UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA

DIPLOMOVÁ PRÁCE

Regresní analýza pomocí metody dílčích nejmenších čtverců

Katedra matematické analýzy a aplikací matematiky
Vedoucí diplomové práce: **doc. RNDr. Karel Hron, Ph.D.**
Vypracovala: **Bc. Dominika Mikšová**
Studijní program: N1103 Aplikovaná matematika
Studijní obor: Aplikace matematiky v ekonomii
Forma studia: prezenční
Rok odevzdání: 2016

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Bc. Dominika Mikšová

Název práce: Regresní analýza pomocí metody dílčích nejmenších čtverců

Typ práce: Diplomová práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

Vedoucí práce: doc. RNDr. Karel Hron, Ph.D.

Rok obhajoby práce: 2017

Abstrakt: Práce se zabývá regresí pomocí metody nejmenších dílčích čtverců, která nachází největší využití v případech, kdy je v datovém souboru více proměnných než pozorování a také je vhodná při redukci dimenze dat. Standardní metoda nejmenších čtverců v této situaci nelze použít. Na metodu nejmenších dílčích čtverců navazuje robustní a řídká metoda, které zajišťují lepší interpretační vlastnosti. V závěru práce jsou jednotlivé metody demonstrovány na reálných a nasimulovaných datech, následně je uvedeno srovnání kvality predikce uvedených metod. Praktické aplikace jsou realizovány pomocí statistického softwaru R.

Klíčová slova: metoda nejmenších dílčích čtverců, metoda hlavních komponent, regrese, chemometrie, robustní metoda, řídká metoda, software R, singulární rozklad matice, prahový parametr

Počet stran: 68

Počet příloh: 0

Jazyk: český

BIBLIOGRAPHICAL IDENTIFICATION

Author: Bc. Dominika Mikšová

Title: Regression analysis using the partial least squares method

Type of thesis: Masters's

Department: Department of Mathematical Analysis and Application of Mathematics

Supervisor: doc. RNDr. Karel Hron, Ph.D.

The year of presentation: 2017

Abstract: The thesis deals with regression modelling using the partial least squares method. It finds the largest application in cases where a dataset contains more variables than observations. It is also possible to apply this method for dimension reduction. Note that the standard least squares method cannot be used for this purpose. The thesis continues with other approaches, such as the partial least squares method applications of robustness and sparsity constraints, which provide better interpretations. Finally, these methods are demonstrated on real and simulated data followed by the comparison of quality of prediction for previously mentioned methods. Practical applications are conducted in statistical software R.

Key words: partial least squares method, principal component analysis, regression, chemometrics, robust method, sparse method, software R, singular value decomposition, tresholding parameter

Number of pages: 68

Number of appendices: 0

Language: Czech

Prohlášení

Prohlašuji, že jsem diplomovou práci zpracovala samostatně pod vedením pana doc. RNDr. Karla Hrona, Ph.D. a všechny použité zdroje jsem uvedla v seznamu literatury.

V Olomouci dne ....

podpis

Obsah

Seznam obrázků

Poděkování

Na tomto místě si zaslouží především poděkování vedoucí mé diplomové práce doc. RNDr. Karel Hron, Ph.D., a to za významné rady, spolupráci a drahocenný čas, který mi věnoval v průběhu pravidelných konzultací. Poděkování patří také mé rodině a kolegům v zaměstnání, kteří mě po celou dobu psaní mé diplomové práce podporovali a vycházeli vstříc.

Úvod

Ve statistické analýze dat můžeme narazit na různé datové soubory. Nejčastěji se však v matematicko-statistickém prostředí setkáváme s daty, kde máme k dispozici více pozorovaných hodnot než proměnných. V této situaci je vhodné použít při regresní analýze již známou, dobře propracovanou standardní metodu nejmenších čtverců. Nicméně musíme připustit, že ne vždy máme k dispozici předpoklad dostatečného počtu pozorování, a to zejména kvůli finanční a časové náročnosti sběru dat. Další příčinou může také být řídký výskyt určitého jevu, a tak je obtížné dosáhnout velkého počtu pozorování.

Právě metoda nejmenších dílčích čtverců má široké využití v případech, kdy je tento předpoklad porušen. Metoda je předně využívána v oblasti chemometrie, protože v této vědní disciplíně je přirozenou vlastností datového souboru zvlášť velký počet proměnných. Proto bude pro mě dílčím cílem též proniknout určitou mírou do problematiky chemometrie, jež je věda používající analytických oborů, jako je vícerozměrná statistika, aplikovaná matematika a informatika, za účelem řešení problémů v chemii, biochemii, ale také medicíně. Avšak primárním úkolem mé práce bude zejména aplikování zmíněné metody regrese na datové soubory, a to pomocí vhodného statistického softwaru, čímž je software R. Demonstrované příklady vychází z reálných dat, která mi poskytl Ústav molekulární a translační medicíny v Olomouci.

Metoda nejmenších dílčích čtverců nabízí hned několik možností, jak provést kvantifikaci regresního vztahu, přesněji řečeno existují různé algoritmy vedoucí k odhadu neznámých parametrů, což popisuje právě první kapitola spolu s teoretickými a matematickými aspekty zmíněné metody. Nadcházející dvě kapitoly obsahují rozšíření základní metody, konkrétně se jedná o robustní a tzv. řídkou metodu nejmenších dílčích čtverců. Na samotný závěr představíme praktické příklady znázorňující aplikační potenciál uvedené metody.

1 Metoda dílčích nejmenších čtverců

Počátky metody nejmenších dílčích čtverců sahají do roku 1975, kdy se jí zabýval statistik Herman Wold [18]. Primárně ji začal používat v oblasti ekonometrie, později jeho syn Svante Wold a další [19] však našli uplatnění zejména v chemometrii. Dokonce lze říci, že v tomto oboru se zmíněná metoda často používá jako určité dogma. Pro srovnání, počátky klasické metody nejmenších čtverců spadají do roku 1795. Protože se tedy jedná o poměrně novou metodu, setkáváme se s ní prozatím především v cizojazyčné literatuře. V textu tak bude nadále často používaná zkratka PLS, která vychází z anglického termínu *Partial Least Squares*.

Jedná se o třídu metod sloužících k modelování vztahů mezi vysvětlujícími a vysvětlovanými proměnnými prostřednictvím skrytých neboli latentních proměnných. Hlavní úloha tak tkví především v odhadech parametrů v regresním modelu. Nesmíme však opomenout ani další užitečné vlastnosti, jako je zajištění redukce dimenze dat, nebo odstranění multikolinearity čili závislosti mezi sloupci v matici plánu $\boldsymbol{X}$, jenž odpovídají jednotlivým vysvětlujícím proměnným. Zmíněné vlastnosti patří mezi velmi žádané ve statistické analýze. Než ovšem překročíme k zavedení jednotlivých modelů, je zapotřebí zavést označení výsledných odhadů, aby nedošlo k nedorozumění. Pro odhady spojené s metodou nejmenších čtverců použijeme symbol $\hat{\boldsymbol{\beta}}$, avšak pro výsledný vektor odhadů parametrů vycházející z metody PLS včetně její robustní a řídké modifikace využijeme symbolu $\boldsymbol{b}$. Navíc, v literatuře, která se zabývá PLS, je zvykem nerozlišovat matici/vektor neznámých parametrů a příslušný odhad, budeme se tedy držet tohoto přístupu. Rozlišovat se to bude v případě, kdy by mohlo dojít k mylnému pochopení. V této kapitole je čerpáno zejména ze zdrojů [10], [11], [13], [14], [15], [16] a [17].

1.1 Teoretické aspekty

Nejdříve si ukážeme základní dělení metody, a to na PLS1 a PLS2. Rozdíl spočívá ve struktuře vysvětlované proměnné. Jednodušší podobu představuje model PLS1, kde závisle proměnná je pouze jednorozměrná, naproti tomu, jestliže předpokládáme více než jednu závisle proměnnou, uvažujeme model PLS2. Nicméně, pro snadnější uchopení teoretických a také praktických vlastností, v práci bude uveden pouze model PLS1, kde se pracuje s vektorem hodnot vysvětlované proměnné $\boldsymbol{y}$, nikoliv s maticí $\boldsymbol{Y}$.

- **PLS1** $\rightarrow \underbrace{\boldsymbol{y}}_{(n\times 1)} = \underbrace{\boldsymbol{X}}_{(n\times p)}\underbrace{\boldsymbol{b}}_{(p\times 1)} + \underbrace{\boldsymbol{e}}_{(n\times 1)}$, kde $\boldsymbol{y}$ je závisle proměnná, matice $\boldsymbol{X}$ je tzv. matice plánu, jejíž sloupce tvoří vysvětlující nezávisle proměnné, vektor $\boldsymbol{b}$ pak představuje vektor neznámých koeficintů a nakonec vektor chyb $\boldsymbol{e}$.

- **PLS2** $\rightarrow \underbrace{\boldsymbol{Y}}_{(n\times q)} = \underbrace{\boldsymbol{X}}_{(n\times p)}\underbrace{\boldsymbol{B}}_{(p\times q)} + \underbrace{\boldsymbol{E}}_{(n\times q)}$, model pro vícerozměrnou vysvětlovanou proměnnou, díky čemuž je v modelu zahrnuta již matice neznámých parametrů $\boldsymbol{B}$, namísto matice a také matice chyb $\boldsymbol{E}$. Příslušné rozměry n, p a q odpovídají počtu pozorování, počtu vysvětlujících proměnných a u PLS2 počtu vysvětlovaných proměnných, v tomto pořadí.

