymbol{x}_j - \boldsymbol{T}_i)'.$$
   (b) V případě, že $|\boldsymbol{C}_i| \neq 0$, vypočítáme vzdálenosti $\boldsymbol{x}_j$ od $\boldsymbol{T}_i$, tj. $d_i(j) = d(\boldsymbol{x}_j; \boldsymbol{T}_i, \boldsymbol{C}_i), j = 1, 2, \ldots, n$.
   (c) Najdeme takovou permutaci π množiny $\{1, \ldots, n\}$, pro kterou platí $d_i(\pi(1)) \leq d_i(\pi(2)) \leq \ldots \leq d_i(\pi(n))$.
   (d) Položíme $H_{i+1} = \{\pi(1), \pi(2), \ldots, \pi(h)\}$.
3. Opakujeme C-krok, dokud pro nějaké $i = m$ nenastane jedna z možností $|\boldsymbol{C}_m| = 0$ nebo $|\boldsymbol{C}_m| = |\boldsymbol{C}_{m-1}|$.

Celý postup od výběru počáteční množiny H_i až po nalezení $\boldsymbol{C}_m$ několikrát zopakujeme a jako MCD-odhad vezmeme tu dvojici $(\boldsymbol{T}_m, \boldsymbol{C}_m)$, pro kterou je $|\boldsymbol{C}_m|$ minimální.

Krok 3 vždy vede k nalezení $\boldsymbol{C}_m$, neboť existuje jen konečně mnoho h-prvkových podmnožin $\{1, \ldots, n\}$ a platí $|\boldsymbol{C}_i| \leq |\boldsymbol{C}_{i-1}|$, přičemž rovnost nastane právě tehdy, když $\boldsymbol{T}_i = \boldsymbol{T}_{i-1}$ a $\boldsymbol{C}_i = \boldsymbol{C}_{i-1}$ (viz Rousseeuw a Van Driesen, 1999). Autoři také uvádějí jinou možnost volby počáteční množiny H_1:

1.* Vezmeme množinu J tvořenou $p+1$ náhodně vybranými body z $\boldsymbol{X}$ a vypočítáme $\boldsymbol{T}_0 = \frac{1}{p+1} \sum_{j \in J} \boldsymbol{x}_j$ a $\boldsymbol{C}_0 = \frac{1}{p+1} \sum_{j \in J} (\boldsymbol{x}_j - \boldsymbol{T}_0)(\boldsymbol{x}_j - \boldsymbol{T}_0)'$. Spočteme vzdálenosti $d_0(j) = d(\boldsymbol{x}_j; \boldsymbol{T}_0, \boldsymbol{C}_0)$, seřadíme je podle velikosti $d_0(\pi(1)) \leq d_0(\pi(2)) \leq \ldots \leq d_0(\pi(n))$ a položíme $H_1 = \{\pi(1), \ldots, \pi(h)\}$.

Autoři tuto volbu upřednostňují, neboť pro hodně kontaminovaná data se zvyšuje pravděpodobnost, že do H_1 budou vybrána i odlehlá pozorování. Autoři dále uvádějí dva způsoby, jak výpočet odhadu pomocí FAST-MCD algoritmu výrazně zrychlit. Zjistili, že už po druhém nebo třetím opakování C-kroku je většinou patrné, zda algoritmus povede k robustnímu nebo nerobustnímu řešení. Proto pro každou počáteční množinu H_1 provedeme pouze dva C-kroky algoritmu, tj. nalezneme množinu H_3. Další C-kroky algoritmu provedeme pouze pro několik málo (např. deset) množin H_3, pro které je $|\boldsymbol{C}_3|$ nejmenší. Dále navrhli autoři zrychlení algoritmu pro velký rozsah výběru n:

- Pro $n \leq 600$ provedeme 500 výběrů H_1 a postupujeme podle algoritmu.
- Pro $600 < n \leq 1500$ rozdělíme pozorvání náhodně do nejvýše čtyř skupin tak, aby každá skupina obsahovala alespoň 300 pozorování a aby skupiny byly pokud možno stejně velké. V každé skupině o n_{skup} pozorováních provedeme několik výběrů H_1 (dohromady ve všech skupinách 500 výběrů), kde místo n a h počítáme s n_{skup} a $h_{skup} = \lfloor n_{skup}(h/n) \rfloor$, a spočteme pro ně dva C-kroky. V každé skupině vybereme deset nejlepších výsledků $(\boldsymbol{T}_{skup}, \boldsymbol{C}_{skup})$. Spojíme skupiny dohromady a dále počítáme již s celým výběrem. Pro každou dvojici $(\boldsymbol{T}_{skup}, \boldsymbol{C}_{skup})$ provedeme dva C-kroky s n a h. Vezmeme opět deset nejlepších výsledků $(\boldsymbol{T}_{spoj}, \boldsymbol{C}_{spoj})$ a s každým z nich pokračujeme s C-kroky. Nakonec najdeme konečný odhad $(\boldsymbol{T}_{kon}, \boldsymbol{C}_{kon})$ s nejmenším diskrimnantem varianční matice.
- Pro $n > 1500$ náhodně vybereme $n_{spoj} = 1500$ pozorování a ta rozdělíme do pěti skupin po $n_{skup} = 300$ pozorováních. Stejně jako v předchozím bodě najdeme pro každou skupinu deset dvojic $(\boldsymbol{T}_{skup}, \boldsymbol{C}_{skup})$, které dávají po dvou C-krocích nejlepší výsledky. Spojíme skupiny pozorování dohromady a pro výslednou skupinu o velikosti n_{spoj} provedeme pro každou z padesáti dvojic $(\boldsymbol{T}_{skup}, \boldsymbol{C}_{skup})$ dva C-kroky, tentokrát počítáme s n_{spoj} a $h_{spoj} = \lfloor n_{spoj}(h/n) \rfloor$. Vezmeme opět deset nejlepších výsledků $(\boldsymbol{T}_{spoj}, \boldsymbol{C}_{spoj})$ a s každým z nich pokračujeme s C-kroky s tím, že nyní už do výpočtů zahrneme všechna pozorování, tj. počítáme s n a h. Jako konečné řešení $(\boldsymbol{T}_{kon}, \boldsymbol{C}_{kon})$ vezmeme to s nejmenším diskriminantem varianční matice.

Váženou obdobu MCD-odhadu navrhl Kalina (2012). Metoda je však výpočetně velmi náročná pro data s větší dimenzí.

2.1.3 M-odhad

Odhad, který popíšeme v této kapitole, vychází z konceptu maximální věrohodnosti. Pro $\boldsymbol{X}$ výběr z eliptického rozdělelní s hustotou (2.10) má věrohodnostní funkce pro odhad parametrů $\boldsymbol{\mu}$ a $\boldsymbol{\Sigma}$ tvar

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{|\boldsymbol{\Sigma}|^{n/2}} \prod_{i=1}^{n} g(d^2(\boldsymbol{x}_i; \boldsymbol{\mu}, \boldsymbol{\Sigma})).$$

Zlogaritmováním a vynásobením -2 převedeme úlohu na minimalizaci funkce

$$l(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = n \log |\boldsymbol{\Sigma}| + \sum_{i=1}^{n} \rho(d^2(\boldsymbol{x}_i; \boldsymbol{\mu}, \boldsymbol{\Sigma})),$$

kde $\rho(y) = -2\log(g(y))$. Za předpokladu, že $g(y)$ je spojitá a diferencovatelná, dospějeme derivováním k soustavě rovnic

$$\begin{aligned} \sum_{i=1}^{n} w(d^2(\boldsymbol{x}_i; \boldsymbol{\mu}, \boldsymbol{\Sigma}))(\boldsymbol{x}_i - \boldsymbol{\mu}) &= \mathbf{0} \\ \frac{1}{n} \sum_{i=1}^{n} w(d^2(\boldsymbol{x}_i; \boldsymbol{\mu}, \boldsymbol{\Sigma}))(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})' &= \boldsymbol{\Sigma}, \end{aligned}$$

kde $w(y) = -2g'(y)/g(y)$. Maximálně věrohodný odhad $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ je pak řešením této soustavy. Podotkněme, že pro p-rozměrné normální rozdělení $\mathsf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ je $w(s) = 1$, a tudíž dostáváme výběrový průměr a $(n-1)/n$-násobek výběrové varianční matice jako maximálně věrohodné odhady střední hodnoty a varianční matice.

Maronna (1976) vytvořil M-odhad, který je zobecněním maximálně věrohodného odhadu.

Definice 2.9 *Nechť* $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ *je náhodný výběr z* $P_{\boldsymbol{\mu},\boldsymbol{\Sigma}} \in \mathcal{P}_p$, $n \geq p+1$. *Pak definujeme* M-odhad $(\boldsymbol{T}_M, \boldsymbol{C}_M) \in (\mathbb{R}^p, \mathrm{SPD}(p))$ *jako řešní soustavy rovnic*

$$\frac{1}{n} \sum_{i=1}^{n} w_1(d_i)(\boldsymbol{x}_i - \boldsymbol{T}) = \mathbf{0} \tag{2.20}$$

$$\frac{1}{n} \sum_{i=1}^{n} w_2(d_i^2)(\boldsymbol{x}_i - \boldsymbol{T})(\boldsymbol{x}_i - \boldsymbol{T})' = \boldsymbol{C}, \tag{2.21}$$

kde $d_i = d(\boldsymbol{x}_i; \boldsymbol{T}, \boldsymbol{C})$ *a* w_1, w_2 *jsou reálné funkce na* $[0, \infty)$ *splňující:*

(i) funkce $w_i(y), i = 1, 2$ *jsou nezáporné, nerostoucí a spojité,*

(ii) funkce $\varphi_i(y) = y w_i(y), i = 1, 2$ *jsou omezené,*

(iii) funkce $\varphi_2(y)$ *je rostoucí na intervalu, kde* $\varphi_2(y) < K_2$ *pro* $K_2 = \sup_{y \geq 0} \varphi_2(y)$,

(iv) existuje y_0 *takové, že* $\varphi_2(y_0^2) > p$ *a* $w_1(y) > 0$ *pro* $y \leq y_0$ *(a tedy* $K_2 > p$*).*

Vlastnosti M-odhadů popsal Maronna (1976). Dokázal, že pro výběr z eliptického rozdělení a při splnění dodatečného předpokladu

(v) $K_2 > pn/(n-p)$,

existuje řešení soustavy (2.20), (2.21). Dále uvádí i postačující podmínky pro jednoznačnost tohoto řešení.

Jestliže M-odhad $(\boldsymbol{T}_{\mathrm{M}}, \boldsymbol{C}_{\mathrm{M}})$ existuje, pak pro jeho bod selhání platí

$$\epsilon^*(\boldsymbol{T}_{\mathrm{M}}, \boldsymbol{C}_{\mathrm{M}}) \leq \min(1/K_2, 1 - p/K_2)$$

(viz Maronna (1976)). Protože funkce $\min(1/K_2, 1-p/K_2)$ nabývá svého maxima pro $K_2 = p+1$, platí

$$\epsilon^*(\boldsymbol{T}_{\mathrm{M}}, \boldsymbol{C}_{\mathrm{M}}) \leq 1/(p+1)$$

a tudíž není vhodné používat M-odhad v případě větší dimenze p.

Za funkce $w_i = 1, 2$ v definici M-odhadu můžeme zvolit například

$$w_i(y) = \psi_i(y)/y,$$

kde $\psi_1 = \psi_H(y, k)$, $\psi_2 = \psi_H(y, k^2)$ a

$$\psi_H(y) = \min\{y, \max\{y, -k\}\}$$

je tzv. *Huberova ψ-funkce* a k je konstanta.

M-odhad je afinně ekvivariantní. Řešením soustavy rovnic

$$\begin{aligned} \mathsf{E}_P\, w_1(d)(\boldsymbol{x} - \boldsymbol{T}) &= \boldsymbol{0} \\ \mathsf{E}_P\, w_2(d^2)(\boldsymbol{x} - \boldsymbol{T})(\boldsymbol{x} - \boldsymbol{T})' &= \boldsymbol{C}, \end{aligned}$$

kde $d = d(\boldsymbol{x}; \boldsymbol{T}, \boldsymbol{C})$, získáme odpovídající statistický funkcionál $(\boldsymbol{T}_{\mathrm{M}}(P), \boldsymbol{C}_{\mathrm{M}}(P))$. M-odhad $(\boldsymbol{T}_{\mathrm{M}n}, \boldsymbol{C}_{\mathrm{M}n})$ konverguje v pravděpodobnosti k $(\boldsymbol{T}_{\mathrm{M}}(P), \boldsymbol{C}_{\mathrm{M}}(P))$. Za dodatečných předpokladů Maronna (1976) ukázal, že pro výběr z eliptického rozdělení jsou $\boldsymbol{T}_{\mathrm{M}n}$ a $\boldsymbol{C}_{\mathrm{M}n}$ asymptoticky nezávislé a že limitní rozdělení $\sqrt{n}(\boldsymbol{T}_{\mathrm{M}n} - \boldsymbol{T}_{\mathrm{M}}(P), \boldsymbol{C}_{\mathrm{M}n} - \boldsymbol{C}_{\mathrm{M}}(P))$ je mnohorozměrné normální s nulovou střední hodnotou. Pro odhad střední hodnoty $\boldsymbol{T}_{\mathrm{M}n}$ rovněž odvodil jeho asymptotickou varianční matici, která má tvar $a/b\boldsymbol{C}_{\mathrm{M}}(P)$, kde

$$a = p^{-1}\, \mathsf{E}\, \varphi_i^2 \left(d(\boldsymbol{x}; \boldsymbol{T}_{\mathrm{M}}(P), \boldsymbol{C}_{\mathrm{M}}(P))\right),$$

$$b = \mathsf{E}\left(w_1\left(d(\boldsymbol{x}; \boldsymbol{T}_{\mathrm{M}}(P), \boldsymbol{C}_{\mathrm{M}}(P))\right)(1 - p^{-1}) + \varphi_i'\left(d(\boldsymbol{x}; \boldsymbol{T}_{\mathrm{M}}(P), \boldsymbol{C}_{\mathrm{M}}(P))\right) p^{-1}\right).$$

Pro případ sférického rozdělení $P_{\boldsymbol{0}, s^2\boldsymbol{I}}$ také odvodil jeho influenční funkci a globální citlivost

$$\begin{aligned} IF(\boldsymbol{x}; \boldsymbol{T}_M, P) &= w_1(\|\boldsymbol{x}\|/s)\boldsymbol{x}/b, \\ \gamma^*(\boldsymbol{T}, P) &= K_1 s/b. \end{aligned}$$

Rovnici (2.20) z definice 2.9 lze přeformulovat na tvar

$$\frac{\sum_{i=1}^n w_1(d_i)\boldsymbol{x_i}}{\sum_{i=1}^n w_1(d_i)} = \boldsymbol{T}.$$

Toho lze využít při výpočtu M-odhadu pomocí iterativního algoritmu. Označme $\boldsymbol{T}_{(0)}$ a $\boldsymbol{C}_{(0)}$ počáteční odhady algoritmu. V $(m+1)$-ním kroku algoritmu spočteme odhady $\boldsymbol{T}_{(m+1)}$, $\boldsymbol{C}_{(m+1)}$ jako

$$\begin{aligned}\boldsymbol{T}_{(m+1)} &= \frac{\sum_{i=1}^{n} w_1(d_{i,(m)})\boldsymbol{x}_i}{\sum_{i=1}^{n} w_1(d_{i,(m)})} \\ \boldsymbol{C}_{(m+1)} &= \frac{1}{n}\sum_{i=1}^{n} w_2(d^2_{i,(m)})(\boldsymbol{x}_i - \boldsymbol{\mu}_{(m+1)})(\boldsymbol{x}_i - \boldsymbol{\mu}_{(m+1)})',\end{aligned}$$

kde $d_{i,(m)} = d(\boldsymbol{x}_i; \boldsymbol{\mu}_{(m)}, \boldsymbol{\Sigma}_{(m)})$. Pro tzv. *monotónní* M-odhady, kdy $d^2 w_2(d^2)$ je nerostoucí, konverguje tento algoritmus k jedinému řešení a výběr počátečních hodnot $\boldsymbol{T}_{(0)}$ a $\boldsymbol{C}_{(0)}$ tak ovlivní pouze počet kroků algoritmu nikoliv jeho výsledek.

2.1.4 S-odhad

Nyní popíšeme třídu odhadů, které zavedl Davies (1987) jako spojitou verzi MVE-odhadu. Používá trochu odlišnou definici, která je však ekvivalentní s definicí 2.10, pokud oslabíme některé její předpoklady (viz Lopuhaä, 1989).

Definice 2.10 *Nechť $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ je náhodný výběr z $P_{\boldsymbol{\mu},\boldsymbol{\Sigma}} \in \mathcal{P}_p$. Nechť $\rho : \mathbb{R} \to [0, \infty)$ je symetrická funkce, má spojitou derivaci ψ, $\rho(0) = 0$ a nechť existuje $0 < c < \infty$ takové, že ρ je rostoucí na $[0, c]$ a konstantní na $[c, \infty)$. Pak definujeme* S-odhad $(\boldsymbol{T}_S, \boldsymbol{C}_S) \in (\mathbb{R}^p, \mathrm{SPD}(p))$ *jako řešení úlohy minimalizace $|\boldsymbol{C}|$ vzhledem k podmínce*

$$\frac{1}{n}\sum_{i=1}^{n} \rho(d_i) = \frac{1}{n}\sum_{i=1}^{n} \rho(((\boldsymbol{x}_i - \boldsymbol{T})'\boldsymbol{C}^{-1}(\boldsymbol{x}_i - \boldsymbol{T}))^{1/2}) = b. \tag{2.22}$$

Konstantu b v (2.22) můžeme volit v souladu s předpokládaným rozdělením. Tak pro eliptické rozdělení s hustotou (2.10) můžeme vzít

$$b = \mathsf{E}\,\rho((\boldsymbol{X} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu})).$$

Označme $a = \sup \rho$, pak konstantu c můžeme volit tak, aby $0 < b/a = r \geq (n-p)/(2n)$. Lopuhaä a Rousseeuw (1991) dokázali, že při této volbě je bod selhání S-odhadu

$$\epsilon^*_n(\boldsymbol{T}_{\mathrm{S}}; \boldsymbol{X}) = \epsilon^*_n(\boldsymbol{C}_{\mathrm{S}}; \boldsymbol{X}) = \lceil nr \rceil / n,$$

kde $\lceil k \rceil$ značí horní celou část k, tj. nejmenší $z \in \mathbb{Z}$ takové, že $k \leq z$. Pro $r = (n-p)/(2n)$ dosahuje bod selhání maximální hodnoty $\lfloor (n-p+1)/2 \rfloor / n$, a tedy $\epsilon^*(\boldsymbol{T}_{\mathrm{S}}, \boldsymbol{C}_{\mathrm{S}}) = 1/2$.