Základní princip metody PLS se zakládá na spojení metod hlavních komponent a mnohorozměrné lineární regrese. Nyní se blíže podíváme na hlavní myšlenky metody nejmenších dílčích čtverců.

Na úplném počátku vytvoříme umělé proměnné pro matici $\boldsymbol{X}$, popřípadě v modelu PLS2 i pro matici $\boldsymbol{Y}$. Dále přijde na řadu aplikace kritéria, čímž je maximalizace kovariance mezi x-ovými skóry a závisle proměnnou. Použitím kritéria dosáhneme vysoké kvality predikce závisle proměnné. Následuje deflace či peeling, což představuje očištění matice plánu. Jinými slovy, z první komponenty (latentní proměnné) je odstraněna informace. Získáváme residuální matici, navazujeme opět odvozením další PLS komponenty. Nyní si můžeme odpovědět na otázku: Proč metoda ve svém názvu obsahuje pojem „dílčích čtverců“? Od-

pověď vychází právě z odvození další komponenty na základě použití dílčí informace o x-ových, resp. y-ových skórech v modelu PLS2. Předchozí kroky opakujeme až do doby, kdy není dosáhnuto žádného dalšího zlepšení predikce hodnot vektoru $\boldsymbol{y}$. Optimální počet PLS komponent určíme pomocí křížové validace - *cross validation*, o které bude pojednáno později v samostatné podkapitole.

1.2 Výchozí metody

PLS můžeme chápat jako kompromis metody nejmenších čtverců (MNČ) a regrese pomocí metody hlavních komponent. Právě tyto metody si tedy v další části textu nastíníme.

1.2.1 Metoda nejmenších čtverců

Velmi známá metoda, jejíž základní úloha spočívá v určení odhadů neznámých parametrů minimalizací součtů čtverců odchylek skutečných hodnot od odhadnutých.

$$\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_p x_{ip})^2 \to \min, \tag{1.1}$$

kde Y_i značí hodnotu vysvětlované proměnné i-tého pozorování, β_0 absolutní člen, $\beta_1, \ldots, \beta_p$ příslušné odhady a x_{ip} i-té pozorování proměnné p. Díky předchozí minimalizaci dostáváme odhad neznámých regresních koeficientů,

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}. \tag{1.2}$$

Výsledné odhady se pak řadí mezi nejlepší nestranné lineární odhady. Další výhodou je, že pro výpočet odhadů stačí použít jeden exaktní vzorec, avšak za velmi striktních předpokladů. Důležitá je přítomnost více pozorování než počtu proměnných, a dále žádáme silný předpoklad, aby sloupce matice $\boldsymbol{X}$ byly lineárně nezávislé.

1.2.2 Regrese pomocí metody hlavních komponent

Zde se využívá zkratky PCR, která náleží anglickému termínu *Principal Component Regression*. U této metody za startovací krok považujeme klasickou metodu hlavních komponent, která se primárně využívá pro redukci dimenze dat prostřednictvím výsledných skórů a zátěží [2]. V prvním kroku rozložíme matici $\boldsymbol{X}$ na matici skórů $\boldsymbol{T}$ a matici zátěží $\boldsymbol{P}$,

$$\underset{(n\times p)}{\boldsymbol{X}} = \underset{(n\times a)}{\boldsymbol{T}}\underset{(a\times p)}{\boldsymbol{P}^T} + \underset{(n\times p)}{\boldsymbol{E_X}}, \tag{1.3}$$

kde např. rozměry matice $\boldsymbol{T}$ odpovídají n řádkům a počet sloupců je roven a hlavním komponentám pro $j = 1, \ldots, a$. Matice chyb $\boldsymbol{E_X}$ přísluší právě rozložené matici $\boldsymbol{X}$. Pokračujeme dosazením rozložené matice $\boldsymbol{X}$ do původního modelu

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e}. \tag{1.4}$$

Poté si označíme výraz $\boldsymbol{P}^T\boldsymbol{b}$ jako vektor regresních koeficientů $\boldsymbol{g}$, kde $\boldsymbol{e}_T$ představuje vektor chyb

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e} = (\boldsymbol{T}\boldsymbol{P}^T)\boldsymbol{b} + \boldsymbol{e}_T = \boldsymbol{T}\boldsymbol{g} + \boldsymbol{e}_T. \tag{1.5}$$

Je zapotřebí zdůraznit, že chyby $\boldsymbol{e}$ a $\boldsymbol{e}_T$ se od sebe liší, protože k modelování bereme pouze několik málo hlavních komponent. To nám vskutku řeší problém s multikolinearitou neboli závislostí vysvětlujících proměnných, jelikož informace o vysoké korelaci proměnných je zakomponována v nových skórových vektorech, které jsou již nekorelované. Na nově vytvořený model $\boldsymbol{y} = \boldsymbol{T}\boldsymbol{g} + \boldsymbol{e}_T$ aplikujeme MNČ, a dostáváme tak vektor odhadů dílčích regresních koeficientů

$$\boldsymbol{g} = (\boldsymbol{T}^T\boldsymbol{T})^{-1}\boldsymbol{T}^T\boldsymbol{y}. \tag{1.6}$$

My však potřebujeme odhadnout vektor regresních parametrů $\boldsymbol{b}$, proto jej vyjádříme z výrazu $\boldsymbol{g} = \boldsymbol{P}^T\boldsymbol{b}$, čímž docílíme požadovaných odhadů neznámých parametrů

$$\boldsymbol{b}_{PCR} = \boldsymbol{P}\boldsymbol{g}. \tag{1.7}$$

1.2.3 Srovnání regrese metodou nejmenších dílčích čtverců s metodami nejmenších čtverců a hlavních komponent

U metody PLS není potřeba silných předpokladů, jako je tomu u MNČ, máme na mysli absenci multikolinearity - závislosti mezi sloupci v matici plánu a také počet pozorování vyšší než počet proměnných. Ovšem daň za porušení těchto předpokladů je taková, že výsledné odhady u PLS nemůžeme považovat za nejlepší nestranné lineární odhady. Dále, ve srovnání s MNČ, algoritmy u metody PLS pro výpočet odhadů neznámých parametrů pracují na postupném principu, nestačí tedy aplikovat jeden exaktní vzorec.

Společným a také hlavním cílem metod PLS a PCR je dosáhnutí nejlepší predikce, na druhém místě pak vysvětlení co největší variability vysvětlujících proměnných (prediktorů). Obecná metoda hlavních komponent má za ústřední cíl právě vysvětlení co největší celkové variability pomocí několika málo hlavních komponent. PCR pracuje pouze s x-ovými skóry, čímž se podstatně odlišuje od PLS, kde naopak dochází k modelování vztahů mezi dvěma bloky proměnných, z čehož plyne, že PLS bere v úvahu vztah mezi x-ovými a y-ovými skóry. Díky použití vztahu mezi oběma bloky tak dostáváme obecně menší počet hlavních komponent. Nicméně, výhodou u PCR je, že tato metoda vyústí v nekorelované skóry a zároveň v ortogonální zátěže, vedoucí následně k ortonormálnímu systému souřadnic. Naproti tomu, PLS nám dokáže poskytnout pouze jedno, nebo druhé. Například daní za to, že odvozené zátěže budou ortogonální, skóry již ortonormální (nekorelované) neobdržíme. Případ ortogonálních (resp. ortonormálních) zátěží je preferován obzvláště z geometrického hlediska, např. při mapování, další výhodu shledáváme v zachování metrických vlastností. Výsledný obraz pak nebude pokřivený, jak by tomu bylo při použití nekorelovaných skórů. Paradoxně ortonormální skóry jsou v praxi využívány častěji, nicméně se jeví výhodné alespoň z hlediska fixace na metodu hlavních komponent.

1.3 Matematické aspekty

Nyní si představíme technické záležitosti výše popsaných teoretických aspektů. Na úplném začátku jsou sloupce matice $\boldsymbol{X}$, popř. matice $\boldsymbol{Y}$ centrovány, jež jsou dále modelovány právě umělými proměnnými (komponentami). Dekompozice matice $\boldsymbol{X}$ na skóry a zátěže má formálně stejnou podobu jako ve vztahu (1.3). V modelu PLS1 lze rozepsat spojitost mezi x-ovými skóry a vektorem $\boldsymbol{y}$ tzv. vnitřním lineárním vztahem

$$\underset{(n\times 1)}{\boldsymbol{y}} = \underset{(n\times a)}{\boldsymbol{T}}\,\underset{(a\times 1)}{\boldsymbol{d}} + \underset{(n\times 1)}{\boldsymbol{h}}, \tag{1.8}$$

kde $\boldsymbol{d}$ představuje vektor regresních parametrů a vektor $\boldsymbol{h}$ znázorňuje příslušné chyby. Můžeme vidět, že u modelu PLS1 nejsou potřeba y-ové skóry, nýbrž přímo y-ové proměnné.

Pro nalezení regresního vztahu budeme aplikovat kritérium použité na nově vytvořené proměnné, klíčová úloha proto zní - dosáhnout maximální kovariance mezi skóry a n-rozměrným vektorem $\boldsymbol{y}$, jelikož se zabýváme pouze modelem PLS1, nikoliv PLS2. Kritérium matematicky zapisujeme

$$\mathrm{cov}(\boldsymbol{t}, \boldsymbol{y}) \to \max. \tag{1.9}$$

Jednou z možností, jak odhadnout kovarianci mezi vektory $\boldsymbol{t}$ a $\boldsymbol{y}$, je například pomocí vztahu $\boldsymbol{t}^T\boldsymbol{y}/(n-1)$ za podmínky $\| \boldsymbol{t} \| = 1$, jež nám zajistí jednoznačnost řešení. Přitom připomeňme, že pro vektor $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$ je euklidovská norma neboli délka vektoru číslo dané vztahem

$$\| \boldsymbol{x} \| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} = \sqrt{\sum_{i=1}^{n} x_i^2}. \tag{1.10}$$

Vyřešením maximalizační úlohy výše tak získáme první vektor skórů. Další vektory skórů získáme analogicky, které však musí být navíc ortonormální k předchozím získaným vektorům, tedy pro ně platí: $\boldsymbol{t}_j^T\boldsymbol{t}_l = 0$. Z technických důvodů je někdy lepší, jak uvidíme v příští kapitole, zavést další zátěžový vektor $\boldsymbol{w}$ pro x-ové proměnné, pro y-ové proměnné pak vektor $\boldsymbol{c}$, ten však u PLS1

ve většině algoritmů nebude potřeba. Maximalizační kritérium můžeme potom přepsat do následující formulace,

$$\mathrm{cov}(\boldsymbol{Xw}, \boldsymbol{y}) \to \max, \tag{1.11}$$

při dodržení podmínky $\| \boldsymbol{Xw} \| = 1$.

Právě kovariance kombinuje vysoký rozptyl jednotlivých složek v matici $\boldsymbol{X}$ a vysokou korelaci mezi složkami matice $\boldsymbol{X}$ a vektorem $\boldsymbol{y}$. Vysoký (vysvětlený) rozptyl zajišťuje PCR, naproti tomu vysokou korelaci MNČ.

1.4 Algoritmy

Tato podkapitola bude věnována jednotlivým výpočetním algoritmům metody PLS1, kterých existuje celá řada. V praktické části pak budou demonstrovány na konkrétních příkladech. Nutno však ještě zmínit, že vysvětlující proměnné mohou být různého typu - spojité, diskrétní, ale i kategoriální. Pro každou z metod vycházející z PLS existuje i varianta pro klasifikaci, a to v situaci, kdy vysvětlovaná proměnná nabývá hodnot 0 nebo 1. V těchto případech pak označujeme PLS-DA, resp. O-PLS-DA, kde DA značí *Discriminant Analysis* – klasifikační analýza. Tato varianta bude představena v praktické části na reálném příkladu.

Jelikož v následujícím textu bude potřeba znát, jakým způsobem probíhá rozklad matice, nejdříve si obecně definujeme pojem singulární rozklad matice s využitím literatury [16]. Jedná se o vysoce využívaný nástroj jak ve statistice, tak i v matematické oblasti. Naštěstí jej nemusíme počítat ručně, protože softwary nabízí již knihovny s jeho implementací. V anglické literatuře se potkáváme s názvem *Singular Value Decomposition* (dále jen SVD).

Podle SVD je každá matice $\boldsymbol{X}$ velikosti $n \times p$ rozložena na součin tří matic, konkrétně

$$\boldsymbol{X} = \boldsymbol{T}_0 \cdot \boldsymbol{S} \cdot \boldsymbol{P}^T, \tag{1.12}$$

kde pro centrovanou matici $\boldsymbol{X}$ platí, že matice $\boldsymbol{T}_0$ velikosti $n \times p$ obsahuje normované skóry o délce 1 vypočtené metodou hlavních komponent. Matice $\boldsymbol{S}$ je

diagonální maticí o velikosti $p \times p$, kde na hlavní diagonále najdeme tzv. singulární hodnoty, které jsou rovny směrodatným odchylkám $\sqrt{\lambda_j}$ příslušných skórů. Těmito směrodatnými odchylkami rozumíme vlastní čísla matice $\boldsymbol{X}$. Pro úplnost zbývá popsat matici $\boldsymbol{P}^T$, což je transponovaná matice zátěží o velikosti $p \times p$. Z výše uvedeného plyne pro matici skórů vztah $\boldsymbol{T} = \boldsymbol{T}_0 \cdot \boldsymbol{S}$.

Daní za obdržení univerzálního algoritmu bude však vysoká početní náročnost. V případě, kdy máme vzhledem k počtu pozorování příliš velký počet proměnných, je určitě výhodnější, především kvůli časové náročnosti, přistoupit ke snadnějšímu výpočtu. Ten uvažuje matici $\boldsymbol{T}_0$ jako matici $\boldsymbol{X}\boldsymbol{X}^T$, matice $\boldsymbol{P}$ je potom rovna matici $\boldsymbol{X}^T\boldsymbol{X}$. Obě matice mají stejná vlastní čísla, jejichž odmocniny tvoří diagonální prvky matice $\boldsymbol{S}$. Z tohoto důvodu jsou vlastní vektory $\boldsymbol{T}_0$ spočteny z matice $\boldsymbol{X}\boldsymbol{X}^T$ s využitím rozptylů jednotlivých skórů z matice $\boldsymbol{S}^2$. S dosavadními informacemi snadno získáme matici zátěží, a to

$$\boldsymbol{P} = \boldsymbol{X}^T \cdot \boldsymbol{T}_0 \cdot \boldsymbol{S}^{-1} = \boldsymbol{X}^T \cdot \boldsymbol{T} \cdot \boldsymbol{S}^{-2}. \tag{1.13}$$

Software R poskytuje několik příkazů pro vypočtení singulárního rozkladu, nejpoužívanějším je však `svd(X)`. Detailnější popis funkcí a jejich použití bude uvedeno v praktické části práce.

1.4.1 Jádrový algoritmus

Jedním z možných přístupů, jak vyřešit regresní úlohu metodou dílčích nejmenších čtverců, je jádrový algoritmus. V anglických textech se setkáme s názvem *Kernel Algorithm*. Kořeny této metody zasahují do roku 1993. Pojmenování jádrový nese, protože dochází k rozkladu tzv. jádrových matic, přesněji výrazu $\boldsymbol{X}^T\boldsymbol{Y}$ na vlastní vektory a jelikož se zabýváme pouze metodou PLS1, zjednodušíme výraz dále na matici $\boldsymbol{X}^T\boldsymbol{y}$.

Algoritmus směřuje k nekorelovaným skórům, nikoliv k ortogonálním zátěžím. Nyní si ukážeme podrobnější postup vedoucí k odhadu regresních parametrů $\boldsymbol{b}$.

1. Nejprve s využitím SVD spočítáme vlastní vektor $\boldsymbol{w}_1$ odpovídající největšímu vlastnímu číslu λ výrazu $\boldsymbol{X}^T\boldsymbol{y}\boldsymbol{y}^T\boldsymbol{X}$ za podmínky $\| \boldsymbol{X}\boldsymbol{w}_1 \| = 1$.

2. Dostáváme tak první vektor skórů $\boldsymbol{t}_1 = \boldsymbol{X}\boldsymbol{w}_1$.

3. Zátěžový vektor spočítáme pomocí MNČ regrese, jež vychází z rozkladu matice $\boldsymbol{X}$ na součin matice skórů a zátěží

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^T + \boldsymbol{E}_{\boldsymbol{X}};$$

uvažujeme-li pouze první vektor skórů a zátěží, píšeme

$$\boldsymbol{X} = \boldsymbol{t}_1\boldsymbol{p}_1^T,$$

rovnici výše roznásobíme výrazem $\boldsymbol{t}_1^T$

$$\boldsymbol{t}_1^T\boldsymbol{X} = (\boldsymbol{t}_1^T\boldsymbol{t}_1)\boldsymbol{p}_1^T,$$

pomocí jednoduché úpravy pak dostáváme zátěžový vektor

$$\boldsymbol{p}_1^T = (\boldsymbol{t}_1^T\boldsymbol{t}_1)^{-1}\boldsymbol{t}_1^T\boldsymbol{X}.$$

Díky podmínce normality zapisujeme $\boldsymbol{p}_1^T = \boldsymbol{t}_1^T\boldsymbol{X} = (\boldsymbol{X}\boldsymbol{w}_1)^T\boldsymbol{X} = \boldsymbol{w}_1^T\boldsymbol{X}^T\boldsymbol{X}$.