S-odhad je afinně ekvivariantní. Častou volbou funkce ρ je tzv. *Tukeyho dvouváhová funkce*

$$\rho(y) = \begin{cases} \frac{y^2}{2} - \frac{y^4}{2c^2} + \frac{y^6}{6c^4}, & \text{pro } |y| \leq c, \\ \frac{c^2}{6}, & \text{pro } |y| \geq c. \end{cases} \tag{2.23}$$

Lopuhaä (1989) ukázal souvislost S-odhadů s M-odhady v tom smyslu, že řešení úlohy minimalizace $|\boldsymbol{C}|$ vzhledem k (2.22) je zároveň řešením soustavy

„M-typu“

$$\frac{1}{n}\sum_{i=1}^{n} u(d_i)(\boldsymbol{x}_i - \boldsymbol{T}) = \mathbf{0}$$

$$\frac{1}{n}\sum_{i=1}^{n} (pu(d_i)(\boldsymbol{x}_i - \boldsymbol{T})(\boldsymbol{x}_i - \boldsymbol{T})' - v(d_i)\boldsymbol{C}) = \mathbf{0},$$

kde $v(y) = \psi(y)y - \rho(y)y + b$ a $u(y) = \psi(y)/y$. Tohoto faktu využívá při odvození influenční funkce a asymptotických vlastností S-odhadů. Za dodatečných předpokladů (mimo jiné existence $\psi'(y)$, spojitost a omezenost $\psi'(y)$ a $u(y)$) ukázal, že influenční funkce S-odhadu je stejného typu jako pro M-odhady. S-odhad je konzistentní, tj. $(\boldsymbol{T}_{\mathrm{S}n}, \boldsymbol{C}_{\mathrm{S}n})$ konverguje skoro jistě k odopovídajícímu funkcionálu $(\boldsymbol{T}_{\mathrm{S}}(P), \boldsymbol{C}_{\mathrm{S}}(P))$, který získáme jako řešení úlohy minimalizace $|\boldsymbol{C}|$ vzhledem k podmínce

$$\mathsf{E}\,\rho\left(\sqrt{(\boldsymbol{x} - \boldsymbol{T})'\boldsymbol{C}^{-1}(\boldsymbol{x} - \boldsymbol{T})}\right) = b.$$

Rovněž asymptotické chování je podobné jako pro M-odhady. Tedy limitní rozdělení $\sqrt{n}(\boldsymbol{T}_{\mathrm{S}n} - \boldsymbol{T}_{\mathrm{S}}(P), \boldsymbol{C}_{\mathrm{S}n} - \boldsymbol{C}_{\mathrm{S}}(P))$ je normální s nulovou střední hodnotou.

Rocke (1996) poukázal na nedostatek S-odhadu pro běžně používané funkce $\rho(y)$. Pro rostoucí dimenzi se i přes vysoký bod selhání ztrácí jeho schopnost reálně označit pozorování za odlehlé. Tuto schopnost měří pomocí tzv. *asymptotické pravděpodobnosti zamítnutí* (*asymptotic rejection probability*, dále APR). Navrhl proto použití upravené dvouváhové funkce (*translated biweight* nebo také *t-biweight*)

$$\rho_t(y) = \begin{cases} \frac{y^2}{2}, & \text{pro } 0 \le y < M, \\ \frac{M^2}{2} - \frac{M^2(M^4 - 5M^2c^2 + 15c^4)}{30c^4} & \\ +y^2(\frac{1}{2} + \frac{M^4}{2c^4} - \frac{M^2}{c^2}) + y^3(\frac{4M}{3c^2} - \frac{4M^3}{3c^4}) & \\ +y^4(\frac{3M^2}{2c^4} - \frac{1}{2c^2}) - \frac{4My^5}{5c^4} + \frac{y^6}{6c^4}, & \text{pro } M \le y \le M + c, \\ \frac{M^2}{2} + \frac{c(5c + 16M)}{30}, & \text{pro } y > M + c. \end{cases} \tag{2.24}$$

Konstanty c a M je pak možné volit tak, aby se dosáhlo požadovaného bodu selhání a zároveň i APR (detaily viz Rocke, 1996).

Pro výpočet S-odhadu je možné použít itertivní algoritmus podobný tomu pro výpočet M-odhadu. V tomto případě však algoritmus vede k nalezení lokálního minima a výběr počátečních hodnot $\boldsymbol{T}_{(0)}$ a $\boldsymbol{C}_{(0)}$ je nyní důležitý. Doporučuje se volit MVE-odhad, který vykazuje dobré výsledky co se týče vychýlení, a to i pro větší dimenze p. Pro definici asymptotického vychýlení odhadů střední hodnoty a disperzní matice a pro určité konkrétní numerické výsledky viz Maronna a kol. (2006).

2.1.5 SD-odhad

Stahel (1981) a Donoho (1982) nezávisle na sobě představili odhad, který převádí problém nalezení robustního mnohorozměrného odhadu na jednorozměrný případ. Předtím, než tento odhad definujeme, zavedeme nejprve následující značení. Pro výběr $X = (x_1, x_2, \ldots, x_n)$ z jednorozměrného rozdělení budeme uvažovat jednorozměrné robustní statistiky $\mu(X)$ pro odhad střední hodnoty a $\sigma(X)$ pro

odhad rozptylu. Budeme předpokládat, že $\mu(X)$ je afinně ekvivariantní a $\sigma(X)$ je translačně invariantní a škálově ekvivariantní (tj. $\mu(aX+b) = a\mu(X)+b$ a $\sigma(aX+b) = |a|\sigma(X)$, pro libovolné $a, b \in \mathbb{R}$). Tyto vlastnosti splňují např. *medián*

$$\mathrm{med}(X) = (x_{(\lceil n/2 \rceil)} + x_{(\lfloor n/2 \rfloor + 1)})/2$$

a *mediánová absolutní odchylka od mediánu*

$$\mathrm{MAD}(X) = \mathrm{med}(|x_i - \mathrm{med}(X)|),$$

kde $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ jsou setřízené hodnoty X.

Definice 2.11 *Nechť $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ je náhodný výběr z $P_{\boldsymbol{\mu},\boldsymbol{\Sigma}} \in \mathcal{P}_p$, $n \geq p+1$. Nechť $w : \mathbb{R} \to [0, \infty)$ je spojitá nerostoucí omezená funkce a $y^2 w(y)$ je omezená funkce. Potom definujeme* Stahelův-Donohův odhad *(dále* SD-odhad*)*

$$\begin{aligned} \boldsymbol{T}_{SD}(\boldsymbol{X}) &= \frac{1}{n}\sum_{i=1}^{n} w_i \boldsymbol{x}_i \\ \boldsymbol{C}_{SD}(\boldsymbol{X}) &= \frac{1}{n-1}\sum_{i=1}^{n} \frac{w_i (\boldsymbol{x}_i - \boldsymbol{T}_{SD}(\boldsymbol{X}))(\boldsymbol{x}_i - \boldsymbol{T}_{SD}(\boldsymbol{X}))'}{\sum_{i=1}^{n} w_i}, \end{aligned}$$

kde $w_i = w(\mathrm{od}(\boldsymbol{x}_i, \boldsymbol{X}))$ a pro $\boldsymbol{y} \in \mathbb{R}^p$ je

$$\mathrm{od}(\boldsymbol{y}, \boldsymbol{X}) = \sup_{\boldsymbol{z} \in \mathbb{R}^p, \|\boldsymbol{z}\|=1} \frac{|\boldsymbol{z}'\boldsymbol{y} - \mu(\boldsymbol{z}'\boldsymbol{X})|}{\sigma(\boldsymbol{z}'\boldsymbol{X})}. \tag{2.25}$$

Myšlenka SD-odhadu spočívá v tom, že pokud je nějaké pozorování odlehlé, pak bude odlehlé i v nějaké jednorozměrné projekci. SD-odhad hledá směr $\boldsymbol{z}$, ve kterém se tato odlehlost nejvíce projeví. Funkce (2.25) pak měří míru odlehlosti pozorování $\boldsymbol{y}$ od výběru $\boldsymbol{X}$. Váhová funkce w tak logicky odlehlejším pozorováním přiřadí menší váhu a zmenší tak jejich vliv na výsledný odhad.

Díky požadavkům, které klademe na $\mu(X)$ a $\sigma(X)$ je $\mathrm{od}(\boldsymbol{y}, \boldsymbol{X})$ afinně invariantní, a SD-odhad je tudíž afinně ekvivariantní. Tyler (1994) dokázal, že pro výběr v obecné pozici může bod selhání SD-odhadu dosáhnout maximální hodnoty $\lfloor (n-p+1)/2 \rfloor / n$, a tedy $\epsilon^*(\boldsymbol{T}_{SD}, \boldsymbol{C}_{SD}) = 1/2$. Této maximální hodnoty je dosaženo např. pro

$$\mu(X) = \mathrm{med}(X), \quad \sigma(X) = \frac{1}{2}(s_{(k_1)} + s_{(k_2)}),$$

kde $s_{(i)}, i = 1, 2, \ldots, n$ jsou setřízené hodnoty $s_i = |x_i - \mathrm{med}(X)|$, $k_1 = \lceil (n+p)/2 \rceil$ a $k_2 = \lfloor (n+p)/2 \rfloor + 1$ (viz Gather a Hilker, 1997).

Maronna a Yohai (1995) porovnali několik typů váhových funkcí. Nejlepších výsledků v jejich simulacích dosahovala váhová funkce Huberova typu

$$w(y) = \min\left\{1, \left(\frac{c}{y}\right)^2\right\},$$

kde $c = \sqrt{\chi^2_p(0,95)}$. Další možnosti, jak volit váhovou funkci w, stejně jako další vlastnosti SD-odhadu, lze nalézt v knize Maronna a kol. (2006) a v ní uvedených odkazech.

V praxi není samozřejmě možné SD-odhad spočítat přesně. Maronna a Yohai (1995) používají pro výpočet SD-odhadu následující algoritmus. Uvažujme podvýběr $\tilde{\boldsymbol{X}}$ výběru $\boldsymbol{X}$ o velikosti p. Najdeme směr $\boldsymbol{z}, \|\boldsymbol{z}\| = 1$ ortogonální k nadrovině generované $\tilde{\boldsymbol{X}}$. Označme $\mathcal{Z}$ množinu všech takových $\boldsymbol{z}$, sestrojených na základě všech možných podvýběrů $\tilde{\boldsymbol{X}}$. Pokud je dimenze p a rozsah výběru n dostatečně malý, hledáme supremum v (2.25) přes množinu $\mathcal{Z}$. Pokud by bylo takové hledání výpočetně příliš náročné, omezíme se na podmnožinu $\mathcal{Z}_N$ obsahující N náhodně vybraných směrů ze $\mathcal{Z}$.

2.1.6 OGK-odhad

Gnanadesikan a Kettenring (1972) přistupují k odhadu varianční matice po jednotlivých složkách. Jejich odhad je založen na robustifikaci vztahu pro kovarianci dvou reálných náhodných veličin Y a Z

$$\mathsf{cov}(Y, Z) = \frac{1}{4}\left(\mathsf{var}(Y + Z) - \mathsf{var}(Y - Z)\right).$$

Definice 2.12 *Nechť $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ je náhodný výběr z $P_{\boldsymbol{\mu},\boldsymbol{\Sigma}} \in \mathcal{P}_p$, $n \geq p+1$ a označme $\boldsymbol{X}' = (\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^p)$. Nechť $s(X)$ je robustní odhad směrodatné odchylky pro náhodný výběr $X = (x_1, x_2, \ldots, x_n)$ z jednorozměrného rozdělení. Pak definujeme* Gnanadesikanův-Kettenringův odhad *(dále* GK-odhad*) varianční matice jako $\boldsymbol{C}_{GK} = (c_{jk})_{j,k=1}^{p}$, kde*

$$c_{jk} = \begin{cases} s^2(\boldsymbol{x}^j) & j = k, \\ s(\boldsymbol{x}^j)s(\boldsymbol{x}^k)\frac{1}{4}\left(s^2\left(\frac{\boldsymbol{x}^j}{s(\boldsymbol{x}^j)} + \frac{\boldsymbol{x}^k}{s(\boldsymbol{x}^k)}\right) - s^2\left(\frac{\boldsymbol{x}^j}{s(\boldsymbol{x}^j)} - \frac{\boldsymbol{x}^k}{s(\boldsymbol{x}^k)}\right)\right) & j \neq k. \end{cases}$$

Nedostatkem $\boldsymbol{C}_{GK}$ odhadu je to, že není afinně ekvivariantní. Na druhou stranu k jeho výpočtu nepotřebujeme odhad střední hodnoty. To může ušetřit výpočetní čas v situacích, kdy nás odhad střední hodnoty nezajímá (např. při analýze hlavních komponent). Pokud je situace opačná, můžeme odhad $\boldsymbol{\mu}$ obdobně jako $\boldsymbol{C}_{GK}$ sestrojit po složkách. Jestliže máme pro výběr X z jednorozměrného rozdělení robustní odhad střední hodnoty $t(X)$, pak můžeme pro výběr $\boldsymbol{X}$ definovat odhad střední hodnoty $\boldsymbol{\mu}$ jako $\boldsymbol{T}_{GK} = (t(\boldsymbol{x}^1), t(\boldsymbol{x}^2), \ldots, t(\boldsymbol{x}^p))'$.

Odhad $\boldsymbol{C}_{GK}$ není obecně pozitivně definitní matice. V knize Maronna a kol. (2006) je uveden algoritmus, jak tento nedostatek odstranit a získat zároveň odhad střední hodnoty. Tento algoritmus zde nyní uvedeme.

1. Označme $\boldsymbol{D} = \mathrm{diag}(s(\boldsymbol{x}^1), s(\boldsymbol{x}^2), \ldots, s(\boldsymbol{x}^p))$, tj. matici s prvky $d_{ii} = s(\boldsymbol{x}^i)$ a $d_{ij} = 0, i \neq j$, a sestrojme $\boldsymbol{Y} = \boldsymbol{D}^{-1}\boldsymbol{X}$. Sloupce matice $\boldsymbol{Y}$ nyní tvoří jednotlivá normalizovaná pozorování $\boldsymbol{y}_i = \boldsymbol{D}^{-1}\boldsymbol{x}_i, i = 1, 2, \ldots, n$ a pro její řádky platí $\boldsymbol{y}^{j\prime} = \boldsymbol{x}^{j\prime}/s(\boldsymbol{x}^j), j = 1, 2, \ldots, p$.

2. Sestavíme matici $\boldsymbol{R} = (r_{ij})$, kde

$$r_{jk} = \begin{cases} 1 & j = k, \\ \frac{1}{4}\left(s^2(\boldsymbol{y}^j + \boldsymbol{y}^k) - s^2(\boldsymbol{y}^j - \boldsymbol{y}^k)\right) & j \neq k. \end{cases}$$

   Matice $\boldsymbol{R}$ je vlastně odhadem korelační matice výběru $\boldsymbol{X}$ a zároveň odhadem varianční matice $\boldsymbol{Y}$.

3. Pro matici $\boldsymbol{R}$ sestrojíme její spektrální rozklad $\boldsymbol{R} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}'$, kde $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$, $U = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_p)$, $\lambda_i, i = 1, 2, \ldots, p$ jsou valstní čísla matice $\boldsymbol{R}$ a $\boldsymbol{\alpha}_i, i = 1, 2, \ldots, p$, jim odpovídající vlastní vektory. Protože $\boldsymbol{R}$ není obecně pozitivně definitní, nemusí být vlastní čísla λ_i nezáporná.

4. Definujme matici $\boldsymbol{Z} = \boldsymbol{U}'\boldsymbol{Y} = \boldsymbol{U}'\boldsymbol{D}^{-1}\boldsymbol{X}$. Dostáváme tak $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{Z}$, kde $\boldsymbol{A} = \boldsymbol{D}\boldsymbol{U}$.

5. Pro řádky matice $\boldsymbol{Z}$ spočteme $s(\boldsymbol{z}^j)$ a $t(\boldsymbol{z}^j), j = 1, 2, \ldots, p$ a definujme $\boldsymbol{\Gamma} = \mathrm{diag}(s^2(\boldsymbol{z}^1), s^2(\boldsymbol{z}^2), \ldots, s^2(\boldsymbol{z}^p))$, $\boldsymbol{t} = (t(\boldsymbol{z}^1), t(\boldsymbol{z}^2), \ldots, t(\boldsymbol{z}^p))'$.

6. Nyní definujeme *ortogonalizovaný Gnanadesikanův-Kettenringův odhad* (dále *OGK-odhad*) střední hodnoty $\boldsymbol{\mu}$ a varianční matice $\boldsymbol{\Sigma}$

$$\begin{aligned} \boldsymbol{T}_{\mathrm{OGK}}(\boldsymbol{X}) &= \boldsymbol{A}\boldsymbol{t} && (2.26) \\ \boldsymbol{C}_{\mathrm{OGK}}(\boldsymbol{X}) &= \boldsymbol{A}\boldsymbol{\Gamma}\boldsymbol{A}'. && (2.27) \end{aligned}$$

Pokud by matice $\boldsymbol{R}$ z kroku 2 byla skutečnou varianční maticí $\boldsymbol{Y}$, byla by pozitivně definitní a platilo by $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p > 0$. Pak by řádky matice $\boldsymbol{Z}$ byly hlavními komponentami $\boldsymbol{Y}$ (viz kap. 4.1). Lze tedy očekávat, že $\boldsymbol{z}^j$ by mohly být nekorelované, nebo alespoň méně korelované než původní veličiny $\boldsymbol{x}^j$.

Dalšího zlepšení OGK-odhadu můžeme docílit iteracemi. Označme (2.26) a (2.27) jako $\boldsymbol{T}^{(0)}_{\mathrm{OGK}}(\boldsymbol{X})$ a $\boldsymbol{C}^{(0)}_{\mathrm{OGK}}(\boldsymbol{X})$. Předpokládejme, že na počátku $(k+1)$-ního iteračního kroku máme odhady $\boldsymbol{T}^{(k)}_{\mathrm{OGK}}(\boldsymbol{X})$, $\boldsymbol{C}^{(k)}_{\mathrm{OGK}}(\boldsymbol{X})$ a matici $\boldsymbol{A}$ spočtenou v k-té iteraci algoritmu v bodě 4. Označme $\boldsymbol{Z}^{(k)}$ odpovídající matici $\boldsymbol{Z}^{(k)} = \boldsymbol{A}^{-1}\boldsymbol{X}$. Pak spočteme odhad $(\boldsymbol{T}_{\mathrm{OGK}}(\boldsymbol{Z}^{(k)}), \boldsymbol{C}_{\mathrm{OGK}}(\boldsymbol{Z}^{(k)}))$ a data transformujeme zpět do původního prostoru

$$\begin{aligned} \boldsymbol{T}^{(k+1)}_{\mathrm{OGK}}(\boldsymbol{X}) &= \boldsymbol{A}\boldsymbol{T}_{\mathrm{OGK}}(\boldsymbol{Z}^{(k)}) \\ \boldsymbol{C}^{(k+1)}_{\mathrm{OGK}}(\boldsymbol{X}) &= \boldsymbol{A}\boldsymbol{C}_{\mathrm{OGK}}(\boldsymbol{Z}^{(k)})\boldsymbol{A}'. \end{aligned}$$

Iterace jsou výpočetně velmi náročné a v praxi se provádějí jen dvě. Maronna a kol. (2006) uvádějí, že následující iterace dle výsledků simulací nevedou ke zlepšení.

Jako $t(X)$ a $s(X)$ se používají

$$\begin{aligned} t(X) &= \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i} \\ s^2(X) &= \frac{\mathrm{MAD}^2(X)}{n} \sum_{i=1}^n \rho\left(\frac{x_i - t(X)}{\mathrm{MAD}(X)}\right), \end{aligned}$$

kde $w_i = w\left((x_i - \mathrm{med}(X))/\,\mathrm{MAD}(X)\right)$, $w(y) = \max\left(0, (1 - (y/k)^2)^2\right)$ a $\rho(y) = \min(c^2, y^2)$ s volbou konstant $k = 4,5$ a $c = 3$.

Maronna a kol. (2006) uvádějí, že pokud bod selhání odhadů $t(X)$ a $s(X)$ není menší než ϵ, pak ani bod selhání odhadu $(\boldsymbol{T}_{\mathrm{OGK}}(\boldsymbol{X}), \boldsymbol{C}_{\mathrm{OGK}}(\boldsymbol{X}))$ není menší než ϵ.

Kapitola 3

Robustní klasifikační analýza

V kapitole 1 jsme popsali klasifikační pravidla 4 a 5, která jsou založena na použití výběrových odhadů $\overline{\boldsymbol{x}}^i$ a $\boldsymbol{S}^i$ v (1.5) a (1.6). Protože jsou tyto odhady citlivé na přítomnost odlehlých pozorování, nemusí být výběrová klasifikační pravidla 4 a 5 příliš spolehlivá, pokud výběry $\boldsymbol{X}^i$ nějaká odlehlá pozorování obsahují. V této kapitole se zaměříme na možnosti robustifikace těchto klasifikačních pravidel.

3.1 Robustní kvadratická klasifikační analýza

Jednoduchou metodou, jak robustifikovat kvadratickou klasifikační analýzu, je nahradit výběrové odhady v (1.5) robustními odhady. Tento postup navrhuje například McLachlan (1992). Předpokládejme tedy, že máme pro každou skupinu π^i robsutní odhad střední hodnoty a varianční matice $(\boldsymbol{T}^i, \boldsymbol{C}^i)$. Můžeme použít libovolný z robustních odhadů z kapitoly 2.1. Pak získá kvadratické klasifikační pravidlo následující tvar.

Pravidlo 6 *Pozorování $\boldsymbol{x}$ zařadíme do skupiny π^i, jestliže*

$$d^i_{QR}(\boldsymbol{x}) > d^j_{QR}(\boldsymbol{x}) \text{ pro všechna } j = 1, 2, \ldots, k, j \neq i, \tag{3.1}$$

kde $d^i_{QR}(\boldsymbol{x}) = -\frac{1}{2}\ln|\boldsymbol{C}^i| - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{T}^i)'(\boldsymbol{C}^i)^{-1}(\boldsymbol{x} - \boldsymbol{T}^i) + \ln \hat{p}^i$.

V kapitole 5 je na základě simulací provedeno srovnání QDA a jeho robustní verze pro různé volby $(\boldsymbol{T}, \boldsymbol{C})$. Pro snadnější orientaci budeme jednotlivé robustní verze pravidla 6 označovat QDA a zkratkou příslušného použitého odhadu. Tak například QDA-MVE bude odkazovat na použití $(\boldsymbol{T}_{\mathrm{MVE}}, \boldsymbol{C}_{\mathrm{MVE}})$ v (3.1).

V kapitole 6 pak porovnáme kvadratická klasifikační pravidla sestavená na základě reálných dat.

3.2 Robustní lineární klasifikační analýza

V případě lineární klasifikační analýzy se nabízí otázka, jak odhadnout varianční matici $\boldsymbol{\Sigma}$. Předpokládejme, že pro každou skupinu $\pi^i, i = 1, 2, \ldots, k$ již máme robustní odhady střední hodnoty a varianční matice $\boldsymbol{T}^i(\boldsymbol{X}^i)$ a $\boldsymbol{C}^i(\boldsymbol{X}^i)$. Jednoduchým řešením je sestrojit robustní analogii sdružené výběrové varianční

matice (1.4), tedy

$$\boldsymbol{C}^A = \frac{1}{(N-k)} \sum_{i=1}^{k} (n^i - 1)\boldsymbol{C}^i,$$

kde $N = \sum_{i=1}^{k} n^i$. Tento postup označme jako metodu A a v souladu s tímto značením budeme psát také $\boldsymbol{T}^{A,i}(\boldsymbol{X}^i) = \boldsymbol{T}^i(\boldsymbol{X}^i)$.

Další metodou, označme ji B, kterou využily například Hubert a Van Driesen (2004), je centrovat data ve skupinách, čímž získáme sdružený výběr, na jehož základě pak spočteme odhad sdružené varianční matice. Definujme tedy matici centrovaných pozorování $\boldsymbol{Y} = (\boldsymbol{Y}^1, \boldsymbol{Y}^2, \ldots, \boldsymbol{Y}^k)$, kde $\boldsymbol{Y}^i = \boldsymbol{X}^i - \boldsymbol{T}^i \mathbf{1}'_{n^i}$. Jako odhad sdružené výběrové matice vezmeme robustní odhad $\boldsymbol{C}^B = \boldsymbol{C}(\boldsymbol{Y})$. Zároveň můžeme odhady střední hodnoty v jednotlivých skupinách upravit pomocí odhadu střední hodnoty centrovaného výběru $\boldsymbol{T}(\boldsymbol{Y})$, tj. $\boldsymbol{T}^{B,i}(\boldsymbol{X}^i) = \boldsymbol{T}^i(\boldsymbol{X}^i) + \boldsymbol{T}(\boldsymbol{Y})$.

Třetí možností (metoda C) je přizpůsobit rovnice pro výpočet robustního odhadu tak, abychom dostali přímo robustní odhad sdružené varianční matice. Obdobným způsobem potom můžeme upravit i algoritmy pro jejich výpočet. Tento postup uplatnili například Hawkins a McLachlan (1997) pro případ MCD-odhadu. Jejich MWCD-odhad (z anglického *minimum within-group covariance determinant estimamtor*) je stejně jako MCD-odhad založen na nalezení takové h-prvkové podmnožiny H bodů z $\boldsymbol{X} = (\boldsymbol{X}^1, \boldsymbol{X}^2, \ldots, \boldsymbol{X}^k)$, pro niž má sdružená varianční matice $\boldsymbol{C}^C_{\mathrm{MWCD}}$ minimální determinant. Označme $\delta^i_j = \mathrm{I}\{\boldsymbol{x}^i_j \in H\}$, pak

$$\begin{aligned}
\boldsymbol{T}^{C,i}_{\mathrm{MWCD}}(\boldsymbol{X}^i) &= \frac{\sum_{j=1}^{n^i} \delta^i_j \boldsymbol{x}^i_j}{\sum_{j=1}^{n^i} \delta^i_j} \\
\boldsymbol{C}^{C,i}_{\mathrm{MWCD}}(\boldsymbol{X}) &= \frac{\sum_{i=1}^{k}\sum_{j=1}^{n^i} \delta^i_j (\boldsymbol{x}^i_j - \boldsymbol{T}^{C,i}_{\mathrm{MWCD}}(\boldsymbol{X}^i))(\boldsymbol{x}^i_j - \boldsymbol{T}^{C,i}_{\mathrm{MWCD}}(\boldsymbol{X}^i))'}{\sum_{i=1}^{k}\sum_{j=1}^{n^i} \delta^i_j}.
\end{aligned}$$

Hledat řešení přes všechny možné h-prvkové podmnožiny by bylo výpočetně příliš náročné. Hubert a Van Driesen (2004) doporučují následující algoritmus.

1. V každé skupině spočteme MCD-odhad střední hodnoty $\boldsymbol{T}^i_{\mathrm{MCD}}(\boldsymbol{X}^i)$.
2. Pozorování centrujeme $\boldsymbol{Y} = (\boldsymbol{Y}^1, \boldsymbol{Y}^2, \ldots, \boldsymbol{Y}^k)$, $\boldsymbol{Y}^i = \boldsymbol{X}^i - \boldsymbol{T}^i_{\mathrm{MCD}} \mathbf{1}'_{n^i}$.
3. Spočteme $\boldsymbol{C}_{\mathrm{MCD}}(\boldsymbol{Y})$ a označíme $H = \bigcup_{i=1}^{k} H^i$, kde H^i značí množinu těch $\boldsymbol{x}^i_j \in \boldsymbol{X}^i$, pro které jim odpovídající $\boldsymbol{y}^i_j \in \boldsymbol{Y}$ minimalizují MCD kritérium.
4. V každé skupině vypočítáme odhady $\boldsymbol{T}^i(\boldsymbol{X}^i) = \sum_{j=1}^{n^i} \boldsymbol{x}^i_j \mathrm{I}\{\boldsymbol{x}^i_j \in H^i\}/|H^i|$.
5. Položíme $\boldsymbol{T}^i_{\mathrm{MWCD}} = \boldsymbol{T}^i(\boldsymbol{X}^i)$ a $\boldsymbol{C}^i_{\mathrm{MWCD}} = \boldsymbol{C}_{\mathrm{MCD}}(\boldsymbol{Y})$.

Kroky 2 až 4 můžeme několikrát zopakovat, autorky však používají jenom jednu iteraci. Nevýhodou tohoto algoritmu je, že může selhat pro malé skupiny.

V závislosti na typu robustního odhadu $(\boldsymbol{T}, \boldsymbol{C})$ a použité metodě (A,B nebo C) pro pro jeho výpočet dostává lineární klasifikační pravidlo následující podobu.

Pravidlo 7 *Pozorování $\boldsymbol{x}$ zařadíme do skupiny π^i, jestliže*

$$d^i_{LR}(\boldsymbol{x}) > d^j_{LR}(\boldsymbol{x}) \text{ pro všechna } j = 1, 2, \ldots, k, j \neq i \tag{3.2}$$

kde $d^i_{LR}(\boldsymbol{x}) = \boldsymbol{T}^{i\prime}\boldsymbol{C}^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{T}^{i\prime}\boldsymbol{C}^{-1}\boldsymbol{T}^i + \ln \hat{p}^i$.

Obdobně jako pro kvadratickou klasifikaci budeme různé verze pravidla 7 označovat zkratkou příslušného odhadu. Tak například při použití MWCD-odhadu označíme příslušné robustní pravidlo LDA-MWCD.

Todorov a Pires (2007) porovnali přesnost linárních klasifikačních pravidel pro různé typy odhadů jak na reálných datech tak v rozsáhlých simulacích. Konkrétně porovnávali klasický výběrový odhad (tj. $\overline{\boldsymbol{x}}$ a $\boldsymbol{S}$), MCD-odhad (pro jeho výpočet použili metody A,B), MWCD-odhad (spočtený dvěma typy algoritmů), OGK-odhad (získaný pomocí metody B), S-odhad s Tukeyho dvouváhovou funkcí (2.23) a s t-biweight funkcí (2.24) a S-odhad spočtený na základě algoritmu vytvořeného pro regresi popsaný v Salibian-Barrera a Yohai (2006). Na příkladě reálných dat, která neobsahovala odlehlá pozorování, ukázali (pomocí ApER a CV odhahů chyby klasifikace) srovnatelné chování všech robustních i klasické lineární klasifikace. V případě reálných dat s odlehlými pozorováními dávaly všechny robustní metody lepší výsledky než klasická metoda lineární klasifikace. Mezi robustními metodami pak nejnižší chyby klasifikace dosáhl MWCD-odhad pro oba použité algoritmy (detailní výsledky viz Todorov a Pires, 2007, kap. 3). Autoři provedli celou řadu simulací s různými kombinacemi vstupních parametrů p, k, n^i a s různým podílem a typem kontaminace dat (detaily viz Todorov a Pires, 2007, kap. 4). Z jejich závěrů vyplývá, že pro nekontaminovaná data dává robustní lineární klasifikace obdobné výsledky jako klasická. Nejlepších výsledků pro kontaminovaná data dosahuje OGK-odhad.

V kapitole 5 jsou pak uvedeny výsledky simulací, jejichž parametry byly nastaveny obdobným způsobem jako pro kvadratickou klasifikaci.

Kapitola 4

Redukce dimenzionality

Pro výpočet některých odhadů popsaných v kapitole 2.1, a tedy i pro klasifikaci na nich založenou, potřebujeme, aby byl rozsah výběru n dostatečně velký. Typicky vyžadujeme $n \geq p+1$. Tento předpoklad však nemusí být vždy splněn. Ale i když tomu tak je, můžou být výpočty odhadů pro velké p příliš náročné. Kalina (2013) také ukázal, že robustní klasifikační metody nejsou vhodné pro data s velkou dimenzí. Proto je někdy třeba před samotnou klasifikací zredukovat dimenzi p.

4.1 Analýza hlavních komponent

Nejčastěji používanou metodou pro redukci dimenzionality je tzv. *analýza hlavních komponent*, kterou budeme dále značit PCA (z anglického *principal component analysis*). Základní myšlenka PCA je transformovat původní data tak, abychom na novém souboru pozorovali $q < p$ nových veličin, tzv. hlavních komponent. Hlavní komponenty volíme jako linární kombinace původních p veličin tak, aby byly vzájemně nekorelované a aby několik prvních komponent vysvětlovalo co největší část variability obsažené v původním souboru dat. Tuto metodu nyní detailněji popíšeme. Budeme přitom vycházet z knihy Anděl (1985).

Předpokládejme tedy, že máme náhodný vektor $\boldsymbol{X}_p = (X_1, X_2, \ldots, X_p)'$ s rozdělením s varianční maticí $\boldsymbol{\Sigma} \in \mathrm{SPD}(p)$. Budeme předpokládat, že její vlastní čísla jsou různá. Označme je $\lambda_1 > \lambda_2 > \ldots > \lambda_p > 0$ a jim příslušející vlastní vektory $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_r$. Pro první komponentu hledáme vektor koeficientů lineární kombinace $\boldsymbol{c_1}, \boldsymbol{c_1}'\boldsymbol{c_1} = 1$ tak, aby náhodná veličina $\boldsymbol{c_1}'\boldsymbol{X}_p$ měla maximální rozptyl, tj.

$$\mathsf{var}(\boldsymbol{c_1}'\boldsymbol{X}_p) = \max_{\boldsymbol{c},\boldsymbol{c}'\boldsymbol{c}=1} \{\mathsf{var}(\boldsymbol{c}'\boldsymbol{X}_p)\}. \tag{4.1}$$

Hledáme vlastně směr, ve kterém jsou data nejvariabilnější. Podmínku (4.1) splňuje $\boldsymbol{c}_1 = \boldsymbol{\alpha}_1$ (viz např. Anděl, 1985, kap. XVII.2). Dostáváme tak první hlavní komponentu $Z_1 = \boldsymbol{\alpha}_1'\boldsymbol{X}_p$, pro kterou platí

$$\mathsf{var}\, Z_1 = \mathsf{var}(\boldsymbol{\alpha}_1'\boldsymbol{X}_p) = \mathsf{var}(\boldsymbol{\alpha}_1'\boldsymbol{\Sigma}\boldsymbol{\alpha}_1) = \lambda_1.$$

Další komponenty $Z_i, i = 2, 3, \ldots, p$ hledáme postupně jako lineární kombinance $\boldsymbol{c}_i'\boldsymbol{X}_p, \boldsymbol{c}_i'\boldsymbol{c}_i = 1$, tak, aby byly nekorelované s dosud nalezenými komponentami $Z_1, \ldots, Z_{i-1}$ a zároveň aby měli mezi těmito možnými kombinacemi maximální rozptyl, tj.

$$\mathsf{var}(\boldsymbol{c_i}'\boldsymbol{X}_p) = \max_{\boldsymbol{c},\boldsymbol{c}'\boldsymbol{c}=1,\boldsymbol{c}\perp\boldsymbol{c}_1,\ldots,\boldsymbol{c}\perp\boldsymbol{c}_{i-1}} \{\mathsf{var}(\boldsymbol{c}'\boldsymbol{X}_p)\}. \tag{4.2}$$

Tyto podmínky jsou splněny pro $\boldsymbol{c}_i = \boldsymbol{\alpha}_i$ a pro i-tou komponentu $Z_i = \boldsymbol{\alpha}_i' \boldsymbol{X}_p$ platí $\mathsf{var}(Z_i) = \lambda_i$ (viz např. Anděl, 1985, kap. XVII.2).