4. Pokračujeme odvozením dalších komponent při stálém dodržení maximalizačního kritéria (1.9). Nastává tedy deflace neboli očištění matice $\boldsymbol{X}$

$$\boldsymbol{X}_1 = \boldsymbol{X} - \boldsymbol{t}_1\boldsymbol{p}_1^T = \boldsymbol{X} - \boldsymbol{t}_1\boldsymbol{t}_1^T\boldsymbol{X} = (\boldsymbol{I} - \boldsymbol{t}_1\boldsymbol{t}_1^T)\boldsymbol{X}.$$

5. Další vektory skórů $\boldsymbol{t}_j$ a zátěží $\boldsymbol{p}_j$ pro $j = 2, \ldots, a$, získáme obdobně. S tím, že nyní už vycházíme z matice $\boldsymbol{X}_1$, nikoliv z matice $\boldsymbol{X}$; matice $\boldsymbol{X}_1$ je následně vždy dále očišťována o hodnoty $\boldsymbol{t}_j\boldsymbol{p}_j^T$. Algoritmus končí obdržením vektorů skórů a zátěží pro poslední možnou komponentu a. Jako kritérium volby a může být použita křížová validace, která je představena v kapitole 1.5.

6. Na samotný závěr se dostáváme k odhadu regresních parametrů

$$\boldsymbol{b} = \boldsymbol{W}(\boldsymbol{P}^T\boldsymbol{W})^{-1}\boldsymbol{y}.$$

Nutno zmínit, že algoritmus výše je vhodnější při velkém počtu pozorování, pokud tomu tak není, přistupujeme k tzv. rozšířené jádrové metodě. Myšlenka spočívá ve spočtení počátečního vektoru $\boldsymbol{w}_1$ z odlišného výrazu, a to $\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{y}\boldsymbol{y}^T$, pokračování algoritmu je dále naprosto analogické.

Samozřejmostí je existence jádrového algoritmu také pro PLS2, který vyžaduje navíc odvození vektorů skórů a zátěží matice $\boldsymbol{Y}$.

1.4.2 NIPALS algoritmus

Z historického hlediska se řadí mezi nejstarší algoritmy použité v rámci metody PLS. Název vychází z anglických termínů *Nonlinear Iterative Partial Least Squares*. Nyní tento algoritmus představíme podrobněji. Předpokládejme n-rozměrnou vysvětlovanou proměnnou, tedy pouze vektor $\boldsymbol{y}$.

1. Nejdříve si inicializujeme matici $\boldsymbol{X}_1$ a vektor $\boldsymbol{y}_1$

$$\boldsymbol{X}_1 = \boldsymbol{X}, \boldsymbol{y}_1 = \boldsymbol{y}.$$

2. Dále spočítáme vektor $\boldsymbol{w}_j$, jenž představuje následující kombinaci

$$\boldsymbol{w}_j = \boldsymbol{X}_j^T\boldsymbol{y}_j/(\boldsymbol{y}_j^T\boldsymbol{y}_j),$$

   a upravíme dělením normou vektoru $\boldsymbol{w}_j \to \boldsymbol{w}_j = \boldsymbol{w}_j/ \parallel \boldsymbol{w}_j \parallel$.

3. Dostáváme tak požadovaný vektor skórů $\boldsymbol{t}_j = \boldsymbol{X}\boldsymbol{w}_j$.

4. Obdobně určíme prvek c_j, nikoliv vektor, jako by tomu bylo u metody PLS2,

$$c_j = \boldsymbol{y}_j^T\boldsymbol{t}_j/(\boldsymbol{t}_j^T\boldsymbol{t}_j).$$

5. Následuje výpočet vektoru zátěží $\boldsymbol{p}_1$

$$\boldsymbol{p}_j = \boldsymbol{X}_j^T\boldsymbol{t}_j/(\boldsymbol{t}_j^T\boldsymbol{t}_j).$$

6. Poté algoritmus vyžaduje očištění matice $\boldsymbol{X}$

$$\boldsymbol{X}_{j+1} = \boldsymbol{X}_j - \boldsymbol{t}_j \boldsymbol{p}_j^T.$$

Postup opakujeme pro všechna j.

7. Na závěr vypočítáme odhady regresních parametrů z původního regresního vztahu $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e}$, odhad vychází z MNČ

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \rightarrow \boldsymbol{b} = \boldsymbol{W}(\boldsymbol{P}^T\boldsymbol{W})^{-1}\boldsymbol{c}.$$

1.4.3 SIMPLS algoritmus

Zkratka náleží anglickému pojmenování *Statistically Inspired Modification of the PLS*. Rozdíl oproti předchozím algoritmům tkví v deflaci, jež není provedena na původní matici $\boldsymbol{X}$. Očištění probíhá na vektoru $\boldsymbol{s} = \boldsymbol{X}^T\boldsymbol{y}$, ke shodě tudíž dochází jen u první komponenty. Krátce představíme postup řídce používaného algoritmu SIMPLS:

1. Určíme matici $\boldsymbol{s}_0 = \boldsymbol{X}^T\boldsymbol{y}$ a poté opakujeme kroky 2 až 6 pro $j = 2, \ldots, a$.

2. Pro $j > 1$, $\boldsymbol{s}_j = \boldsymbol{s}_{j-1} - \boldsymbol{p}_{j-1}(\boldsymbol{p}_{j-1}^T\boldsymbol{p}_{j-1})^{-1}\boldsymbol{p}_{j-1}^T\boldsymbol{s}_{j-1}$.

3. Vypočteme vektor $\boldsymbol{w}_j$, který je u PLS1 roven $\boldsymbol{s}_j$; tento krok řeší maximalizační úlohu. Takto obdržený vektor jednoduše znormujeme $\boldsymbol{w}_j = \boldsymbol{w}_j / \parallel \boldsymbol{w}_j \parallel$.

4. Prostřednictvím projekce matice $\boldsymbol{X}$ na optimální směr jsou získány jednotlivé skóry $\boldsymbol{t}_j = \boldsymbol{X}\boldsymbol{w}_j$.

5. Proběhne úprava j-tého vektoru skórů $\boldsymbol{t}_j = \boldsymbol{t}_j / \parallel \boldsymbol{t}_j \parallel$.

6. V předposledním kroku použijeme regresi metodou nejmenších čtverců k obdržení zátěží $\boldsymbol{p}_j = \boldsymbol{X}_j^T\boldsymbol{t}_j$.

7. Nakonec opět odvodíme požadovaný odhad regresních koeficientů, který má následující podobu

$$\boldsymbol{b} = \boldsymbol{W}\boldsymbol{T}^T\boldsymbol{y}.$$

Algoritmus se od předchozích liší dále maticí $\boldsymbol{W}$, jejíž prvky neboli váhy jsou zde přímo vyprodukovány z matice $\boldsymbol{X}$, nikoliv z očištěné matice, jako tomu bylo v předešlých algoritmech.

1.4.4 O-PLS algoritmus

Modifikací algoritmu NIPALS můžeme rozumět algoritmus O-PLS, jehož název je odvozen z názvosloví *Orthogonal Projections To Latent Structures*. Patří mezi nejmodernější postupy metody PLS používané především v chemometrii. Hlavní myšlenkou O-PLS je rozdělení matice vstupních proměnných $\boldsymbol{X}$ na dvě části. První část je tvořena těmi proměnnými, které nejsou ortogonální k výstupní proměnné $\boldsymbol{y}$, druhá část je tvořena proměnnými z matice $\boldsymbol{X}$, jež jsou ortogonální k vektoru $\boldsymbol{y}$, tedy jsou nekorelované. Poté lze matici $\boldsymbol{X}$ vyjádřit ve tvaru

$$\boldsymbol{X} = \boldsymbol{T}_p \boldsymbol{P}_p^T + \boldsymbol{T}_o \boldsymbol{P}_o^T + \boldsymbol{E}, \tag{1.14}$$

kde $\boldsymbol{T}_o$ reprezentuje skóry a $\boldsymbol{P}_o$ zátěže ortogonální části, $\boldsymbol{T}_p$ a $\boldsymbol{P}_p$ pak pro neortogonální část. Následuje deflace v podobě $\boldsymbol{X} - \boldsymbol{T}_o \boldsymbol{P}_o^T$. Tímto přístupem eliminujeme nekorelovanou část informace směřující k ortogonálním zátěžím. Jelikož při regresi je nežádoucí mít lineární nezávislost mezi dvěma bloky proměnných, stává se tento přístup velmi atraktivním z toho pohledu, že nepodává zkreslené výsledky a považujeme je za důvěryhodné.

1.4.5 Algoritmus vlastních vektorů

V cizojazyčné literatuře se setkáme s názvem *The Eigenvector Algorithm*, který patří k dalšímu, ačkoliv méně známému postupu pro výpočet regrese metody nejmenších dílčích čtverců. Tento způsob výpočtu představuje jednodušší variantu jádrového algoritmu. Hlavní myšlenka a také rozdíl oproti jádrovému algoritmu spočívá v řešení regresní úlohy vypočítáním všech vlastních vektorů, které přísluší největším a vlastním číslům, kde a představuje žádoucí počet PLS komponent. Algoritmus vede na $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_a$ vektorů ortogonálních zátěží v x-ovém prostoru, jež jsou dány právě vlastními vektory příslušející matici $\boldsymbol{X}^T \boldsymbol{y} \boldsymbol{y}^T \boldsymbol{X}$.

Lineární kombinací $\boldsymbol{t}_j = \boldsymbol{X}\boldsymbol{p}_j$ získáme výsledné x-ové skóry. Analogicky, v případě PLS2 by byly vypočteny také ortogonální zátěže y-ového prostoru $\boldsymbol{q}_1, ..., \boldsymbol{q}_a$, a to z matice $\boldsymbol{Y}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{Y}$. Významnou odlišností od předchozích algoritmů nacházíme v očištění matice, která se v tomto postupu reálně neuskuteční. Následkem absence deflace nezískáme nekorelované skóry, naopak dostáváme, jak již bylo řečeno, ortogonální zátěže.