Označme $\sigma^2 = \sum_{i=1}^{p} \mathsf{var}\, X_i$ celkovou variabilitu obsaženou ve vektoru $\boldsymbol{X}_p$. Protože zároveň platí $\sigma^2 = \sum_{i=1}^{p} \lambda_i$ a $\lambda_i > \lambda_{i+1}, i = 1, 2, \ldots, p-1$, můžeme pro další analýzu dat uvažovat pouze prvních q komponent, které vysvětlují pro nás dostačující variabilitu dat. Většinou volíme q takové, aby

$$\frac{\sum_{i=1}^{l} \lambda_i}{\sigma^2} \geq \alpha,$$

kde α bývá $0,9$, $0,95$ nebo $0,99$.

4.2 Robustní analýza hlavních komponent

Tato kapitola uvádí přehled různých robustních metod pro analýzu hlavních komponent, které byly navrženy v posledních letech, a to i včetně algoritmů pro jejich výpočet.

Vzhledem k tomu, že v praxi často neznáme varianční matici $\boldsymbol{\Sigma}$, je zvykem nahradit ji ve vzorcích a výpočtech výběrovou varianční maticí $\boldsymbol{S}$ spočtenou na základě výběru $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$. Tento postup se uplatňuje i pro PCA. Úlohu (4.1),(4.2) v tomto případě můžeme zapsat ve tvaru

$$s^2(\boldsymbol{c_1}'\boldsymbol{X}) = \max_{\boldsymbol{c},\boldsymbol{c}'\boldsymbol{c}=1} \{s^2(\boldsymbol{c}'\boldsymbol{x}_1, \boldsymbol{c}'\boldsymbol{x}_2, \ldots, \boldsymbol{c}'\boldsymbol{x}_n)\}, \tag{4.3}$$

$$s^2(\boldsymbol{c_i}'\boldsymbol{X}) = \max_{\boldsymbol{c},\boldsymbol{c}'\boldsymbol{c}=1,\boldsymbol{c}\perp\boldsymbol{c}_1,\ldots,\boldsymbol{c}\perp\boldsymbol{c}_{i-1}} \{s^2(\boldsymbol{c}'\boldsymbol{x}_1, \boldsymbol{c}'\boldsymbol{x}_2, \ldots, \boldsymbol{c}'\boldsymbol{x}_n)\}, \tag{4.4}$$

kde s^2 značí výběrový rozptyl.

Výběrová varianční matice je však citlivá vůči odlehlým pozorováním a spolu s ní tedy i celá analýza hlavních komponent. V zásadě se nabízí několik možností, jak ji robustifikovat. Místo odhadu $\boldsymbol{S}$ můžeme použít některý z robustních odhadů varianční matice z kapitoly 2.1. Touto metodou se zabývali např. Croux a Haesbroeck (2000). Pokud tedy takový robustní odhad $\boldsymbol{C} \in \mathrm{SPD}(p)$ máme, nalezneme vlastní čísla této matice $\lambda_1(\boldsymbol{C}) \geq \lambda_2(\boldsymbol{C}) \geq \ldots \geq \lambda_p(\boldsymbol{C}) > 0$ a jim odpovídající vlastní vektory $\boldsymbol{c}_1(\boldsymbol{C}), \boldsymbol{c}_2(\boldsymbol{C}), \ldots, \boldsymbol{c}_p(\boldsymbol{C})$. Tak nalezneme směry, do kterých budeme data promítat. Pokud budeme používat tuto metodu, označíme ji PCA se zkratkou příslušného odhadu (např. PCA-MVE při použití $\boldsymbol{C}_{\mathrm{MVE}}$).

Tento postup je však nevhodný pro velká p. V takovém případě je lepší nahradit rozptyl v (4.3) a (4.4) nějakým robustním jednorozměrným odhadem rozptylu, označme ho $S(Y)$, kde Y značí výběr z jednorozměrného rozdělení. Takovým odhadem může být například $\mathrm{MAD}(Y)$. Řešíme tedy úlohu

$$S(\boldsymbol{c_1}'\boldsymbol{X}) = \max_{\boldsymbol{c},\boldsymbol{c}'\boldsymbol{c}=1} \{S(\boldsymbol{c}'\boldsymbol{x}_1, \boldsymbol{c}'\boldsymbol{x}_2, \ldots, \boldsymbol{c}'\boldsymbol{x}_n)\}, \tag{4.5}$$

$$S(\boldsymbol{c_i}'\boldsymbol{X}) = \max_{\boldsymbol{c},\boldsymbol{c}'\boldsymbol{c}=1,\boldsymbol{c}\perp\boldsymbol{c}_1,\ldots,\boldsymbol{c}\perp\boldsymbol{c}_{i-1}} \{S(\boldsymbol{c}'\boldsymbol{x}_1, \boldsymbol{c}'\boldsymbol{x}_2, \ldots, \boldsymbol{c}'\boldsymbol{x}_n)\}. \tag{4.6}$$

Výhodou této metody (budeme ji značit PP podle anglického termínu *Projection Pursuit*) je to, že nemusíme počítat odhad celé varianční matice, ale jenom prvních několik komponent, které pro další analýzu potřebujeme. Narozdíl od (4.3), (4.4), kde řešení nalezneme jednoduše pomocí vlastních vektorů matice $\boldsymbol{S}$,

je nenalezení řešení (4.5), (4.6) náročná optimalizační úloha. Croux a Ruiz-Gazen (2005) sestavili algoritmus, který tuto úlohu řeší. Budeme ho nazývat PROJ algoritmus a analýzu hlavních komponent provedenou pomocí tohoto algoritmu označíme PCA-PROJ.

K provedení PROJ algoritmu potřebujeme také odhad střední hodnoty - budeme uvažovat nějaký robustní odhad $\boldsymbol{T}(\boldsymbol{X})$. Označme q počet komponent, které chceme spočítat. Pak odhad k-tého vlastního vektoru $\boldsymbol{c}_k(\boldsymbol{X}), k = 1, 2, \ldots, q$, spočteme následujícím způsobem.

1. Pro $k = 1$ položme $\boldsymbol{x}_i^k = \boldsymbol{x}_i - \boldsymbol{T}(\boldsymbol{X}), i = 1, 2, \ldots, n$, pro $k = 2, \ldots, q$, položme $\boldsymbol{x}_i^k = \boldsymbol{x}_i^{k-1} - \boldsymbol{y}_i^{k-1}\boldsymbol{c}_{k-1}(\boldsymbol{X}), i = 1, 2, \ldots, n$.

2. Sestavíme množinu $A^k = \{\boldsymbol{x}_i^k / \|\boldsymbol{x}_i^k\|, i = 1, 2, \ldots, n\}$.

3. Za odhad $\boldsymbol{c}_k(\boldsymbol{X})$ vezmeme

$$\boldsymbol{c}_k(\boldsymbol{X}) = \arg\max_{\boldsymbol{c} \in A^k} S(\boldsymbol{c}'\boldsymbol{x}_1^k, \boldsymbol{c}'\boldsymbol{x}_2^k, \ldots, \boldsymbol{c}'\boldsymbol{x}_n^k),$$

   tj. to $\boldsymbol{c} \in A^k$, které maximalizuje $S(\boldsymbol{c}'\boldsymbol{x}_1^k, \boldsymbol{c}'\boldsymbol{x}_2^k, \ldots, \boldsymbol{c}'\boldsymbol{x}_n^k)$. Navíc sestrojíme projekce do tohoto směru $y_i^k = \boldsymbol{c}_k{}'\boldsymbol{x}_i^k$.

Odhad vlastních hodnot pak získáme jako

$$\lambda_k(\boldsymbol{X}) = S(\boldsymbol{c}_k(\boldsymbol{X})'\boldsymbol{x}_1, \boldsymbol{c}_k(\boldsymbol{X})'\boldsymbol{x}_2, \ldots, \boldsymbol{c}_k(\boldsymbol{X})'\boldsymbol{x}_n).$$

Pro $p = q$ můžeme navíc sestrojit i odhad varianční matice

$$\sum_{k=1}^{p} \lambda_k(\boldsymbol{X})\boldsymbol{c}_k(\boldsymbol{X})\boldsymbol{c}_k(\boldsymbol{X})'.$$

Jako odhad $\boldsymbol{T}(\boldsymbol{X})$ volí autoři L_1-odhad definivaný jako

$$\boldsymbol{T}_{L_1}(\boldsymbol{X}) = \arg\min_{\boldsymbol{T} \in \mathbb{R}^p} \sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{T}\|. \tag{4.7}$$

PROJ algoritmus nemusí dobře fungovat v situaci, kdy je rozsah výběru n relativně malý vzhledem k p. Croux, Filzmoser a Oliveria (2007) ukázali, že pro každé $k > n/2$ bude $\lambda_k(\boldsymbol{X}) = 0$ nezávisle na typu dat a volbě robustního odhadu $S(Y)$ s vysokým bodem selhání. Tento nedostatek odstranili v jejich GRID algoritmu, který odlišným způsobem hledá směry, přes které maximalizovat. Stejně jako PCA-PROJ zavedeme zkratku PCA-GRID pro analýzu hlavních komponent provedenou pomocí GRID algoritmu. Tento algoritmus nyní popíšeme.

Nejprve přeindexujeme řádky $\boldsymbol{X}$ tak, aby pro $S_j = S(x_1^j, x_2^j, \ldots, x_n^j), j = 1, 2, \ldots, p$ platilo $S_1 \geq S_2 \geq \ldots \geq S_p$. Označme $\boldsymbol{e}_i, i = 1, 2, \ldots, p$ bazické vektory, tj. vektory s hodnotou 1 na i-tém místě a 0 jinde. Pak můžeme předpokládat, že platí $S(\boldsymbol{e}_1) \geq S(\boldsymbol{e}_2) \geq \ldots \geq S(\boldsymbol{e}_p)$. První optimální směr nalezneme následujícím způsobem.

1. Označme $\boldsymbol{c}_{1,0} = \boldsymbol{e}_1$.

2. Pro $i = 1, 2, \ldots, N_c$, kde N_c je konstanta, opakujeme následující cyklus pro $j = 1, 2, \ldots, p$. (Autoři používají $N_c = 10$.)

- Nalezneme úhel $\theta_{i,j}$, pro který platí

$$\theta_{i,j} = \arg\max_{\theta \in A_i} S\left((\cos\theta \boldsymbol{c}_{i,j-1} + \sin\theta \boldsymbol{e}_j)' \boldsymbol{X}\right),$$

  kde $A_i = \{\pi/2^{i-1}(1/2 - k/N_g), k = 0, 1, \ldots, N_g - 1\}$ a N_g je konstanta. (Autoři volí $N_g = 10$.)
- zpřesníme odhad vlastního vektoru $\boldsymbol{c}_{i,j} = \cos\theta_{i,j}\boldsymbol{c}_{i,j-1} + \sin\theta_{i,j}\boldsymbol{e}_j$.

Položíme $\boldsymbol{c}_{i+1,0} = \boldsymbol{c}_{i,p}$.

V hlavním cyklu začíne se směrem, ve kterém předpokládáme největší variabilitu. Při každém dalším průchodu cyklem ho dál lokálně zpřesňujeme, tentokrát na polovičním intervalu $[-\pi/2^i, \pi/2^i)$. Ve vnořeném cyklu pak upravujeme j-tou souřadnici $\boldsymbol{c}_{i,j}$. Výsledkem algoritmu je pak nalezený optimální směr $\boldsymbol{c}_1 = \boldsymbol{c}_{N_c,p}$. Jako vedlejší produkt navíc v posledním průchodu oběma cykly dostáváme odhad vlastního čísla λ_1.

Další optimální směry nalezneme obdobným způsobem. Předpokládejme, že jsme již nalezli $(k-1)$-ní optimální směr $\boldsymbol{c}_{k-1}$. Pak na datech provedeme Householderovu transformaci tak, aby se směr $\boldsymbol{c}_{k-1}$ zobrazil na $\boldsymbol{e}_{k-1}$. Opět provedeme GRID algoritmus s tím rozdílem, že ve vnořeném cyklu pracujeme pouze s $j = p - k + 1, \ldots, p$. Jako odhad vlastních vektorů pak vezmeme optimální směry $\boldsymbol{c}_i$ trasformované zpátky do původního prostoru.

Hubert, Rousseeuw a Vanden Branden (2005) využili myšlenky obou metod - projekce i robustního odhadu varianční matice. Jejich ROBPCA algoritmus probíhá ve třech krocích. Jeho detailní popis by byl příliš dlouhý, proto ho zde uvedem pouze ve zkrácené podobě. Podrobný popis lze nalézt v Hubert a kol. (2005).

Nejprve pomocí rozkladu matice $\boldsymbol{X}'$ podle singulárních hodnot transformujeme data do podprostoru generovaného $\boldsymbol{X}$. Tento krok je obzvlášť přínosný pokud $p \geq n$, neboť na transformovaných datech pozorujeme nejvýše $n-1$ veličin. V druhém kroce spočteme pro všechny body transformovaného výběru jejich odlehlost analogickou (2.25) pro SD-odhad. jako robustní odhady μ a σ vezmeme jednorozměrný MCD-odhad střední hodnoty a rozptylu. Maximalizaci v (2.25) provádíme přes všechny směry určené dvojicemi bodů z výběru. Pokud je rozsah výběru n příliš velký, vybereme náhodně jen určitý počet směrů (autoři používají 250). Vybereme h bodů z výběru, které mají nejmenší odlehlost a spočteme jejich výběrovou varianční matici $\boldsymbol{S}_h$. Na základě spektrálního rozkladu $\boldsymbol{S}_h$ určíme počet komponent q, které budeme dále počítat, a vlastní vektory odpovídající q největším vlastním číslům matice $\boldsymbol{S}_h$. Data promítneme do prostoru generovaného těmito vlastními vektory. Ve třetím kroce pak spočteme MCD-odhad transformavaných dat pomocí upraveného FAST-MCD algoritmu. Pomocí spektrálního rozkladu konečného odhadu varianční matice pak získáme její vlastní vektory. Ty pak transformujeme zpět do původního prostoru a získáme tak robustní odhady vlastních vektorů původních dat.

Locantore a kol. (1999) představili další metodu robustifikace PCA, tzv. *sférickou analýzu hlavních komponent* (dále SPCA z anglického termínu *spherical principal component analysis*). Základní myšlenkou SPCA je promítnout data na kulovou plochu s jednotkovým poloměrem a provést klasickou PCA na takto

transformovaných datech. Jako střed koule je vhodné vzít nějaký robustní odhad střední hodnoty, v této práci budeme používat L_1-odhad (4.7).

L_1-odhad lze nalézt jako řešení úlohy

$$\frac{\partial}{\partial \boldsymbol{T}} \sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{T}\| = \sum_{i=1}^{n} \frac{\boldsymbol{x}_i - \boldsymbol{T}}{\|\boldsymbol{x}_i - \boldsymbol{T}\|} = \boldsymbol{0}. \tag{4.8}$$

Označme $\boldsymbol{x}_i^{\boldsymbol{T}}$ projekci bodu $\boldsymbol{x}_i$ na sféru se středem v bodě $\boldsymbol{T}$ a poloměrem 1, tj.

$$\boldsymbol{x}_i^{\boldsymbol{T}} = \boldsymbol{T} + \frac{\boldsymbol{x}_i - \boldsymbol{T}}{\|\boldsymbol{x}_i - \boldsymbol{T}\|}. \tag{4.9}$$

Z (4.8) a (4.9) vidíme, že L_1-odhad je zároveň průměrem bodů $\boldsymbol{x}_i^{\boldsymbol{T}_{L_1}}, i = 1, 2, \ldots, n$. L_1-odhahd tedy můžeme nalézt tak, že umístíme střed jednotkové koule do nějakého bodu $\boldsymbol{T}_0$, promítneme data na sféru koule a poté koulí pohybujeme v prostoru tak dlouho, dokud střed koule nesplyne s průměrem promítnutých bodů. Tento střed je potom L_1-odhadem. Locantore a kol. (1999) uvádějí iterační algoritmus pro výpočet $\boldsymbol{T}_{L_1}$ založený na této myšlence. Tento algoritmus zde nyní popíšeme.

Za výchozí odhad $\boldsymbol{T}_0$ vezmeme výběrový průměr $\overline{\boldsymbol{x}}$. Označme $\boldsymbol{T}_{j-1}$ odhad středu získaný v $(j-1)$-ním iteračním kroce. Odhad $\boldsymbol{T}_j$ získáme tak, že se do něj posuneme z bodu $\boldsymbol{T}_{j-1}$ ve směru průměru promítnutých bodů $\sum_{i=1}^{n} \boldsymbol{x}_i^{\boldsymbol{T}_{j-1}}/n$. Abychom se posunuli více, pokud jsou data více rozptýlená, převážíme délku kroku harmonickým průměrem vzdáleností bodů nepromítnutých na sféru od jejího středu $\boldsymbol{T}_{j-1}$. Při značení $w_i = 1/\|\boldsymbol{x}_i - \boldsymbol{T}_{j-1}\|$ tedy máme

$$\begin{aligned} \boldsymbol{T}_j &= \boldsymbol{T}_{j-1} + \frac{\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i^{\boldsymbol{T}_{j-1}} - \boldsymbol{T}_{j-1}}{\frac{1}{n}\sum_{i=1}^{n} w_i} = \boldsymbol{T}_{j-1} + \frac{\sum_{i=1}^{n} w_i(\boldsymbol{x}_i - \boldsymbol{T}_{j-1})}{\sum_{i=1}^{n} w_i} \\ &= \frac{\sum_{i=1}^{n} w_i \boldsymbol{x}_i}{\sum_{i=1}^{n} w_i}. \end{aligned}$$

Jako kritérium pro ukončení iterací zvolili autoři situaci, kdy se $\boldsymbol{T}_{j-1}$ a $\boldsymbol{T}_j$ liší o méně než 10^{-6} nebo pokud byla spočtena dvacátá iterace.

Kapitola 5

Simulace

Pro výpočet robustních odhadů, klasifikační analýzu i analýzu hlavních komponent byla navržena řada algoritmů. Všechny zmiňované v této práci byly souhrně implementovány pro program R v knihovně rrcov. Todorov a Filzmoser (2009) popsali strukturu této knihovny, odhady a metody v ní implementované. Také uvedli řadu příkladů, jak s touto knihovnou pracovat. Výpočty v této a následující kapitole byly provedeny v R právě pomocí této knihovny a pokud nebude uvedeno jinak, používali jsme přednastavené hodnoty parametrů.