1.5 Kvalita modelu

Model s odhadnutými parametry je třeba podrobit dalšímu ověření, zda je model v dostatečném souladu s daty. K tomu slouží několik ukazatelů, které řadíme mezi míry vhodnosti modelu. K nejpoužívanějším patří charakteristiky odvozené z reziduálního součtu čtverců, například index determinace R^2. Nicméně, vzhledem k naší problematice se nám bude ze široké nabídky nejvíce hodit křížová validace. Dále je možno prověřit model pomocí testů, například test kvality regrese či testy významnosti koeficientů.

Křížová validace

Křížová validace, ale alternativně třeba také metoda bootstrap, je aplikována při zjišťování vhodného počtu komponent v modelu z hlediska predikce vysvětlované proměnné. Při křížové validaci je odstraněna vždy jedna pozorovaná hodnota (či obecně množina hodnot), vytvoří se tak podmnožina a na základě hodnot zbývajících pozorování je predikce provedena pro příslušnou podmnožinu. Takto je postup zopakován tolikrát, dokud není postupně odstraněno každé pozorování, dokud tedy nejsou vypočteny odhady pro všechny možné podmnožiny.

Nyní si představíme křížovou validaci matematicky. Myšlenka spočívá v rozdělení množiny dat na dvě skupiny, první z nich pojmenujeme jako trénovací, druhou pak testovací. Data mohou být náhodně rozdělena do několika segmentů, kde trénovací data, jež vychází ze všech segmentů, jsou formovány pouze do jednoho segmentu. Testovací datová množina je pak tvořená objekty z vynechaného

segmentu. Tímto způsobem vypadá množina určená pro tzv. trénování

$$\boldsymbol{X}_{TRAIN} = \boldsymbol{T}_{TRAIN} \cdot \boldsymbol{P}_{TRAIN}^{T} + \boldsymbol{E}_{TRAIN}, \tag{1.15}$$

s využitím a komponent, kde $\boldsymbol{T}_{TRAIN}$ označuje prvních a skórů a $\boldsymbol{P}_{TRAIN}$ prvních a zátěžových vektorů. Poté dataset příslušející testové množině získáme jako

$$\boldsymbol{T}_{TEST} = \boldsymbol{X}_{TEST} \cdot \boldsymbol{P}_{TRAIN}, \tag{1.16}$$

$$\boldsymbol{E}_{TEST} = \boldsymbol{X}_{TEST} - \boldsymbol{T}_{TEST} \cdot \boldsymbol{P}_{TRAIN}^{T}. \tag{1.17}$$

Jestliže bychom použili pouze jediné rozdělení datového souboru na množinu testovací a trénovací, výsledky by byly velmi zkreslené a nepodávaly by žádnou informaci. Z tohoto důvodu křížová validace pracuje na principu, že každý segment je jednou využit jako testovací a podruhé jako trénovací. Celá procedura je zopakována pro odlišné počty komponent.

Speciálním variantou křížové validace je tzv. „odlož jeden mimo“ (angl. *leave-one-out* - LOO). Interpretace tohoto případu je, že v každé z n iterací je jedno pozorování (podmnožina) použito na testování a zbylých $n - 1$ podmnožin na trénování. Vyplývá, že u této varianty křížové validace provedeme k možných kombinací, což je z časového hlediska velmi náročné, je tedy spíše vhodná pro menší datové datové soubory. Výhodu shledáváme ve skutečnosti, že oproti jakékoliv jiné k-násobné křížové validaci (LKO) dostaneme vždy pouze jednu výslednou hodnotu úspěšnosti. Pro velké datové soubory se však doporučuje používat 10-násobnou křížovou validaci [22]. Mezi tyto typy křížové validace patří například metoda k-fold či jackknife.

Kvalitu predikce budeme posuzovat především na základě ukazatelů SEP, MSEP, RMSEP a indexu determinace R^2, proto si nyní definujeme jejich vztahy. Začneme se čtvercovou chybou predikce

$$\text{SEP} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(e_i - \overline{e})^2}, \tag{1.18}$$

kde $\overline{e}$ značí aritmetický průměr reziduí e_i. Střední čtvercová chyba predikce je dána vztahem

$$\text{MSEP} = \frac{1}{n}\sum_{i=1}^{n} e_i^2, \tag{1.19}$$

RMSEP je definován jako odmocnina ze střední čtvercové chyby predikce,

$$\text{RMSEP} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2}. \tag{1.20}$$

Připomeňme, že pro rezidua $e_i = y_i - \hat{y}_i$; přitom preferujeme nižší hodnoty kritérií před vyššími [4]. Nakonec poslední nástroj index determinace

$$R^2 = 1 - \frac{SSE}{SST}, \tag{1.21}$$

kde SSE představuje residuální součet čtverců $\sum_{i=1}^{n} e_i^2$ a SST celkový součet čtverců $\sum_{i=1}^{n}(y_i - \overline{\boldsymbol{y}})^2$. R^2 nabývá hodnot z intervalu [0,1], nyní preferujeme hodnoty blízké 1, jelikož v tomto případě bude kvalita regresního modelu vysoká.

2 Robustní metoda

V dnešní době se řadí užití robustních alternativ ke standardním statistickým metodám mezi velmi oblíbené. V regresním kontextu na popularitě získávají právě z důvodu redukce vlivu odlehlých hodnot při zkoumání závislosti proměnných v datovém souboru. V kapitole bude čerpáno především z literatury [3] a [16].

V první části této kapitoly nastíníme přehledově problematiku robustní regresní analýzy. Dále přejdeme k samotné aplikaci robustní metody pro nás stěžejní metodu nejmenších dílčích čtverců.

2.1 Úvod do robustní statistiky

Jak již bylo zmíněno výše, robustní statistika umožňuje pracovat s daty, která jsou zcela přirozeně zatížena odchylkami od ideálních statistických modelů. V praxi například nemusí nutně všechna data ležet kolem regresní přímky, dále může být porušen dodatečný předpoklad normality reziduí, že rozptyl náhodných chyb je konstantní a další omezující podmínky ve standardním regresním modelu.

Cílem je tedy najít takový model, jehož předpoklady jsou splněny pro většinu dat, nikoliv splnění striktního požadavku pro všechna data z datového souboru. Odlehlé hodnoty, které se v modelu vyskytují, mohou být produktem nepřesných a také chybných měření. Právě robustní statistika poskytuje nástroj sloužící k potlačení vlivu těchto nežádoucích hodnot. Lze říci, že do modelu je zahrnuta hlavní část dat, která má vypovídající hodnotu pro další analýzu.

V regresní analýze rozlišujeme dva druhy odlehlých pozorování (angl. *outliers*),

- *odlehlá hodnota ve směru y-ové proměnné* - regrese pomocí MNČ je citlivá k vychýleným hodnotám ve směru y-ové proměnné, v anglické literatuře se setkáme s pojmem *vertical outliers*,

- *odlehlá hodnota ve směru x-ové proměnné* - MNČ se zde stává velmi citlivou k odlehlým hodnotám ve směru x-ové souřadnice, můžeme se setkat i s pojmem vlivná pozorování, nebo také *leverage points*.

2.1.1 M-odhady polohy

Problematika robustní regresní analýzy je velmi rozsáhlá, proto bude uvedena pouze část podstatná pro navazující kapitolu. Mezi nejrozšířenější odhady pracující s „rozumnou“ většinou datového souboru patří M-odhady, jimiž se bude zabývat tato kapitola. Dále stojí za zmínku L-odhady a R-odhady, zájemce však odkazujeme na doporučenou literaturu [9], jelikož tato problematika je již nad rámec diplomové práce.

Předpokládejme náhodný výběr hodnot $x_1, \ldots, x_n$, které jsou nezávislé a stejně rozdělené náhodné veličiny s distribuční funkcí F a hustotou f. Problém zde vystává v odhadu charakteristiky polohy t. Odhad je založen na metodě maximální věrohodnosti (z angl. *maximum likelihood estimator*, dále jen MLE), tedy vychází z věrohodnostní funkce, tj. součinu marginálních hustot, kterou maximalizujeme

$$\prod_{i=1}^{n} f(x_i - t) = \max_t. \tag{2.1}$$

Zlogaritmováním výrazu (2.1) dostáváme součet místo součinu, dále přidáním minusového znaménka minimalizujeme funkci, jelikož v mnoha situacích jsou algoritmy napsány právě pomocí minimalizace, nikoliv maximalizace,

$$\sum_{i=1}^{n} [-\ln f(x_i - t)] = \min_t. \tag{2.2}$$

V dalším kroku výraz zderivujeme podle jednotlivých parametrů rozdělení a položíme součet roven nule. Tímto postupem hledáme stacionární bod, jenž je klíčový pro požadovaný odhad

$$\sum_{i=1}^{n} \frac{-f'}{f}(x_i - t) = 0. \tag{2.3}$$

Nyní si předchozí úvahu zobecníme zavedením obecné funkce ρ

$$\sum_{i=1}^{n} \rho(x_i - t) = \min, \tag{2.4}$$

kde v předchozím případě $\rho = -\ln f$. Tuto funkci publikoval poprvé Huber v roce 1964 [7]. Derivací obdržíme novou funkci $\psi = \rho'$, neboli $\psi = \frac{f'}{f}$ výše, poté parametr t splňuje implicitní tvar rovnice

$$\sum_{i=1}^{n} \psi(x_i - t) = 0. \tag{2.5}$$

Je důležité poznamenat, že každý MLE je také M-odhadem, opačné tvrzení však neplatí.