Pro výpočet klasické nerobustní QDA (resp. LDA) jsou v knihovně rrcov implementovány třídy QdaClassic (resp. LdaClassic). Robustní QDA je implementována ve třídě QdaCov. Volbou parametru method pak můžeme nastavit typ odhadu, který se pro výpočet QDA použje. Co se týče robustní LDA, jsou v knihovně implementovány pouze varianty pro MCD-odhad. Nalezneme je ve třídě Linda, kde pomocí parametru method můžeme vybrat metodu pro výpočet sdružené varianční matice a typ použitého algoritmu. Ostatní robustní varianty LDA budeme počítat přímo podle (3.2) v pravidle 7. Pro výpočet robustních odhadů $(\boldsymbol{T}, \boldsymbol{C})$ přitom budeme používat třídu CovRobust. Pro konkrétní typ odhadu je vždy definovaná příslušná podtřída - například CovMve pro výpočet $(\boldsymbol{T}_{\mathrm{MVE}}, \boldsymbol{C}_{\mathrm{MVE}})$.

Analýza hlavních komponent je implementována v třídě Pca. Ta zahrnuje podtřídy PcaClassic pro klasickou nerobustní PCA a PcaCov pro PCA založenou na robustním odhadu varianční matice, kde se volba typu odhadu nastaví pomocí parametru cov.control. Dále zahrnuje podtřídy PcaLocantore pro SPCA a podtřídy PcaProj, PcaGrid, PcaHubert pro výpočet PCA pomocí algoritmů PROJ, GRID a ROBPCA.

5.1 Kvadratická klasifikační analýza

V této kapitole se pokusíme na základě simulací porovnat vliv různých odhadů střední hodnoty a varianční matice na kvadratickou klasifikaci. Budeme uvažovat situaci, kdy $p = 6$, klasifikujeme do $k = 3$ skupin a předpokládané rozdělení ve skupinách je $\mathsf{N}(\boldsymbol{\mu}^i, \boldsymbol{\Sigma}^i), i = 1, 2, 3$, kde $\boldsymbol{\mu}^1 = (0, 0, 0, 0, 0, 0)$, $\boldsymbol{\mu}^2 = (1, 1, 1, 0, 0, 0)$, $\boldsymbol{\mu}^3 = (0, 0, 0, 1, 1, 1)$ a

$$\boldsymbol{\Sigma}^1 = \begin{pmatrix} \boldsymbol{P}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{P}_2 \end{pmatrix},$$

kde

$$\boldsymbol{P}_1 = \begin{pmatrix} 1 & 0,95 & 0,3 \\ 0,95 & 1 & 0,1 \\ 0,3 & 0,1 & 1 \end{pmatrix}, \boldsymbol{P}_2 = \begin{pmatrix} 1 & -0,499 & -0,499 \\ -0,499 & 1 & -0,499 \\ -0,499 & -0,499 & 1 \end{pmatrix},$$

$\boldsymbol{\Sigma}^2 = \mathrm{diag}(0,9; 0,75; 0,6; 0,45; 0,3; 0,15)$, $\boldsymbol{\Sigma}^3 = \boldsymbol{I}$.

Budeme porovnávat nerobstní výběrové pravidlo QDA a různé verze robustního kvadratického pravidla. Konkrétně to budou pravidla QDA-MVE, QDA-MCD, QDA-SD, QDA-OGK, QDA-S1 a QDA-S2, kde QDA-S1 používá S-odhad s dvouváhovou funkcí (2.23) a QDA-S2 zase S-odhad s funkcí t-biweight (2.24).

Fungování klasifikačních pravidel budeme porovnávat na základě odhadu jejich chyby klasifikace AER pro různě velké a různě kontaminované výběry. Pro i-tou skupinu sestrojíme náhodný výběr $\boldsymbol{X}^i_{tr}$ o rozsahu n^i z rozdělení

$$(1-\epsilon^i)\,\mathsf{N}(\boldsymbol{\mu}^i, \boldsymbol{\Sigma}^i) + \epsilon^i\,\mathsf{N}(\boldsymbol{\mu}^{i*}, \boldsymbol{\Sigma}^{i*}), \tag{5.1}$$

kde ϵ značí podíl kontaminovaných dat ve výběru. Tento zápis znamená, že vygenerujeme $(1-\epsilon^i)n^i$ bodů z $\mathsf{N}(\boldsymbol{\mu}^i, \boldsymbol{\Sigma}^i)$ a $\epsilon^i n^i$ bodů z $\mathsf{N}(\boldsymbol{\mu}^{i*}, \boldsymbol{\Sigma}^{i*})$. Na základě tréninkového výběru $\boldsymbol{X}_{tr} = (\boldsymbol{X}^1_{tr}, \boldsymbol{X}^2_{tr}, \boldsymbol{X}^3_{tr})$ sestrojíme odhady $\boldsymbol{T}(\boldsymbol{X}_{tr})$ a $\boldsymbol{C}(\boldsymbol{X}_{tr})$. Vygenerujeme testovací výběr $\boldsymbol{X}_{ts} = (\boldsymbol{X}^1_{ts}, \boldsymbol{X}^2_{ts}, \boldsymbol{X}^3_{ts})$, kde $\boldsymbol{X}^i_{ts}$ je náhodný výběr o rozsahu n z nekontaminovaného rozdělení $\mathsf{N}(\boldsymbol{\mu}^i, \boldsymbol{\Sigma}^i)$. Podle příslušného pravidla pak klasifikujeme jednotlivé body z $\boldsymbol{X}_{ts}$ do skupin a odhadneme chybu klasifikace pomocí AER_{ts}. Tento postup zopakujeme J-krát. Označme $AER^j_{ts}, j = 1, 2, \ldots, J$ chybu klasifikace spočtenou v j-tém kroce. Pro porovnání vlivu použitého odhadu na klasifikaci použijeme průměrnou chybu klasifikace a její výběrovou směrodatnou odchylku

$$\overline{AER}_{ts} = \frac{1}{J}\sum_{j=1}^{J} AER^j_{ts} \tag{5.2}$$

$$sd = \sqrt{\frac{1}{J-1}\sum_{j=1}^{J}(AER^j_{ts} - \overline{AER}_{ts})^2}. \tag{5.3}$$

Označme $\boldsymbol{\mu}^4 = (0,0,5,5,0,0)$, $\boldsymbol{\mu}^5 = (0,0,0,0,5,5)$, $\boldsymbol{\mu}^6 = (5,5,0,0,0,0)$ a $\boldsymbol{\Sigma}^4 = \mathrm{diag}(0,5; 0,5; 0,5; 0,5; 0,5; 0,5)$. Budeme zkoumat následující situace.

A. $n^i = 500, \epsilon^i = 0, i = 1,2,3, \boldsymbol{\mu}^{1*} = \boldsymbol{\mu}^4, \boldsymbol{\mu}^{2*} = \boldsymbol{\mu}^5, \boldsymbol{\mu}^{3*} = \boldsymbol{\mu}^6, \boldsymbol{\Sigma}^{i*} = \boldsymbol{\Sigma}^4, i = 1,2,3.$

B. $n^i = 500, \epsilon^i = 1/5, i = 1,2,3, \boldsymbol{\mu}^{1*} = \boldsymbol{\mu}^4, \boldsymbol{\mu}^{2*} = \boldsymbol{\mu}^5, \boldsymbol{\mu}^{3*} = \boldsymbol{\mu}^6, \boldsymbol{\Sigma}^{i*} = \boldsymbol{\Sigma}^4, i = 1,2,3.$

C. $n^1 = 750, n^2 = 500, n^3 = 250, \epsilon^i = 1/5, i = 1,2,3, \boldsymbol{\mu}^{1*} = \boldsymbol{\mu}^4, \boldsymbol{\mu}^{2*} = \boldsymbol{\mu}^5, \boldsymbol{\mu}^{3*} = \boldsymbol{\mu}^6, \boldsymbol{\Sigma}^{i*} = \boldsymbol{\Sigma}^4, i = 1,2,3.$

D. $n^i = 500, i = 1,2,3, \epsilon^1 = 1/5, \epsilon^2 = 1/4, \epsilon^3 = 1/3, \boldsymbol{\mu}^{1*} = \boldsymbol{\mu}^4, \boldsymbol{\mu}^{2*} = \boldsymbol{\mu}^5, \boldsymbol{\mu}^{3*} = \boldsymbol{\mu}^6, \boldsymbol{\Sigma}^{i*} = \boldsymbol{\Sigma}^4, i = 1,2,3.$

E. $n^i = 500, \epsilon^i = 1/5, i = 1,2,3, \boldsymbol{\mu}^{1*} = \boldsymbol{\mu}^1, \boldsymbol{\mu}^{2*} = \boldsymbol{\mu}^2, \boldsymbol{\mu}^{3*} = \boldsymbol{\mu}^3, \boldsymbol{\Sigma}^{i*} = 25\boldsymbol{\Sigma}^i, i = 1,2,3.$

		QDA	QDA -MVE	QDA -MCD	QDA -S1	QDA -S2	QDA -SD	QDA -OGK
A	AER_{ts}	0,0517	0,0524	0,0520	0,0519	0,0522	0,0524	0,0532
	sd	0,0053	0,0055	0,0056	0,0055	0,0055	0,0054	0,0057
B	AER_{ts}	0,2029	0,0543	0,0542	0,0702	0,0534	0,0542	0,0670
	sd	0,0108	0,0055	0,0055	0,0067	0,0055	0,0056	0,0086
C	AER_{ts}	0,2038	0,0541	0,0539	0,0703	0,0532	0,0541	0,0672
	sd	0,0113	0,0060	0,0060	0,0073	0,0059	0,0060	0,0092
D	AER_{ts}	0,2137	0,0715	0,0684	0,0831	0,0701	0,0703	0,0717
	sd	0,0113	0,0071	0,0090	0,0078	0,0071	0,0071	0,0102
E	AER_{ts}	0,2398	0,0545	0,0544	0,0578	0,0537	0,0537	0,0532
	sd	0,0198	0,0059	0,0057	0,0060	0,0056	0,0057	0,0060

Tabulka 5.1: Průměrná chyba kvadratické klasifikace a její směrodatná odchylka v závislosti na použitém klasifikačním pravidle a typu kontaminace vstupních dat.

Situace A nám umožní porovnat chování nerobustního výběrového pravidla a různých robustních pravidel pro nekontaminovaná data. V situacích B a C porovnáváme pravidla pro výběry kontaminované 1/5 odlehlých pozorování. V případě B mají výběry ve skupinách stejný rozsah, v případě C různý. D zachycuje situaci, kdy je podíl kontaminovaných dat ve výběrech různý a v případě E kontaminujeme výběry pouze radiálními odlehlými pozorováními. Jako odhad pravděpodobností $\hat{p}^i, i = 1, 2, 3$ v pravidlech volíme vždy 1/3, rozsahy tréninkových výběrů $n = 500$ a počet opakování $J = 250$.

Výsledky simulací jsou shrnuty v tabulce 5.1. Z těchto výsledků vidíme, že pro nekontaminovaná data (situace A) dávají nerobustní pravidlo i všechna robustní pravidla prakticky stejnou chybu klasifikace (přibližně 5 %). V případě kontaminovaných vstupních dat (situace B-E) se výsledky výrazně liší pro robustní a nerobustní pravidla. Při použití nerobustního výběrového pravidla je chybně klasifikována pětina až čtvrtina pozorování, zatímco pro žádné z robustních pravidel nepřekročí chyba klasifikace 9 %. Mezi robustními pravidly jsou jen malé rozdíly. Pouze QDA-S1 a QDA-OGK dávají v situacích B, C a D o trochu horší výsledky. Ve srovnání s nerobustním výběrovým pravidlem je však tento rozdíl nepatrný. S výjimkou sitauce D (rozdílný podíl kontaminace) jsou navíc chyby klasifikace pro kontaminovaná a nekontaminovaná data velmi blízké.

Z výsledků plyne, že použití robustních odhadů v klasifikaci je nejen žádoucí v případě, kdy analyzujeme kontaminovaná data, ale není ani na závadu, pokud data kontaminovaná nejsou. Jestliže tedy nemáme jistotu, že data neobsahují odlehlá pozorování, je pro sestavení klasifikačního pravidla lepší použít některý z robustních odhadů.

Nyní se zaměříma na to, jak bude klasifikace ovlivněna, pokud na datech provedeme nejprve analýzu hlavních komponent. Porovnáme různé robustní a nerobustní metody PCA v kombinaci s následnou robustní či nerobustní klasifikací. Budeme uvažovat stejné výchozí situace A-E jako v případě simulací bez PCA. Na tréninkovém výběru vygenerovaném podle (5.1) provedeme klasickou nerobustní PCA a její robustní varianty popsané v kapitole 4.2. Pro robustní PCA založenou na robustním odhadu varainční matice budeme zkoumat PCA-MVE,

		A		B		C		E	
		QDA	QDA -S2	QDA	QDA -S2	QDA	QDA -S2	QDA	QDA -S2
PCA	AER_{ts}	0,279	0,280	0,564	0,382	0,594	0,338	0,543	0,407
	sd	0,045	0,045	0,013	0,011	0,014	0,012	0,032	0,014
PCA-MVE	AER_{ts}	0,390	0,392	0,488	0,333	0,537	0,416	0,425	0,331
	sd	0,023	0,024	0,039	0,057	0,018	0,015	0,066	0,049
PCA-MCD	AER_{ts}	0,399	0,401	0,534	0,415	0,536	0,418	0,520	0,401
	sd	0,012	0,012	0,016	0,016	0,020	0,015	0,029	0,012
PCA-S1	AER_{ts}	0,346	0,348	0,533	0,357	0,541	0,422	0,412	0,323
	sd	0,033	0,033	0,031	0,024	0,014	0,015	0,056	0,041
PCA-S2	AER_{ts}	0,344	0,345	0,479	0,316	0,533	0,413	0,401	0,315
	sd	0,036	0,036	0,036	0,046	0,018	0,014	0,062	0,047
PCA-SD	AER_{ts}	0,410	0,412	0,530	0,395	0,525	0,410	0,508	0,399
	sd	0,013	0,013	0,029	0,017	0,015	0,012	0,030	0,013
PCA-OGK	AER_{ts}	0,361	0,363	0,486	0,312	0,526	0,404	0,440	0,350
	sd	0,034	0,034	0,039	0,055	0,015	0,013	0,039	0,033
PCA-PROJ	AER_{ts}	0,297	0,298	0,448	0,234	0,520	0,397	0,407	0,325
	sd	0,047	0,048	0,039	0,067	0,018	0,013	0,044	0,038
PCA-GRID	AER_{ts}	0,290	0,291	0,489	0,312	0,524	0,390	0,394	0,314
	sd	0,044	0,044	0,047	0,068	0,025	0,020	0,049	0,041
ROB PCA	AER_{ts}	0,384	0,385	0,480	0,310	0,533	0,411	0,403	0,311
	sd	0,016	0,016	0,036	0,049	0,017	0,014	0,065	0,048
SPCA	AER_{ts}	0,330	0,331	0,473	0,245	0,556	0,399	0,452	0,372
	sd	0,040	0,040	0,042	0,076	0,016	0,012	0,028	0,025

Tabulka 5.2: Průměrná chyba klasifikace vybraných kvadratických pravidel a její směrodatná odchylka pro různě kontaminovaná vstupní data, na kterých byla nejprve provedena PCA, v závislosti na použité metodě PCA.

PCA-MCD, PCA-SD, PCA-OGK, PCA-S1 a PCA-S2, kde S1 opět značí S-odhad s dvouváhovou funkcí (2.23) a S2 zase S-odhad s funkcí t-biweight (2.24). Pro PP metodu budeme uvažovat obě varianty PCA-PROJ a PCA-GRID. Provedeme také ROBPCA a SPCA.

Po provedení konkrétního typu PCA promítneme vygenerovaný výběr do prostoru určeného prvními dvěma hlavními komponentami. Pro takto transformovaná data sestrojíme odhad střední hodnoty a varianční matice potřebné pro sestavení kvadratických klasifikačních pravidel. Pravidla budou stejného typu jako v případě klasifikace bez předchozího provedení analýzy hlavních komponent, tj. QDA, QDA-MVE, QDA-MCD, QDA-S1, QDA-S2, QDA-SD a QDA-OGK.

Poté vygenerujeme testovací výběr $\boldsymbol{X}_{ts}$ (stejným způsobem jako v předchozích simulacích), také ho promítneme do prostoru určeného prvními dvěma komponentami a aplikujeme na něj klasifikační pravidla.

Celý tento postup zopakujeme J-krát a v každém kroce spočteme chybu klasifikace $AER_{ts}^{j}, j = 1, 2, \ldots, J$ odpovídající použité metodě PCA a jednotlivým klasifikačním pravidlům.Pro porovnání výsledků použijeme opět průměrnou chybu klasifikace (5.2) a její výběrovou směrodatnou odchylku (5.3).

		QDA	QDA -MVE	QDA -MCD	QDA -S1	QDA -S2	QDA -SD	QDA -OGK
PCA	AER_{ts}	0,574	0,386	0,385	0,405	0,384	0,385	0,467
	sd	0,014	0,012	0,012	0,015	0,012	0,012	0,018
PCA-MVE	AER_{ts}	0,542	0,382	0,380	0,398	0,376	0,389	0,463
	sd	0,035	0,051	0,051	0,056	0,051	0,055	0,070
PCA-MCD	AER_{ts}	0,551	0,451	0,450	0,476	0,443	0,458	0,518
	sd	0,019	0,028	0,027	0,023	0,026	0,030	0,027
PCA-S1	AER_{ts}	0,582	0,399	0,397	0,419	0,399	0,419	0,481
	sd	0,022	0,025	0,025	0,029	0,027	0,028	0,050
PCA-S2	AER_{ts}	0,544	0,371	0,369	0,385	0,365	0,378	0,455
	sd	0,031	0,045	0,044	0,048	0,044	0,047	0,066
PCA-SD	AER_{ts}	0,562	0,423	0,423	0,455	0,422	0,435	0,517
	sd	0,019	0,021	0,020	0,025	0,022	0,027	0,030
PCA-OGK	AER_{ts}	0,558	0,367	0,366	0,383	0,366	0,377	0,444
	sd	0,034	0,037	0,037	0,041	0,039	0,041	0,077
PCA-PROJ	AER_{ts}	0,523	0,314	0,314	0,329	0,313	0,321	0,373
	sd	0,044	0,078	0,078	0,089	0,080	0,081	0,116
PCA-GRID	AER_{ts}	0,538	0,357	0,357	0,378	0,357	0,367	0,439
	sd	0,032	0,055	0,055	0,063	0,056	0,057	0,084
ROB PCA	AER_{ts}	0,579	0,410	0,406	0,430	0,409	0,427	0,487
	sd	0,027	0,029	0,029	0,033	0,031	0,029	0,045
SPCA	AER_{ts}	0,548	0,369	0,368	0,389	0,367	0,372	0,460
	sd	0,021	0,024	0,024	0,030	0,026	0,025	0,039

Tabulka 5.3: Průměrná chyba klasifikace jednotlivých kvadratických pravidel a její směrodatná odchylka pro data kontaminovaná způsobem D, na kterých byla nejprve provedena PCA, v závislosti na použité metodě PCA.