2.2 Robustní regrese

Nyní využijeme dosavadní poznatky k odvození robustní regrese. Máme k dispozici vektor $\boldsymbol{y}$ závisle proměnné, dále matici $\boldsymbol{X}$ vysvětlujících proměnných s absolutním členem, regresní model poté bude ve známém tvaru (1.4), kde o složkách e_i vektoru chyb $\boldsymbol{e}$ předpokládáme, že jsou nezávislé stejně rozdělené náhodné veličiny. Pro rezidua, neboli rozdíly mezi vyrovnanými a naměřenými hodnotami, zavedeme (vektorové) označení $\boldsymbol{r} = \boldsymbol{r}(\hat{\boldsymbol{\beta}}) = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$. Konkrétněji, pro i-té reziduum platí vztah

$$r_i = y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}, \tag{2.6}$$

kde $i = 1, \ldots, n$.

Většina regresních metod je založena na minimalizaci součtu druhých mocnin těchto reziduí, což vyústí v klasickou metodu nejmenších čtverců, v literatuře se běžně setkáváme se zkratkou LS odhad, neboť pochází z anglického názvosloví *Least Squares*. Jelikož již dobře známe odhady regresních parametrů při využití této metody, můžeme rovnou napsat vztah pro výpočet vyrovnaných hodnot,

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y}, \tag{2.7}$$

kde matice $\boldsymbol{H}$ je známá pod názvem *hat matrix*, nicméně v češtině ji můžeme pojmenovat jako klobouková matice či více známá projekční matice. Dále v textu jí tedy budeme rozumět výraz $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$.

Na obr. 2.1 dále uvidíme, že regresní parametry spočítané pomocí metody nejmenších čtverců jsou značně ovlivněny odlehlými hodnotami, a to jak v x-ovém, tak v y-ovém směru. MNČ, kvůli její výrazné citlivosti na odlehlé hodnoty, nelze tudíž vždy považovat za vhodnou metodu.

Nahrazením druhých mocnin absolutní hodnotou získáváme tzv. L_1 odhad, a minimalizujeme tak součet absolutních hodnot reziduí. Odpovídající odhad dostává podobu: $\hat{\boldsymbol{\beta}} = \min \sum_{i=1}^{n} |r_i|$. Oproti MNČ má L_1 regrese výhodu v tom, že je citlivá pouze na odlehlé hodnoty v x-ovém směru, je ovšem robustní vůči odlehlým pozorováním v y-ovém směru. Pro výpočet odhadu regresních parametrů v L1 regresi je ovšem, obdobně jako u ostatních regresních metod, potřeba využívat iteračního algoritmu.

2.2.1 Regrese pomocí mediánu čtverců reziduí

Název této regrese je odvozen z anglické terminologie *Least Median of Squares* (dále pouze LMS), jelikož se jedná o minimalizaci mediánu čtverců reziduí. Nastíníme si hlavní myšlenku této regrese, přičemž začneme uspořádáním absolutních hodnot reziduí $|r|_{(1)} \leq \cdots \leq |r|_{(n)}$, dále si definujme nejjednodušší odhad rozptylu chyb v regresním modelu jako $\sigma(\boldsymbol{r}) = |r|_{(h)}$; pro $h = n/2$ reziduí, odtud pak plyne název regresní metody. Příslušnou minimalizací

$$\hat{\boldsymbol{\beta}} = \min_{\boldsymbol{\beta}} \hat{\sigma}(\boldsymbol{r}(\beta)) = \min_{\boldsymbol{\beta}} |r|_{(h)} \tag{2.8}$$

dostáváme vektor regresních parametrů; kde odhad směrodatné odchylky $\hat{\sigma}$ je robustní standardizovaný odhad reziduí. Tato metoda je charakteristická vysokým bodem selhání - 50 %. Bodem selhání rozumíme nejmenší počet pozorování, která mohou zapříčinit, že odhad může mít libovolnou hodnotu, což v našem případě znamená, že až 50 % pozorování lze přesunout bez toho, aby to výrazně ovlivnilo regresní odhady. Navzdory této vlastnosti se metoda LMS řadí mezi méně účinné,

jelikož vede k odhadům s velkým rozptylem. Navíc, příslušný algoritmus je velmi pomalý v případě velkého počtu vysvětlujících proměnných.

2.2.2 Metoda nejmenších useknutých čtverců

V návaznosti na předchozí metodu byla odvozena tzv. hladší alternativa - metoda nejmenších useknutých čtverců. V nadcházejícím textu budeme užívat zkratky LTS z anglického termínu *Least Trimmed Squares*. Tato metoda je založena na následujícím odhadu rozptylu chyb,

$$\sigma(\boldsymbol{r}) = \left(\frac{1}{n} \sum_{i=1}^{h} |r|^2_{(i)} \right)^{1/2}, \tag{2.9}$$

při $h = n$ dostáváme hodnotu charakteristiky RMSEP, jelikož h se zvětšuje oproti předchozí metodě, bod selhání se úměrně snižuje na hodnotu $\frac{n-h}{n}$.

Dalším krokem je minimalizace tohoto součtu, čímž dostáváme námi požadované odhady regresních parametrů:

$$\hat{\boldsymbol{\beta}} = \min_{\boldsymbol{\beta}} \hat{\sigma}(\boldsymbol{r}(\boldsymbol{\beta})) = \min_{\boldsymbol{\beta}} \sum_{i=1}^{h} |r|^2_{(i)}. \tag{2.10}$$

Díky svým vlastnostem se tato metoda stává v poslední době velmi využívanou. Řadí se mezi nejrobustnější variantu LTS, kterou jsme schopni získat.

2.2.3 M-regrese

Hlavní myšlenka spočívá v aplikaci funkce ρ na rezidua, přičemž výsledný součet funkce ρ, o které bylo pojednáno již v kapitole o M-odhadech 2.1.1, minimalizujeme. Pro názornost formulujeme základní vztah pro M-regresi jako

$$\hat{\boldsymbol{\beta}} = \min \sum_{i=1}^{n} \rho(r_i(\boldsymbol{\beta})). \tag{2.11}$$

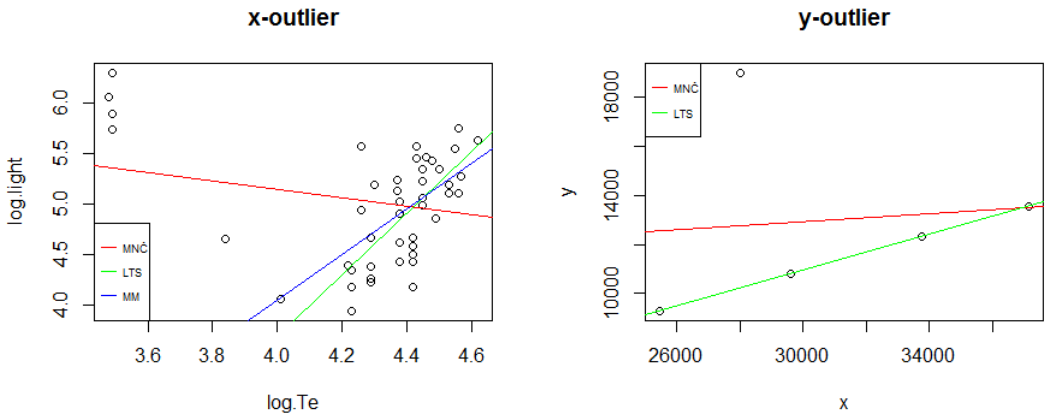
V případě regrese metodou nejmenších čtverců platí $\rho(r) = r^2$, intuitivně pro L_1 odhad zapisujeme $\rho(r) = |r|$. Celá úvaha je tedy založena na redukci vlivu

odlehlých hodnot v důsledku použití vhodné volby funkce ρ, jež je odvozena právě z M-odhadů. Tato regrese bude aplikována dále i v následující kapitole a také v praktické části diplomové práce.

Další možností je regrese pomocí MM-odhadů. Jedná se o metodu, jež vychází z regrese M-odhadů. Značnou výhodu nacházíme v tom, že příslušná regrese je invariantní na změnu měřítka. Odhad je dán vztahem:

$$\hat{\boldsymbol{\beta}} = \min \sum_{i=1}^{n} \rho \left(\frac{r_i(\boldsymbol{\beta})}{\hat{\sigma}} \right). \tag{2.12}$$

Z následující dvojice obrázků 2.1 můžeme vidět, že metoda nejmenších čtverců je značně ovlivněna hodnotami jak v x-ovém, tak v y-ovém směru. Naproti tomu, LTS regrese a také regrese pomocí MM-odhadů nám lépe charakterizuje datový soubor.