Výsledky simulací jsou shrnuty v tabulkách 5.2 a 5.3. S výjimkou kontaminace typu D se průměrné chyby klasifikace pro robustní pravidla prakticky nelišily. Proto jsou v tabulce 5.2 souhrně uvedeny výsledky pro data kontaminovaná způsoby A, B, C a E, přičemž jsou zastoupeny pouze výsledky pro nerobustní QDA a robustní QDA-S2 v kombinaci se všemi použitými metodami PCA. Pro data kontaminovaná zbůsobem D jsou výsledky pro všechny použité kombinace PCA a klasifikačních pravidel uvedeny v tabulce 5.3.

Nejprve se budeme zabývat otázkou, které klasifikační pravidlo je nejlepší používat. Pro nekontaminovaná data (situace A), stejně jako v případě klasifikace bez předchozí PCA, jsou výsledky pro robustní a nerobustní pravidla prakticky stejné. V případě kontaminovaných dat (situace B-E) jsou výsledky při použití nějakého robustního pravidla jednoznačně lepší oproti nerobustnímu pravidlu - rozdíl v špatně klasifikovaných pozorováních je často větší než 10 %. V situacích A,B,C a E není příliš podstatné, které z robustních pravidel použijeme. Jinak tomu je v situaci D, kdy uvažujeme skupiny s různým podílem kontaminace. V tomto případě dávají nejlepší výsledky QDA-MVE, QDA-MCD a QDA-S2. O něco horší chyby klasifikace dosahují QDA-S1 a QDA-SD. Nejhorší výsledky mezi robustními pravidly pak vykazuje QDA-OGK. I tak ale použití QDA-OGK

vede k výrazně nižší průměrné chybě klasifikace než použití nerobustního výběrového pravidla.

Pokud tedy nejprve snižujeme dimenzionalitu dat pomocí analýzy hlavních komponent (ať už robustní či nerobustní), zdá se rozumné použít k následné kvadratické klasifikaci jedno z pravidel QDA-MVE, QDA-MCD nebo QDA-S2. V každém případě bychom měli počítat s tím, že provedení analýzy hlavních komponent klasifikaci negativně ovlivní. Při robustní klasifikaci na původních datech nepřekročila chyba klasifikace 9 % (srovnej tab. 5.1). Oproti tomu po provedení PCA jsme se naopak nedostali pod dolní hranici 23 %. Tento výsledek však není tak překvapivý, neboť při provedení PCA se vzdáváme části informace, která je v datech obsažena.

Nyní se pokusíme posoudit, jak klasifikaci ovlivní různé verze PCA. Budeme uvažovat pouze výsledky, které jsme získali pro robustní klasifikační pravidlo QDA-S2. Pro nekontaminovaná data (situace A) dává mezi všemi metodami nejmenší chybu klasifikace (27, 9 %) klasická nerobustní PCA. Mezi robustními metodami se jí blíží PCA-GRID s 29,1 % a PCA-PROJ s 29,8 % špatně klasifikovaných pozorování. Ostatní robustní metody dávají ještě vyšší chybu klasifikace - v rozmezí mezi 33,1 % a 40,1 %.

Pro výběry stejného rozsahu kontaminované z 1/5 odlehlými pozorováními (situace B) dávjí výrazně nejlepší výsledky PCA-PROJ (23,4 %) a SPCA (24,5 %), relativně dobré výsledky dávají také ROBPCA, PCA-OGK, PCA-GRID a PCA-S2 (31 % - 31,6 %). Pro výběry s nestejným rozsahem a stejným podílem kontaminace (situace C) má nejmenší chybu klasifikace (33,8 %) klasická nerobustní PCA, dobré výsledky dávají také PCA-GRID, PCA-PROJ, SPCA a PCA-OGK (39 % - 40,4 %). Pro různě kontaminovaná data (situace D) se jako nejlepší jeví PCA-PROJ s chybou klasifikace 31,4 %, dále pak PCA-GRID, SPCA a PCA-OGK s chybou klasifikace 36,1 % - 36,8 %. ROBPCA, PCA-GRID a PCA-S2 pak dosahují nejnižší chyby klasifikace (31,1 % a 31,5 %) pro data kontaminovaná radiálními odlehlými pozorováními (situace E). V tomoto případě dosahují dobré chyby klasifikace také PCA-S1, PCA-S2, PCA-PROJ a PCA-MVE (31,6 % - 32,7 %). Proznamenejme, že ve většině případů dávají PCA-MCD a PCA-SD oproti ostatním robustním metodám PCA horší výsledky.

V případě, že data odlehlá pozorování neobsahují, je rozumné držet se klasické nerobustní PCA. Jestliže však nemáme informaci o tom, zda jsou data kontaminovaná, případně jakým způsobem, jeví se jako rozumný přístup k PCA (pro účely následné klasifikace) použít metodu PP (ať už algoritmus PROJ nebo GRID) nebo sférickou analýzu hlavních komponent. Následnou klasifikaci je pak nejlepší provést pomocí klasifikačních pravidel odvozených na základě MVE-odhadu, MCD-odhadu nebo S-odhadu s použitím funkce t-biweight (2.24).

5.2 Lineární klasifikační analýza

Nyní porovnáme vliv různých odhadů střední hodnoty a varianční matice na lineární klasifikaci. Stejně jako v kapitole 5.1 budeme uvažovat klasifikaci do $k = 3$ skupin a $p = 6$. Budeme předpokládat, že rozdělení ve skupině $\pi^i, i = 1, 2, 3$ je normální s parametry $\mathsf{N}(\boldsymbol{\mu}^i, \boldsymbol{\Sigma})$, kde $\boldsymbol{\mu}^1 = (0, 0, 0, 0, 0, 0)$, $\boldsymbol{\mu}^2 = (1, 1, 1, 0, 0, 0)$,

$\boldsymbol{\mu}^3 = (0,0,0,1,1,1)$ a

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{P}_1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{P}_2 \end{pmatrix},$$

kde

$$\boldsymbol{P}_1 = \begin{pmatrix} 1 & 0,95 & 0,3 \\ 0,95 & 1 & 0,1 \\ 0,3 & 0,1 & 1 \end{pmatrix}, \boldsymbol{P}_2 = \begin{pmatrix} 1 & -0,499 & -0,499 \\ -0,499 & 1 & -0,499 \\ -0,499 & -0,499 & 1 \end{pmatrix}.$$

Porovnáme chování nerobustního výběrového pravidla LDA a robustních klasifikačních pravidel sestrojených na základě stejných odhadů jako v případě simulací pro kvadratickou klasifikaci. Konkrétně to tedy budou pravidla LDA-MVE, LDA-MCD, LDA-S1, LDA-S2, LDA-SD a LDA-OGK, kde LDA-S1 (resp. LDA-S2) opět značí konstrukci pravidla na základě S-odhadu s dvouváhovou funkcí (2.23) (resp. t-biweight (2.24)). Při sestavování robustních klasifikačních pravidel budeme pro sestrojení odhadu sdružené varianční matice používat metodu B popsanou v kapitole 3.2.

Fungování jednotlivých pravidel pro různě velké a různě kontaminované výběry budeme porovnávat stejně jako v případě kvadratické klasifikace. Pro každou skupinu vždy vygenerujeme náhodný výběr $\boldsymbol{X}_{tr}^i$ o rozsahu n^i z kontaminovaného rozdělení

$$(1-\epsilon^i)\,\mathsf{N}(\boldsymbol{\mu}^i, \boldsymbol{\Sigma}) + \epsilon^i\,\mathsf{N}(\boldsymbol{\mu}^{i*}, \boldsymbol{\Sigma}^*). \tag{5.4}$$

Na základě tréninkového výběru $\boldsymbol{X}_{tr} = (\boldsymbol{X}_{tr}^1, \boldsymbol{X}_{tr}^2, \boldsymbol{X}_{tr}^3)$ pak sestrojíme klasifikační pravidla. Ta aplikujeme na testovací výběr $\boldsymbol{X}_{ts} = (\boldsymbol{X}_{ts}^1, \boldsymbol{X}_{ts}^2, \boldsymbol{X}_{ts}^3)$, kde $\boldsymbol{X}_{ts}^i$ je vygenerovaný výběr o rozsahu n z nekontaminovaného rozdělení $\mathsf{N}(\boldsymbol{\mu}^i, \boldsymbol{\Sigma})$. Tento postup zopakujeme J-krát, přičemž v každém kroce $j = 1, 2, \ldots, J$ spočítáme odhad chyby klasifikace AER_{ts}^j příslušející danému klasifikačnímu pravidlu. Kvalitu jednotlivých pravidel budeme opět porovnávat na základě průměrné chyby klasifikace $\overline{AER}_{ts}$ (5.2) a její výběrové směrodatné odchylky sd (5.3). Budeme také uvažovat stejné typy kontaminace jako v případě kvadratické klasifikace, tj. nejprve nekontaminovaná data, poté data se stejným podílem kontaminace a stejnými nebo rozdílnými rozsahy výběrů, data s různým podílem kontaminace a data kontaminovaná pouze radiálními odlehlými pozorováními.

Označme $\boldsymbol{\mu}^4 = (0,0,5,5,0,0)$, $\boldsymbol{\mu}^5 = (0,0,0,0,5,5)$, $\boldsymbol{\mu}^6 = (5,5,0,0,0,0)$ a $\boldsymbol{\Sigma}^2 = \mathrm{diag}(0,5; 0,5; 0,5; 0,5; 0,5; 0,5)$. V konkrétních simulacích byly rozsahy výběrů a parametry v (5.4) voleny následujícím způsobem.

A. $n^i = 500, \epsilon^i = 0, i = 1,2,3, \boldsymbol{\mu}^{1*} = \boldsymbol{\mu}^4, \boldsymbol{\mu}^{2*} = \boldsymbol{\mu}^5, \boldsymbol{\mu}^{3*} = \boldsymbol{\mu}^6, \boldsymbol{\Sigma}^* = \boldsymbol{\Sigma}^2$.

B. $n^i = 500, \epsilon^i = 1/5, i = 1,2,3, \boldsymbol{\mu}^{1*} = \boldsymbol{\mu}^4, \boldsymbol{\mu}^{2*} = \boldsymbol{\mu}^5, \boldsymbol{\mu}^{3*} = \boldsymbol{\mu}^6, \boldsymbol{\Sigma}^* = \boldsymbol{\Sigma}^2$.

C. $n^1 = 750, n^2 = 500, n^3 = 250, \epsilon_i = 1/5, i = 1,2,3, \boldsymbol{\mu}^{1*} = \boldsymbol{\mu}^4, \boldsymbol{\mu}^{2*} = \boldsymbol{\mu}^5, \boldsymbol{\mu}^{3*} = \boldsymbol{\mu}^6, \boldsymbol{\Sigma}^* = \boldsymbol{\Sigma}^2$.

D. $n^i = 500, i = 1,2,3, \epsilon_1 = 1/5, \epsilon_2 = 1/4, \epsilon_3 = 1/3, \boldsymbol{\mu}^{1*} = \boldsymbol{\mu}^4, \boldsymbol{\mu}^{2*} = \boldsymbol{\mu}^5, \boldsymbol{\mu}^{3*} = \boldsymbol{\mu}^6, \boldsymbol{\Sigma}^* = \boldsymbol{\Sigma}^2$.

E. $n^i = 500, \epsilon^i = 1/5, i = 1,2,3, \boldsymbol{\mu}^{1*} = \boldsymbol{\mu}^1, \boldsymbol{\mu}^{2*} = \boldsymbol{\mu}^2, \boldsymbol{\mu}^{3*} = \boldsymbol{\mu}^3, \boldsymbol{\Sigma}^* = 25\boldsymbol{\Sigma}$.

		LDA	LDA -MVE	LDA -MCD	LDA -S1	LDA -S2	LDA -SD	LDA -OGK
A	AER_{ts}	0,1559	0,1563	0,1560	0,1561	0,1561	0,1561	0,1562
	sd	0,0090	0,0091	0,0090	0,0090	0,0091	0,0090	0,0090
B	AER_{ts}	0,4371	0,1563	0,1562	0,1561	0,1562	0,1564	0,1615
	sd	0,0145	0,0090	0,0092	0,0093	0,0093	0,0093	0,0119
C	AER_{ts}	0,4408	0,1552	0,1553	0,1551	0,1553	0,1554	0,1615
	sd	0,0145	0,0089	0,0089	0,0090	0,0089	0,0090	0,0128
D	AER_{ts}	0,4846	0,1559	0,1561	0,1562	0,1561	0,1563	0,1780
	sd	0,0136	0,0092	0,0093	0,0092	0,0092	0,0092	0,0122
E	AER_{ts}	0,1600	0,1558	0,1559	0,1558	0,1559	0,1559	0,1560
	sd	0,0089	0,0079	0,0078	0,0079	0,0078	0,0079	0,0082

Tabulka 5.4: Průměrná chyba lineární klasifikace a její směrodatná odchylka v závislosti na použitém klasifikačním pravidle a typu kontaminace vstupních dat.

Jako odhad apriorních pravděpodobností v pravidlech volíme $\hat{p}^i = 1/3, i = 1, 2, 3$, rozsahy testovacích výběrů $n = 500$ a počet opakování $J = 250$.

Výsledky simulací jsou shrnuty v tabulce 5.4. Z tabulky je vidět, že stejně jako v případě kvadratické klasifikace (srovnej tab. 5.1) dávají všechna pravidla pro nekontaminovaná data (situace A) vesměs stejnou chybu klasifikace (asi 16 %). Stejný výsledek dostaneme i pro kantaminaci dat radiálními odlehlými pozorováními (situace E). Pro ostatní typy kontaminace (situace B-D) se výsledky pro nerobustní a robustní pravidla výrazně liší. Zatímco pro robsutní klasifikační pravidla zůstává průměrná chyba klasifikace stále kolem 16 %, pro nerobustní výběrové pravidlo neklesne pod 43 %. LDA-OGK dává v situacích B-D o trochu horší výsledky než ostatní robustní pravidla. Tento rozdíl však oproti nerobustnímu LDA není nijak výrazný.

Vidíme tedy, že použití robustních pravidel výsledky lineární klasifikace nezhorší, pokud data neobsahují odlehlá pozorování. Pokud však data odlehlá pozorování obsahují, je vhodné použít některou z robustních verzí LDA-MVE, LDA-MCD, LDA-S1, LDA-S2, nebo LDA-SD.

Nyní budeme zkoumat, jak lineární klasifikaci ovlivní předchozí provedení analýzy hlavních komponent. Budeme porovnávat stejné metody jako v případě kvadratické klasifikace, tj. PCA-MVE, PCA-MCD, PCA-S1, PCA-S2, PCA-SD, PCA-OGK, PCA-PROJ, PCA-GRID, ROBPCA a SPCA. Postupovat budeme také obdobně, tj. v každém z $j = 1, 2, \ldots, J$, kroků vygenerujeme tréninkový výběr $\boldsymbol{X}_{tr}$ podle jednoho ze schémat A-E. Na výběru provedeme analýzu hlavních komponent, data promítneme do prostoru určeného prvními dvěma komponentami a sestavíme příslušná lineární klasifikační pravidla. Z nekontaminovaného rozdělení vygenerujeme testovací výběr $\boldsymbol{X}_{ts}$, také ho promítneme do prostoru prvních dvou komponent a jednotlivá pozorování transformovaného výběru klasifikujeme do skupin podle sestrojených pravidel.

Srovnání provedeme opět pomocí průměrné chyby klasifikace $\overline{AER}_{ts}$ (5.2) a její výběrové směrodatné odchylky sd (5.3) odpovídající vždy konkrétní kombinaci metody PCA a vybraného klasifikačního pravidla. Stejně jako v případě klasifikace bez předchozí PCA volíme $\hat{p}^i = 1/3, i = 1, 2, 3$, $n = 500$ a $J = 250$.

		A		B			
		LDA	LDA -S2	LDA	LDA -MCD	LDA -S2	LDA -OGK
PCA	AER_{ts}	0,491	0,527	0,761	0,467	0,545	0,539
	sd	0,013	0,012	0,013	0,013	0,015	0,019
PCA-MVE	AER_{ts}	0,515	0,522	0,626	0,535	0,541	0,550
	sd	0,012	0,013	0,022	0,016	0,018	0,017
PCA-MCD	AER_{ts}	0,518	0,523	0,621	0,532	0,541	0,548
	sd	0,012	0,013	0,023	0,016	0,017	0,017
PCA-S1	AER_{ts}	0,517	0,522	0,639	0,513	0,538	0,553
	sd	0,012	0,012	0,027	0,016	0,017	0,018
PCA-S2	AER_{ts}	0,517	0,524	0,624	0,535	0,542	0,550
	sd	0,012	0,012	0,021	0,016	0,017	0,016
PCA-SD	AER_{ts}	0,518	0,524	0,626	0,530	0,537	0,552
	sd	0,013	0,012	0,021	0,016	0,019	0,017
PCA-OGK	AER_{ts}	0,486	0,528	0,632	0,498	0,528	0,557
	sd	0,013	0,011	0,019	0,015	0,015	0,017
PCA-PROJ	AER_{ts}	0,480	0,527	0,620	0,473	0,521	0,540
	sd	0,015	0,014	0,036	0,017	0,016	0,030
PCA-GRID	AER_{ts}	0,481	0,525	0,615	0,476	0,511	0,540
	sd	0,020	0,014	0,051	0,036	0,035	0,042
ROB PCA	AER_{ts}	0,489	0,528	0,625	0,506	0,532	0,547
	sd	0,013	0,011	0,025	0,017	0,018	0,020
SPCA	AER_{ts}	0,482	0,528	0,684	0,473	0,527	0,530
	sd	0,014	0,012	0,028	0,014	0,016	0,029

Tabulka 5.5: Průměrná chyba klasifikace vybraných linerních klasifikačních pravidel a její směrodatná odchylka pro nekontaminovaná data a data kontaminovaná způsobem B, na kterých byla nejprve provedena PCA, v závislosti na použité metodě PCA.