Obrázek 2.1: Vliv odlehlých hodnot v x-ovém a y-ovém směru

2.3 Robustní metoda PLS

U všech algoritmů popsaných výše byl odhad kovariance mezi x-ovými skóry a vektorem $\boldsymbol{y}$ spočítán klasickou výběrovou kovariancí. Nesmíme opomenout, že rozptyly a kovariance jsou silně ovlivněné výskytem odlehlých hodnot, tudíž zde

vyvstává obodbný problém i pro maximalizační úlohu (1.9) prostřednictvím maximalizace výrazu $\boldsymbol{X}^T\boldsymbol{y}$. Z těchto důvodů je vhodné aplikovat robustní odhad kovariance.

Při odvození robustní metody budeme vycházet z metody nejmenších dílčích čtverců, avšak v tom smyslu, že místo „nejmenších čtverců“ použijeme M-regresi, a tak výsledná metoda dostává pojmenování „dílčí robustní M-regrese.“ Neřešíme zde obdobně jako u klasické PLS regrese vstupní regresní problém, neboli neodhadujeme vektor regresních parametrů $\boldsymbol{b}$ přímo,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e}, \tag{2.13}$$

nýbrž opět rozložíme matici $\boldsymbol{X}$ na skóry a zátěže,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e} = (\boldsymbol{T}\boldsymbol{P}^T)\boldsymbol{b} + \boldsymbol{e}_T, \tag{2.14}$$

čímž opět využijeme pouze částečné informace x-ových dat. Jako v předchozí kapitole, také zde implementujeme teoretické aspekty pouze pro případ PLS1, avšak problematika může být rozšířená i pro více vysvětlovaných proměnných (PLS2). Stěžejní úlohou je robustně odhadnout regresní koeficienty $\boldsymbol{g} = \boldsymbol{P}^T\boldsymbol{b}$ ve vztahu (2.14), kde matice skórů $\boldsymbol{T}$ je stále neznámá. Hlavní myšlenka spočívá v aplikaci funkce ρ na rezidua regrese $r_i = y_i - \boldsymbol{t}_i^T\boldsymbol{g}$, kde $\boldsymbol{t}_i$ představují jednotlivé řádky příslušné matici skórů, pro $i = 1, \ldots, n$. Funkce ρ má za důsledek potlačení odlehlých hodnot, jinými slovy, snižuje váhu příliš velkých absolutních reziduí, proto je vhodné minimalizovat funkci $\sum_{i=1}^n \rho(y_i - \boldsymbol{t}_i^T\boldsymbol{g})$, jež lze alternativně zapsat jako minimalizaci součtu $\sum_{i=1}^n w_i^r(y_i - \boldsymbol{t}_i^T\boldsymbol{g})^2$ s vhodně zvolenými váhami reziduí $w_i^r = \rho(r_i)/r_i^2$, z čehož plyne označení součtu jako vážený součet čtverců reziduí. Nikoli pouze velká rezidua, ale také vlivná pozorování, jenž se vyznačují odlehlými hodnotami ve směru x-ové souřadnice, můžou ovlivnit odhady regresních koeficientů. Z tohoto důvodu je potřeba zavést dodatečné váhy pro zredukování dopadu také těchto odlehlých hodnot. Výsledné váhy, přiřazené každému vektoru $\boldsymbol{t}_i$, značíme w_i^t. Oba typy vah lze dále kombinovat, např. jako

$w_i = w_i^r w_i^t$. Regresní koeficienty $\boldsymbol{g}$ pak dostaneme minimalizací

$$\sum_{i=1}^{n} w_i (y_i - \boldsymbol{t}_i^T \boldsymbol{g})^2 = \sum_{i=1}^{n} (\sqrt{w_i} y_i - \sqrt{w_i} \boldsymbol{t}_i^T \boldsymbol{g})^2. \tag{2.15}$$

Můžeme vidět, že namísto klasického součtu čtverců reziduí jsou zde data y_i a x-ové skóry vynásobeny vhodnými váhami, a to $\sqrt{w_i}$, poté přichází na řadu aplikace MNČ. Prakticky probíhá výpočet tak, že počáteční hodnoty vah musíme určit, a v dalších krocích jsou upravovány užitím iteračního algoritmu.

Formulace finální úlohy poté zní robustně odhadnout vektory skórů matice $\boldsymbol{T}$, které jsou potřebné ve vztahu výše. Jak již víme z předchozí kapitoly, konkrétně ze vztahu (1.4), pro vysvětlující proměnné je j-tý vektor skórů dán výrazem $\boldsymbol{t}_j = \boldsymbol{X}\boldsymbol{p}_j$, pro $j = 1, \ldots, a$. Podle maximalizačního kritéria (1.9) jsou vektory zátěží $\boldsymbol{p}_j$ získány postupně pomocí

$$\mathrm{cov}_w(\boldsymbol{X}\boldsymbol{p}, \boldsymbol{y}) \to \max \tag{2.16}$$

v závislosti na příslušných omezeních, a to v podobě jednotkového zátěžového vektoru $\| \boldsymbol{p} \| = 1$. Dále požadujeme nekorelovanost jednotlivých skórů $\mathrm{cov}_w(\boldsymbol{X}\boldsymbol{p}_j, \boldsymbol{X}\boldsymbol{p}_l) = 0$ pro $1 \leq l < j$. Poslední podmínku přepsanou do tvaru $\mathrm{cov}_w(\boldsymbol{u}, \boldsymbol{y}) = 0$ s vektorem $\boldsymbol{u}$ délky n si vysvětlíme podrobněji. Zmíněná kovariance zde zastává tzv. váženou kovarianci, která, jak již napovídá název, je definována pomocí výše popsaných vah vztahem $\mathrm{cov}_w(\boldsymbol{u}, \boldsymbol{y}) = 1/n \sum w_i y_i u_i$. Tímto vztahem je splněna jednak podmínka jednotkového zátěžového vektoru a jednak nekorelovanosti všech předchozích získaných zátěžových vektorů. Nyní můžeme přejít k samotnému závěru, tedy vypočtu požadovaných regresních koeficientů. Předpokládáme, že máme k dispozici všechny vektory zátěží, následně je zapotřebí spočítat skóry nám již známým vztahem $\boldsymbol{T} = \boldsymbol{X}\boldsymbol{P}$. Vyřešením regresního problému, čímž rozumíme vztah (2.15), obdržíme koeficienty $\boldsymbol{g} = \boldsymbol{P}^T \boldsymbol{b}$, a konečně přecházíme k původním regresním parametrům, jež nám poskytuje odvozený vztah: $\boldsymbol{b} = \boldsymbol{P}\boldsymbol{g}$.

3 Řídká regrese

S metodou, o niž se pojednává v následující kapitole, se v česky psaných textech spíše nesetkáme, jelikož se jedná o velmi moderní metodu, jež se stává součástí výzkumu statistiků až v posledních dekádách. Chceme-li najít jakékoliv informace, hledáme v anglicky psaných článcích či knihách termín *sparse regression*.

Nejprve představíme řídkou regresi obecně, dále se seznámíme s implementací na námi známé metody PLS a robustní PLS.Následně pro lepší uchopení bude tato modifikace demonstrována na příkladech v praktické části práce. Kapitola byla sepsána za použití literatury [1], [8], [10] a [16].

Opět předpokládáme lineární model (1.4), odhad parametrů klasickou MNČ (1.2), přičemž vyrovnané hodnoty obdržíme vztahem: $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$. V řadě případů, kdy počet proměnných nabývá velkých hodnot, má matice $\boldsymbol{X}$ tendenci obsahovat vysvětlující proměnné, jež nenesou žádnou užitečnou informaci o datovém souboru, navíc zkreslují výsledek vyrovnaných hodnot nebo žádným způsobem tento výsledek neovlivňují, jsou tedy naprosto zbytečné. K tomuto účelu jsme si již v první kapitole popsali metodu hlavních komponent zajišťující snížení dimenze vysvětlujících proměnných a zpracování pouze relevantní informace. Nicméně, odhady parametru $\hat{\boldsymbol{\beta}}$ zahrnují i takovou podmnožinu parametrů $\{\hat{\beta}_{j_1}, \ldots, \hat{\beta}_{j_{\check{p}}}\}$, kde $\hat{\beta}_{j_i}$ pro $i = 1, \ldots, \check{p}$ jsou velmi malá čísla blízká nule náležející $\check{p}$ nevýznamným proměnným. Nutno však poznamenat, že tato malá čísla se nerovnají přesně nule. Tyto koeficienty stále ovlivňují, byť v omezené míře, regresní model, což je ve statistické analýze nežádoucí. Tento problém řeší právě řídký (*sparse*) odhad $\boldsymbol{b}$, který přiřazuje koeficientům s malou důležitostí přesně nulu.