Výsledky simulací jsou shrnuty v tabulkách 5.5, 5.6 a 5.7. Tabulky vždy obsahují výsledky pro všechny metody PCA. S výjimkou situace D (tabulka 5.7) se pro některá pravidla jejich průměrné chyby klasifikace lišily jen nepatrně, a proto nejsou v tabulkách 5.5 a 5.6 uvedeny kompletní výsledky.

Pro nekontaminovaná data (situace A, viz tabulka 5.5) jsou zastoupeny výsledky pouze pro nerobustní výběrové pravidlo a pro robustní pravidlo LDA-S2, neboť LDA-MCD dávala téměř shodné výsledky jako nerobustní LDA a výsledky pro ostatní robustní pravidla se téměř nelišily od výsledků pro LDA-S2. Z tabulky vidíme, že průměrná chyba klasifikace se vždy pohybovala v rozmezí od 48 % do 53 %. Vidíme tedy, že provedení analýzy hlavních komponent snížilo schopnost pravidla správně klasifikovat pozorování o víc jak 20 % (srovnej s tabulkou 5.4). Nezávisle na volbě metody PCA dávaly o něco lepší výsledky nerobustní výběrové pravidlo a robustní LDA-MCD. Tato klasifikační pravidla pak dávala o něco lepší výsledky v kombinaci s nerobustní PCA nebo s robustními verzemi PCA, které nejsou založeny na robustním odhadu varianční matice. Na ostatní robustní pravidla neměl výběr metody PCA zásadní vliv.

		C				E	
		LDA	LDA -MCD	LDA -S2	LDA -OGK	LDA	LDA -S2
PCA	AER_{ts}	0,826	0,428	0,539	0,512	0,520	0,529
	sd	0,022	0,016	0,019	0,043	0,013	0,013
PCA-MVE	AER_{ts}	0,604	0,533	0,558	0,563	0,516	0,523
	sd	0,025	0,016	0,022	0,017	0,011	0,013
PCA-MCD	AER_{ts}	0,604	0,533	0,557	0,563	0,518	0,523
	sd	0,023	0,016	0,023	0,018	0,013	0,012
PCA-S1	AER_{ts}	0,606	0,533	0,558	0,563	0,493	0,528
	sd	0,023	0,015	0,023	0,017	0,016	0,012
PCA-S2	AER_{ts}	0,606	0,534	0,559	0,565	0,519	0,523
	sd	0,022	0,015	0,025	0,018	0,011	0,012
PCA-SD	AER_{ts}	0,606	0,528	0,549	0,560	0,519	0,524
	sd	0,022	0,015	0,021	0,016	0,019	0,017
PCA-OGK	AER_{ts}	0,608	0,522	0,544	0,557	0,493	0,527
	sd	0,018	0,014	0,019	0,014	0,015	0,011
PCA-PROJ	AER_{ts}	0,560	0,496	0,532	0,564	0,494	0,528
	sd	0,030	0,017	0,014	0,013	0,019	0,013
PCA-GRID	AER_{ts}	0,560	0,496	0,532	0,564	0,493	0,528
	sd	0,030	0,017	0,014	0,013	0,027	0,020
ROB PCA	AER_{ts}	0,606	0,522	0,557	0,563	0,496	0,528
	sd	0,028	0,016	0,022	0,018	0,014	0,012
SPCA	AER_{ts}	0,638	0,538	0,560	0,473	0,499	0,527
	sd	0,040	0,014	0,017	0,014	0,015	0,012

Tabulka 5.6: Průměrná chyba klasifikace vybraných lineárních klasifikačních pravidel a její směrodatná odchylka pro data kontaminovaná způsobem C a E, na kterých byla nejprve provedena PCA, v závislosti na použité metodě PCA.

Velmi podobné výsledky dostáváme i pro data kontaminovaná radiálními odlehlými pozorováními (situace E, tabulka 5.5). Vidíme tedy, že tento typ kontaminace lineární klasifikaci nijak zásadně neovlivní. Uvedeny jsou opět pouze výsledky pro nerobustní výběrové pravidlo, které byly velmi podobné výsledkům pro LDA-MCD, a výsledky pro LDA-S2, které se výrazně nelišily od výsledků pro zbylá robustní pravidla.

Pro data kontaminovaná stejným podílem odlehlých pozorování neuvádíme výsledky pro LDA-MVE, LDA-S1 a LDA-SD, neboť byly velmi podobné výsledkům pro LDA-S2. To se týká jak situace B, kdy jsou rozsahy výběrů stejné (viz tab. 5.5), tak situace C, kdy jsou výběry různě velké (viz tab. 5.6). V obou případech dávají robustní klasifikační pravidla lepší výsledky než nerobustní LDA. Obzvlášť dobrých výsledků dosahuje LDA-MCD. Při jeho použití se dokonce průměrná chyba klasifikace blíží hodnotám pro nekontaminovaná data. Před provedením robustní LDA se jeví vhodné použít klasickou nerobustní PCA, případně PCA-PROJ nebo PCA-GRID.

V případě kontaminace dat různým podílem odlehlých pozorování (situace D, tab. 5.7) byly výsledky o něco různorodější. Opět dosahuje nejhorší cyhby

		LDA	LDA -MVE	LDA -MCD	LDA -S1	LDA -S2	LDA -SD	LDA -OGK
	AER_{ts}	0,747	0,528	0,455	0,528	0,527	0,518	0,630
PCA	*sd*	0,011	0,022	0,014	0,017	0,017	0,016	0,061
PCA-	AER_{ts}	0,663	0,564	0,542	0,555	0,550	0,565	0,566
MVE	*sd*	0,018	0,039	0,015	0,020	0,016	0,020	0,021
PCA-	AER_{ts}	0,665	0,572	0,538	0,564	0,552	0,565	0,576
MCD	*sd*	0,017	0,044	0,017	0,032	0,024	0,020	0,026
PCA-	AER_{ts}	0,683	0,565	0,529	0,558	0,552	0,571	0,573
S1	*sd*	0,022	0,039	0,017	0,020	0,019	0,026	0,025
PCA-	AER_{ts}	0,668	0,571	0,539	0,556	0,552	0,565	0,568
S2	*sd*	0,018	0,043	0,014	0,018	0,016	0,019	0,019
PCA-	AER_{ts}	0,672	0,586	0,535	0,573	0,557	0,573	0,584
SD	*sd*	0,018	0,049	0,016	0,037	0,031	0,019	0,021
PCA-	AER_{ts}	0,691	0,586	0,512	0,590	0,549	0,577	0,602
OGK	*sd*	0,016	0,054	0,017	0,050	0,034	0,022	0,022
PCA-	AER_{ts}	0,722	0,545	0,469	0,552	0,548	0,519	0,710
PROJ	*sd*	0,016	0,030	0,018	0,028	0,042	0,027	0,017
PCA-	AER_{ts}	0,717	0,543	0,465	0,556	0,548	0,527	0,686
GRID	*sd*	0,028	0,065	0,057	0,078	0,079	0,067	0,089
ROB	AER_{ts}	0,681	0,570	0,526	0,557	0,549	0,566	0,574
PCA	*sd*	0,019	0,046	0,018	0,024	0,018	0,021	0,025
	AER_{ts}	0,724	0,532	0,465	0,534	0,519	0,510	0,635
SPCA	*sd*	0,011	0,025	0,013	0,020	0,016	0,017	0,038

Tabulka 5.7: Průměrná chyba klasifikace lineárních klasifikačních pravidel a její směrodatná odchylka pro data kontaminovaná způsobem D, na kterých byla nejprve prvedena PCA, v závislosti na použité metodě PCA.

klasifikace nerobustní výběrové pravidlo (kolem 70 %). Těmto hodnotám se blíží i LDA-OGK, pokud byla na datech nejprve provedena PCA-PROJ nebo PCA-GRID. Výrazně nejlepších výsledků dosáhneme opět použitím klasifikačního pravidla LDA-MCD. V kombinaci s nerobustní PCA, PCA-PROJ nebo PCA-GRID je chyba klasifikace asi 46 %.

Výsledky simulací naznačují, že pro lineární klasifikaci je nejvhodnější použít robustní klasifikaci založenou na MCD-odhadu. Pokud před klasifikací provádíme analýzu hlavních komponent, je vhodné použít PCA-PROJ nebo PCA-GRID.

Kapitola 6

Příklady s reálnými daty

Nyní porovnáme chování jednotlivých kvadratických klasifikačních pravidel a různých metod PCA na dvou souborech reálných dat. Budeme uvažovat stejné varianty QDA, PCA i jejich kombinace jako v kapitole 5.1.

6.1 Chemicko-fyzikální vlastnosti vína

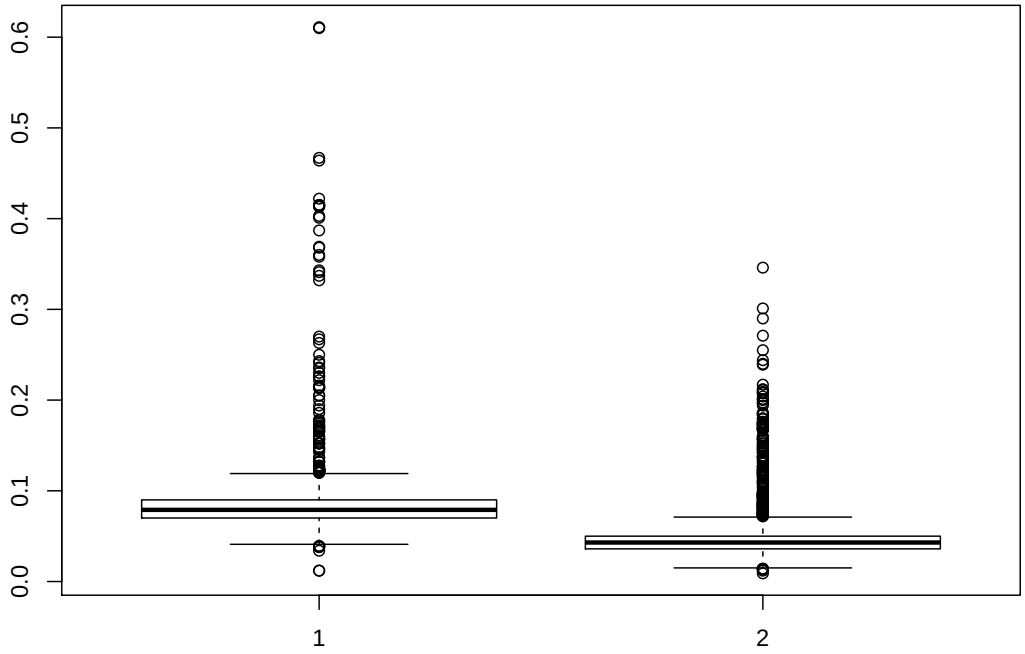
Nejprve se budeme zabývat rozsáhlejším souborem dat, který obsahuje údaje o chemicko-fyzikálních vlastnostech červené a bílé odrůdy portugalského vína *Vinho Verde*. Data ve svém článku použili Cortez a kol. (2009). Pro výzkumné účely je na internetu zpřístupnili Frank a Asuncion (2010) a označili je jako *Wine Quality*.

Data zahrnují naměřené hodnoty 11 veličin jako například pH a hustota, přičemž 1599 pozorování náleží červenému vínu a 4898 bílému. Budeme tedy uvažovat dvě skupiny (π^1 = červené víno a π^2 = bílé víno) do kterých budeme klasifikovat. Máme tak sdružený výběr $\boldsymbol{X} = (\boldsymbol{X}^1, \boldsymbol{X}^2)$, s rozsahy příslušných výběrů $n^1 = 1599$ a $n^2 = 4898$. Jako odhad apriorních pravděpodobností vezmeme $\hat{p}^1 = \hat{p}^2 = 1/2$. Data obsahují odlehlá pozorování, která jsou patrná už při pohledu na jednotlivé proměnné. Pro ilustraci jsou pro obě odrůdy vína na obrázku 6.1 zobrazeny krabicové grafy pro proměnnou *chloridy*. Ta udává hmotnostní koncentraci soli ve víně (v jednotkách g/dm^3).

Jednotlivá klasifikační pravidla budeme porovnávat pomocí odhadů chyb klasifikace ApER (1.7) a AER_{ts} (1.9). Výběry $\boldsymbol{X}_{tr}$ a $\boldsymbol{X}_{ts}$ sestrojíme tak, že pro každé pozorování $\boldsymbol{x}_j^i$ generujeme náhodnou veličinu u z alternativního rozdělení Alt(1/2) a pokud $u = 0$, pak $\boldsymbol{x}_j^i \in \boldsymbol{X}_{tr}$, v opačném případě $\boldsymbol{x}_j^i \in \boldsymbol{X}_{ts}$. Při konkrétní realizaci tohoto postupu bylo do výběru $\boldsymbol{X}_{tr}$ zařazeno 802 pozorování z $\boldsymbol{X}^1$ a 2472 pozorování z $\boldsymbol{X}^2$.

V tabulce 6.1 jsou uvedeny výsledky pro jednotlivá klasifikační pravidla. Z tabulky vidíme, že odhady chyb ApER a AER_{ts} se příliš neliší. Nejlepší výsledky dává nerobustní QDA (pouze 1,7 % špatně klasifikovaných pozorování). QDA-MCD a QDA-SD dávají chybu klasifikace kolem 2,5 %. To je o něco lepší výsledek, než pro ostatní robustní verze QDA s přibližně 3,2 % špatně klasifikovaných pozorování. Ale všechny tyto výsledky jsou velmi dobré.

Nyní se podíváme na to, jak klasifikaci ovlivní předchozí provedení analýzy hlavních komponent. V tabulce 6.2 jsou uvedeny výsledky pro klasifikační pravidla sestavená pro data transformovaná do prostoru prvních dvou komponent získa-



Obrázek 6.1: Krabicové grafy naměřených hodnot proměnné *chlorides* pro červenou (1) a bílou (2) odrůdu vína.

	ApER	AER_{ts}
QDA	0,017	0,016
QDA-MVE	0,030	0,034
QDA-MCD	0,026	0,022
QDA-S1	0,033	0,033
QDA-S2	0,032	0,033
QDA-SD	0,027	0,024
QDA-OGK	0,030	0,031

Tabulka 6.1: Odhad chyby klasifikace ApER a AER_{ts} různých klasifikačních pravidel aplikovaných na skupinu červených a bílých vín.

ných pomocí PCA-S2. Pro ostatní typy PCA (včetně klasické nerobustní PCA) se výsledky lišili pouze nepatrně, a proto je zde neuvádíme. Odhady ApER a AER_{ts} se opět výrazně neliší. Při použití klasické nerobustní QDA bylo chybně klasifikováno 12,6 % pozorování. Tento výsledek předčily všechny robustní varianty QDA. Jejich chyba klasifikace se většinou pohybovala kolem 7,5 %. Pouze QDA-SD a QDA-OGK dávali o něco málo horší výsledek - asi 8,3 % chybně klasifikovaných pozorování.

	ApER	AER_{ts}
QDA	0,134	0,124
QDA-MVE	0,076	0,074
QDA-MCD	0,073	0,071
QDA-S1	0,076	0,073
QDA-S2	0,075	0,075
QDA-SD	0,084	0,083
QDA-OGK	0,083	0,082

Tabulka 6.2: Odhad chyby klasifikace ApER a AER_{ts} různých klasifikačních pravidel aplikovaných na data o víně, na kterých byla nejprve provedena robustní PCA-S2.

6.2 Technické parametry osobních automobilů

Další zkoumaný soubor dat pochází z knihovny *rpart* programu R, kde ho nalezneme pod názvem *car.test.frame*. Soubor obsahuje technické údaje o několika typech aut. Budeme uvažovat 57 pozorování rozdělených do pěti skupin podle typu: *Small*, *Sporty*, *Compact*, *Medium* a *Van*. Počet objektů v jednotlivých skupinách je uveden v tabulce 6.3. Původní data v knihovně obsahují ještě tři pozorování pro typ *Large*, tuto skupinu však pro její malé zastoupení nepoužijeme. Z veličin, měřených na jednotlivých autech, nás budou zajímat veličiny *Price*, *Mileage*, *Weight*, *Displacement* a *HP* (tedy cena, spotřeba, objem motoru a výkon). Budeme porovnávat různá klasifikační pravidla na základě chyb klasifikace

Typ	Počet
Small	13
Sporty	9
Compact	15
Medium	13
Van	7

Tabulka 6.3: Počty vozů v jednotlivých skupinách v datovém souboru.

	$k = 5$	$k = 3$
QDA	0,053	0,049
QDA-MVE	-	0,195
QDA-MCD	-	0,195
QDA-S1	0,228	0,220
QDA-S2	-	0,171
QDA-SD	0,333	0,244
QDA-OGK	0,123	0,098

Tabulka 6.4: Odhad chyby klasifikace ApER různých klasifikačních pravidel pro pět skupin aut a tři skupiny aut.

	PCA	PCA-PROJ	PCA-GRID	ROBPCA	SPCA
QDA	0,404	0,404	0,386	0,404	0,386
QDA-MVE	0,333	0,333	0,316	0,333	0,316
QDA-MCD	0,333	0,351	0,351	0,333	0,351
QDA-S1	0,368	0,368	0,368	0,368	0,368
QDA-S2	0,333	0,333	0,333	0,333	0,333
QDA-SD	0,333	0,333	0,333	0,333	0,333
QDA-OGK	0,281	0,316	0,333	0,298	0,281

Tabulka 6.5: Odhad chyby klasifikace ApER různých klasifikačních pravidel pro data v pěti skupinách, na kterých byla nejprve provedena PCA, v závislosti na použité metodě PCA.

	PCA	PCA-SD	PCA-PROJ	PCA-GRID	ROBPCA
QDA	0,491	0,491	0,491	0,474	0,491
QDA-MVE	0,439	0,439	0,439	0,456	0,439
QDA-MCD	0,491	0,491	0,491	0,491	0,509
QDA-S1	0,491	0,491	0,491	0,491	0,491
QDA-S2	0,509	0,509	0,509	0,491	0,509
QDA-SD	0,491	0,491	0,491	0,491	0,491
QDA-OGK	0,421	0,439	0,404	0,404	0,439

Tabulka 6.6: Odhad chyby klasifikace CV různých klasifikačních pravidel pro data v pěti skupinách, na kterých byla nejprve provedena PCA, v závislosti na použité metodě PCA.

ApER (1.7) a CV (1.8).

Pokud chceme provést klasifikaci do všech pěti skupin bez předchozí PCA, povede se nám sestavit pouze klasické nerobustní pravidlo QDA a robustní klasifikační pravidla QDA-S1, QDA-SD a QDA-OGK. Důvodem je malé zastoupení typů *Sporty* a *Van*. Pokud budeme uvažovat pouze zbylé tři skupiny (*Small*, *Compact*, *Medium*), podaří se nám sestavit také klasifikační pravidla QDA-S2, QDA-MVE a QDA-MCD. V tabulce 6.4 jsou uvedeny ApER odhady chyby klasifikace pro obě situace. Z tabulky je patrné, že nejlepších výsledků dosahuje nerobustní QDA. Velmi dobré výsledky dává také QDA-OGK.

Vzhledem k tomu, že počet veličin ($p = 5$) je relativně malý vzhledem k počtu aut v jednotlivých skupinách, zkusíme před samotnou klasifikací nejprve provést analýzu hlavních komponent. Porovnáme různé metody PCA tak, že data promítneme do prostoru generovaného prvními dvěma komponentami a na transformovaná data aplikujeme jednotlivá klasifikační pravidla.

V tabulkách 6.5 a 6.6 jsou uvedeny příslušné odhady chyby klasifikace ApER a CV pro všech pět skupin aut. V tabulkách nejsou uvedeny výsledky pro PCA založenou na robustním odhadu varianční matice, neboť PCA-MVE, PCA-MCD, PCA-S1, PCA-S2, PCA-SD i PCA-OGK dávaly shodné výsledky jako ROBPCA. V případě odhadu CV se navíc shodovaly i výsledky pro klasickou PCA a sférickou SPCA, a proto jsou v tabulce 6.6 uvedeny jen výsledky pro nerobustní PCA.

QDA	0,171
QDA-MVE	0,244
QDA-MCD	0,244
QDA-S1	0,212
QDA-S2	0,212
QDA-SD	0,195
QDA-OGK	0,146

Tabulka 6.7: Odhad chyby klasifikace ApER různých klasifikačních pravidel pro data ve třech skupinách, na kterých byla nejprve provedena analýza hlavních komponent.

	PCA	PCA-PROJ	PCA-GRID	ROBPCA	SPCA	PCA-MVE	PCA-SD
QDA	0,171	0,171	0,171	0,171	0,171	0,171	0,171
QDA-MVE	0,293	0,268	0,268	0,293	0,268	0,293	0,293
QDA-MCD	0,268	0,293	0,293	0,317	0,293	0,293	0,317
QDA-S1	0,341	0,341	0,341	0,341	0,341	0,341	0,341
QDA-S2	0,317	0,317	0,317	0,317	0,317	0,317	0,317
QDA-SD	0,293	0,293	0,293	0,293	0,293	0,293	0,293
QDA-OGK	0,244	0,268	0,195	0,244	0,268	0,244	0,268

Tabulka 6.8: Odhad chyby klasifikace CV různých klasifikačních pravidel pro data ve třech skupinách, na kterých byla nejprve provedena PCA, v závislosti na použité metodě PCA.

Z obou tabulek je patrné, že se výsledky příliš neliší, co se týče použité metody PCA. Z odhadů použitých při sestavování klasifikačních pravidel pak dávají nejlepší výsledky OGK-odhad a MVE-odhad.

Z tabulek je také patrný velký rozdíl mezi CV a ApER odhadem chyby klasifikace. Projevuje se tak malé zastoupení aut typu *Sporty* a *Van*. Pokud nebudeme tyto skupiny uvažovat a budeme klasifikovat pouze do zbylých tří skupin (*Small*, *Compact*, *Medium*), rozdíl mezi CV a ApER odhadem nebude tak velký, což je patrné z tabulek 6.7 a 6.8. V případě odhadu chyby klasifikace ApER (viz tab. 6.7) se výsledky vůbec nelišily, co se týče použité metody PCA. CV odhady chyby klasifikace po předchozí PCA-MCD a PCA-OGK byly stejné jako po nerobustní PCA a obdobně PCA-S1 a PCA-S2 dávaly stejné odhady CV jako ROBPCA, a proto nejsou v tabulce 6.8 uvedeny. Z tabulek 6.7 a 6.8 vidíme, že nejlepších výsledků pro případ tří skupin dosahuje nerobustní klasifikace QDA a robustní QDA-OGK.

Závěr

V této práci jsme se věnovali klasické klasifikační analýze a metodám, které vedou k její robustifikaci. Použití těchto metod jsme ilustrovali na reálných datech a v simulacích. Viděli jsme, že pro data obsahující odlehlá pozorování dávají robustní metody mnohdy výrazně lepší výsledky než metody nerobustní. V případě, že data odlehlá pozorování neobsahují, jsou výsledky často srovnatelné. Pokud nemáme informaci o tom, zda data odlehlá pozorování obsahují, zdá se lepší použít robustní metodu.

Simulace však nebyly nijak rozsáhlé a bylo by dobré vyzkoušet další nastavení vstupních parametrů. Například pro menší rozsah výběrů by se pro data bez odlehlých pozorování mohla více projevit menší eficience některých robustních odhadů. Výsledky při použití odpovídajících robustních kalsifikačních pravidel by pak mohly být výrazně horší.

Pokud nemáme jistotu, zda data odlehlá pozorování obsahují, můžeme vyzkoušet robustní i nerobustní klasifikaci a podle jejich výsledků se rozhodnout, kterou z nich použijeme. Mohli bychom se také pokusit vybrat správný odhad ještě před sestavením samotných pravidel. K tomu bychom mohli využít odhad střední hodnoty a varianční matice, který navrhli He a Wang (1996). Autoři uvažují dva odhady - jeden s vysokou eficiencí a jeden s vysokým bodem selhání. Pro tyto odhady sestrojili kritérium, na základě kterého se jako výsledný odhad použije buď eficientní odhad nebo odhad s vysokým bodem selhání. Výsledný odhad pak zdědí obě vlastnosti. Je zároveň eficientní a má vysoký bod selhání. Pokud jsou navíc oba odhady afinně ekvivariantní, je afinně ekvivariantní i výsledný odhad. Použití této metody v klasifikaci by bylo patrně výpočetně velmi náročné. Na druhou stranu by mohlo být přínosné v situaci, kdy některé skupiny odlehlá pozorování obsahují a jiné ne.

Také jsme se zabývali otázkou, jak bude kasifikace ovlivněna, pokud jsme nuceni nejprve snížit dimenzionalitu dat pomocí analýzy hlavních komponent. V takovém případě dávají klasifikační pravidla horší výsledky, neboť využíváme jen část informace, která je v datech obsažena. Pokud data neobsahují odlehlá pozorování, můžeme použít klasickou nerobustní analýzu hlavních komponent. Jednak je tato metoda výpočetně méně náročná, jednak je založena na výběrové varianční matici, která je optimálním odhadem varianční matice pro nekontaminovaná data pocházející z mnohorozměrného normálního rozdělení. Pokud data odlehlá pozorování obsahují, jsou výsledky simulací různorodé. Jako dobrý přístup se jeví zkombinovat analýzu hlavních komponent založenou na metodě *projection pursuit* a robustní klasifikaci založenou na MCD-odhadu.

Literatura

Anděl, J. (1985). *Matematická statistika*. SNTL, Praha.

Butler, R.W., Davies, P.L., Jhun, M. (1993). Asymptotics for the minimum covariance determinant estimator. *The Annals of Statistics*, 21 (3), 1385-1400.

Cortez, P., Cardeira, A., Almeida, F., Matos, T., Reis, J. (2009) Modeling wine preferences by data mining from physicochemical properties. *Support Systems*, 47(4), 547-553.

Croux, C., Filzmoser, P., Oliveira, M. R. (2007). Algorithms for projection-pursuit robust principal component analysis, *Chemometrics and Intelligent Laboratory Systems*, 87(2), 218-225.

Croux, C., Haesbroeck, G. (2000). Principal component analysis based on robust estimators of the covariance or correlation matrix: Influence functions and efficiencies. *Biometrika*, 87(3), 603-618.

Croux, C., Ruiz-Gazen, A. (2005). High breakdown estimators for principal components: the projection-pursuit approach revisited. *Journal of Multivariate Analysis*, 95(1), 206-226.

Davies, P. L. (1987). Asymptotic behaviour of S-estimates of multivariate location parameters and dispersion matrices. *The Annals of Statistics*, 15 (3), 1269-1292.

Davies, P. L. (1992) The asymptotics of Rousseeuw's minimum volume ellipsoid estimator. *The Annals of Statistics*, 20(4), 1828-1843.

Donoho, D. L. (1982) *Breakdown properties of multivariate location estimators*. Ph.D. Qualifying paper, Harvard University.

Frank, A., Asuncion, A. (2010). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml], Irvine, CA: University of California, School of Information and Computer Science.

Gather, U., Hilker, T. (1997). A note on Tyler's modification of the MAD for the Stahel-Donoho estimator. *The Annals of Statistics*, 25, 2024-2026.

Gnanadesikan, R., Kettenring, J. R. (1972). Robust estimates, residuals, and outlier detction with multiresponse data. *Biometrics*, 28, 81-124.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346), 383-393.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., Stahel, W. A. (1986). *Robust statistics: The approach based on influence functions*. Wiley, New York.

Hawkins, D. M., McLachlan, G. (1997). High-breakdown linear discriminant analysis. *Journal of the American Statistical Assocciation*, 92(437), 136-143.

He, X., Wang, G. (1996). Cross-checking using the minimum volume ellipsoid estimator. *Statistica Sinica*, 6, 367-374.

Huber, P. J., Ronchetti, E. M. (2009). *Robust statistics, 2nd edition*. Wiley, New York.

Hubert, M., Van Driesen, K. (2004). Fast and robust discriminant analysis. *Computational Statistics and Data Analysis*, 45(2), 301-320.

Hubert, M., Rousseeuw, P. J., Vanden Branden, K. (2005). ROBPCA: a new approach to robust principal component analysis. *Technometrics*, 47(1), 64-79.

Jurečková, J. (2001). *Robustní statistické metody*. Karolinum, Praha.

Kalina, J. (2012). Implicitly weighted methods in robust imige analysis. *Journal of Mathematical Imaging and Vision*, 44(3), 449-462.

Kalina, J. (2013). Classification methods for genetic data. *Biocybernetics and Biomedical Engineering*. Přijato, v tisku.

Locantore, N., Marron, J. S., Simpson, D. G., Tripoli, N., Zhang, J. T., Cohen, K. L., (1999). Robust principal component analysis for functional data. *Test*, 8(1), 1–73.

Lopuhaä, H. P. (1989). On the relation between S-estimators and M-estimators of multivariate location and scatter. *The Annals of Statistics*, 17, 1662-1683.

Lopuhaä, H. P., Rousseeuw, P. J. (1991). Breakdown points of affine equivariant estimators of multivariate location and covariance matricies. *The Annals of Statistics*, 19(1), 229-248.

Maronna, R. A. (1976). Robust M-estimators of multivariate location and scatter. *The Annals of Statistics*, 4(1), 51-67.

Maronna, R. A., Martin, D., Yohai, V. J. (2006). *Robust statistics: Theory and Methods*, Wiley, New York.

Maronna, R. A., Yohai, V. J. (1995). The behavior of the Stahel-Donoho robust multivariate estimator. *Journal of the American Statistical Assocciation*, 90 (429), 330-341.

McLachlan, G. J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, New York.

Pison, G., Van Aelst, S., Willems, G. (2002). Small sample corrections for LTS and MCD. *Metrika*, 55(1-2), 111-123.

Rao, R. C. (1978). *Lineární metody statistické indukce a jejich aplikace*. Academia, Praha.

Rensová, D. (2008). *Klasifikační analýza*. Bakalářská práce, MFF UK, Praha.

Rencher, A. C. (1996). *Multivariate statistical inference and applicatoins*. Wiley, New York.

Rocke, D. M. (1996). Robustness properties of S-estimators of multivariate location and shape in high dimension. *The Annals of Statistics*, 24(3), 1327-1345.

Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. *Mathematical Statistics and Applications, Vol. B* (W. Grossmann, G. Pflug, I. Vincze a W. Wertz, Eds.), Reidel Publishing Company, Dordrecht, 283-297.

Rousseeuw, P. J., Van Driessen, K. (1999). A fast algorithm for the minimum covariance determimant estimator. *Technometrics*, 41(3), 212-223.

Salibian-Barrera, M., Yohai, V. J. (2006). A fast algorithm for S-regression estimates. *Journal of Computational and Graphical Statistics*, 15(2), 414-427.

Stahel, W. A. (1981). Breakdown of covariance estimators. Research report 31, Fachgruppe für Statistik, ETH, Zurich.

Stigler, S. M. (2010). The changing history of robustness. *American Statistician*, 64(4), 277-281.

Todorov, V., Filzmoser, P. (2009). An object-oriented framework for robust multivariate analysis. *Journal of Statistical Software*, 32(3), 1-47.

Todorov, V., Pires, A. M. (2007). Comparative performance of several robust linear discriminant analysis methods. *REVSTAT - Statistical Journal*, 5(1), 63-83.

Tyler, D. E. (1994). Finite-sample breakdown points of projection-based multivariate location and scatter statistics. *The Annals of Statistics*, 22, 1024-1044.

Seznam použitých zkratek

AER	chyba klasifikace konkrétního klasifikačního pravidla (*actual error rate*), str. 6
AER_{ts}	odhad AER spočtený na základě testovacího výběru, str. 6
ApER	přímý odhad chyby klasifikace (*apparent error rate*), str. 6
APR	asymptotická pravděpodobnost zamítnutí (*asymptotic rejection probability*), str. 23
CV	odhad chyby klasifikace spočtený pomocí křížové validace (*cross-validation*), str. 6
GES	globální citlivost (*gross-error sensitivity*), str. 10
GK-odhad	Gnanadesikanův-Kettenringův odhad, str. 25
GRID	algoritmus pro výpočet PCA založený na metodě PP, str. 32
ER	chyba klasifikace (*error rate*), str. 5
FAST-MCD	rychlý algoritmus pro výpočet MCD-odhadu, str. 18
LDA	výběrové lineární klasifikační pravidlo, str. 5
LDA-MCD	robustní lineární klasifikační pravidlo založené na MCD-odhadu, str. 41
LDA-MVE	robustní lineární klasifikační pravidlo založené na MVE-odhadu, str. 41
LDA-MWCD	robustní lineární klasifikační pravidlo založené na MWCD-odhadu, str. 29
LDA-OGK	robustní lineární klasifikační pravidlo založené na OGK-odhadu, str. 41
LDA-S1	robustní lineární klasifikační pravidlo založené na S-odhadu s dvouváhovou funkcí (2.23), str. 41
LDA-S2	robustní lineární klasifikační pravidlo založené na S-odhadu s funkcí tbiweight (2.24), str. 41
LDA-SD	robustní lineární klasifikační pravidlo založené na SD-odhadu, str. 41
LSS	lokální citlivost odhadu (*local shift sensitivity*), str. 10
MAD	mediánová absolutní odchylka od mediánu (*median absolute deviation*), str. 24
MCD-odhad	odhad minimalizující determinant varianční matice (*minimum covariance determinant estimator*), str. 17
med	medián, str. 24
MVE-odhad	odhad minimalizující objem elipsoidu pokrývajícího část dat (*minimum volume ellipsiod estimator*), str. 16
MWCD-odhad	odhad minimalizující determinant sdružené varianční matice (*minimum within-group covariance determinant estimamtor*), str. 28

OGK-odhad	ortogonalizovaný Gnanadesikanův-Kettenringův odhad, str. 26
PCA	analýza hlavních komponent (*principal component analysis*), str. 30
PCA-GRID	analýza hlavních komponent založená na GRID algoritmu, str. 32
PCA-MCD	analýza hlavních komponent založená na MCD-odhadu, str. 38
PCA-MVE	analýza hlavních komponent založená na MVE-odhadu, str. 31
PCA-OGK	analýza hlavních komponent založená na OGK-odhadu, str. 38
PCA-PROJ	analýza hlavních komponent založená na PROJ algoritmu, str. 32
PCA-S1	analýza hlavních komponent založená na S-odhadu s dvouváhovou funkcí (2.23), str. 38
PCA-S2	analýza hlavních komponent založená na S-odhadu s funkcí tbiweight (2.24), str. 38
PCA-SD	analýza hlavních komponent založená na SD-odhadu, str. 38
PP	metoda PCA založená na hledání optimálních jednorozměrných projekcí (*projection pursuit*), str. 31
PROJ	algoritmus pro výpočet PCA založený na metodě PP, str. 32
QDA	výběrové kvadratické klasifikační pravidlo, str. 5
QDA-MCD	robustní kvadratické klasifikační pravidlo založené na MCD-odhadu, str. 36
QDA-MVE	robustní kvadratické klasifikační pravidlo založené na MVE-odhadu, str. 27
QDA-OGK	robustní kvadratické klasifikační pravidlo založené na OGK-odhadu, str. 36
QDA-S1	robustní kvadratické klasifikační pravidlo založené na S-odhadu s dvouváhovou funkcí (2.23), str. 36
QDA-S2	robustní kvadratické klasifikační pravidlo založené na S-odhadu s funkcí tbiweight (2.24), str. 36
QDA-SD	robustní kvadratické klasifikační pravidlo založené na SD-odhadu, str. 36
ROBPCA	algoritmus pro výpočet analýzy hlavních komponent kombinující metodu PP a robustní odhad varianční matice, str. 33
SD-odhad	Stahelův-Donohův odhad, str. 24
SPCA	sférická analýza hlavních komponent (*spherical principal component analysis*), str. 33