Řídká regrese spadá do tzv. penalizačních regresních metod. Zjednodušeně řečeno, tyto metody kladou podmínky na vektor koeficientů $\boldsymbol{b}$, prakticky volí parametr vyjadřující penaltu (sankci), jak nám napovídá název těchto metod. V našem případě aplikujeme tzv. *Lasso odhad*. Tento penalizovaný odhad používá

L_1 penaltu vedoucí k následujícímu vektoru řídkých koeficientů,

$$\min_{\boldsymbol{b}} \| \boldsymbol{y} - \boldsymbol{X}\boldsymbol{b} \|^2 + \lambda_1 \| \boldsymbol{b} \|_1, \tag{3.1}$$

kde $\| b \|_1 = \sum_{j=1}^{p} |b_j|$. Parametr λ_1 určuje řídkost odhadu a nepřímo uvažuje také $\check{p}$, tedy počet koeficientů rovnajících se velmi nízkým číslům. Řídký odhad v kombinaci s lasso odhadem se staly významným statistickým nástrojem nacházející velkého využití v oblastech chemometrie, ale také bioinformatiky, tedy v odvětvích, kde je v datových souborech typickým znakem větší počet proměnných než pozorování.

Nutno připomenout, že se jedná o nerobustní verzi řídké metody, o robustní alternativě bude pojednáno v kapitole 3.2. Nelze opomenout, že absence robustnosti v modelu způsobuje zkreslení predikované hodnoty zapříčiněné přítomností odlehlých hodnot v datech.

3.1 Řídká PLS regrese

Mezi výraznou přednost metody PLS patří získání velmi dobré predikční schopnosti vysvětlované proměnné. Nicméně, tato metoda současně nedokáže garantovat dobré predikční vlastnosti spolu s vhodným výběrem vysvětlujících proměnných, tedy s redukcí dimenze dat. Tohoto úkolu, současně zajistit obě vlastnosti, se ujímá právě *sparse partial least squares regression* - SPLS, čehož budeme dále užívat v textu.

Postup je následující:

1. Najdeme vlastní vektor příslušný největšímu vlastnímu číslu matice $\boldsymbol{M} = \boldsymbol{X}^T\boldsymbol{y}\boldsymbol{y}^T\boldsymbol{X}$, čemuž odpovídá tato maximalizační úloha,

   $$\max_{\boldsymbol{p}} \boldsymbol{p}^T\boldsymbol{M}\boldsymbol{p} \tag{3.2}$$

   za předpokladu splnění podmínek $\boldsymbol{p}^T\boldsymbol{p} = 1$ a $|\boldsymbol{p}| \leq \lambda$.

2. Abychom dostali námi požadovaný řídký odhad, je k tomu potřeba přeformulovat SPLS kritérium (3.2) v předchozím kroku. Daná formulace zajistí

přesné nuly, nikoliv koeficienty blízké nule zkreslující výsledné predikční hodnoty. Dále užitím L_1 penalty na příslušný tzv. náhradní vektor zátěží $\boldsymbol{c}$ namísto původního zátěžového vektoru $\boldsymbol{p}$ požadujeme

$$\min_{\boldsymbol{p},\boldsymbol{c}} -\kappa \boldsymbol{p}^T \boldsymbol{M} \boldsymbol{p} + (1-\kappa)(\boldsymbol{c}-\boldsymbol{p})^T \boldsymbol{M} (\boldsymbol{c}-\boldsymbol{p}) + \lambda_1 \parallel \boldsymbol{c} \parallel_1 + \lambda_2 \parallel \boldsymbol{c} \parallel, \quad (3.3)$$

za podmínek $\boldsymbol{p}^T \boldsymbol{p} = 1$. První z penalt L_1 přísluší řídkosti parametru $\boldsymbol{c}$, druhá penalta L_2 pak bere v úvahu potenciální singularitu ve výrazu $\boldsymbol{M}$. Možnost, která se zde nabízí, je změnit měřítko parametru $\boldsymbol{c}$ tak, abychom obdrželi $\parallel \boldsymbol{c} \parallel = 1$.

Lze formulovat také vztah pro zobecněnou regresi SPLS, změna spočívá ve snaze získat řešení pro $\boldsymbol{p}$ zafixováním parametru $\boldsymbol{c}$ a poté analogicky najít řešení pro $\boldsymbol{c}$ při zafixovaní $\boldsymbol{p}$. První ze zmíněných úloh řešíme pomocí účelové funkce

$$\min_{\boldsymbol{p}} -\kappa \boldsymbol{p}^T \boldsymbol{M} \boldsymbol{p} + (1-\kappa)(\boldsymbol{c}-\boldsymbol{p})^T \boldsymbol{M} (\boldsymbol{c}-\boldsymbol{p}), \qquad \boldsymbol{p}^T \boldsymbol{p} = 1. \quad (3.4)$$

Jestliže pro parametr κ platí $0 < \kappa < 1/2$, rovnice (3.4) se upraví do tvaru

$$\min_{\boldsymbol{p}} (\boldsymbol{Z}^T \boldsymbol{p} - \kappa' \boldsymbol{Z}^T \boldsymbol{c})^T (\boldsymbol{Z}^T \boldsymbol{p} - \kappa' \boldsymbol{Z}^T \boldsymbol{c}), \qquad \boldsymbol{p}^T \boldsymbol{p} = 1, \quad (3.5)$$

kde $\kappa' = (1-\kappa)/(1-2\kappa)$ a $\boldsymbol{Z} = \boldsymbol{X}^T \boldsymbol{y}$. Problém se dále řeší přes Lagrangeovy multiplikátory. Zájemce o detailní popis výpočetního postupu odkazujeme na literaturu [1].

Druhá z výše uvedených možností spočívá v nalezení řešení vektoru $\boldsymbol{c}$ pro zafixovanou hodnotu $\boldsymbol{p}$,

$$\min_{\boldsymbol{c}} (\boldsymbol{Z}^T \boldsymbol{c} - \boldsymbol{Z}^T \boldsymbol{p})^T (\boldsymbol{Z}^T \boldsymbol{c} - \boldsymbol{Z}^T \boldsymbol{p}) + \lambda_1 \parallel \boldsymbol{c} \parallel_1 + \lambda_2 \parallel \boldsymbol{c} \parallel . \quad (3.6)$$

Metoda SPLS často požaduje vysoké hodnoty penalizačního parametru λ_2 ve snaze vyřešit rovnici (3.6), a to z toho důvodu, že matice $\boldsymbol{Z}$ rozměru $q \times p$ je obvykle s nízkými hodnotami q. Jelikož se zabýváme primárně modelem PLS1, q nabývá nejnižší možné hodnoty, tj. $q = 1$ přísluší jednorozměrnému vektoru $\boldsymbol{y}$. Jako možné opatření se nabízí volba parametru $\lambda_2 = \infty$, který zajistí řešení

rovnice (3.6) v podobě jemného hraničního či prahového odhadu. Více o prahových parametrech bude pojednáno v kapitole 3.1.2. Z předchozích úvah plyne řešení pro obecnou vysvětlovanou proměnnou $\boldsymbol{Y}$, jak pro jednorozměrnou, tak i vícerozměrnou. Dále ukážeme, že první SPLS zátěžový vektor lze spočítat v jednom kroku volbou příslušné meze původního PLS vektoru zátěží.

3.1.1 Algoritmus řídké regrese

Nyní si představíme SPLS algoritmus, který ukazuje, jak extrahovat pouze relevantní proměnné a přitom obdržet kvalitní regresní odhad. Nejprve si definujeme potřebné označení. Novou matici indexů významných proměnných z matice $\boldsymbol{X}$ pojmenujeme jako $\mathcal{A}$ zahrnující aktivní proměnné, tato matice $\mathcal{A}$ je obdržena optimalizací dle vztahu (3.3). Dále písmenem K označíme počet skrytých (latentních) proměnných, a nakonec poslední symbol $\boldsymbol{X}_{\mathcal{A}}$ představuje matici plánu příslušející matici $\mathcal{A}$. V této chvíli můžeme přejít k samotnému algoritmu:

1. Vypočteme $\hat{\boldsymbol{b}}_{PLS} = 0, \mathcal{A} = \{\}, k = 1$ a inicializujeme matici $\boldsymbol{y}_1 = \boldsymbol{y}$.
2. Předpokládejme $k < K$,
   (a) najdeme odhad $\hat{\boldsymbol{p}}$ ze vztahu (3.3), kde $\boldsymbol{M} = \boldsymbol{X}^T \boldsymbol{y}_1 \boldsymbol{y}_1^T \boldsymbol{X}$,
   (b) aktualizujeme matici $\mathcal{A}$ jako $\{i : \hat{p}_i \neq 0\} \cup \{j : \hat{\boldsymbol{b}}_i^{PLS} \neq 0\}$,
   (c) fitujeme PLS regresi pomocí matice plánu $\boldsymbol{X}_{\mathcal{A}}$, která čítá k latentních proměnných,
   (d) díky znalosti nových PLS odhadů zátěžových vektorů $\hat{\boldsymbol{p}}$ obměníme odhady parametru $\hat{\boldsymbol{b}}^{PLS}$,
   (e) aktualizujeme vektor $\boldsymbol{y}_1$ jako $\boldsymbol{y}_1 = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{b}}^{PLS}$, následně $k = k + 1$.

Metoda SPLS je především zajímavá tím, že je schopná v jeden čas pracovat s více než jednou aktivní proměnnou. Od ostatních metod se odlišuje tím, že využívá metodu konjungovaných gradientů ke spočtení koeficientů v každém kroku, a také dokáže simultánně vybrat skupinu korelovaných proměnných k další

analýze. Tato problematika nicméně překračuje charakter práce, proto odkazujeme na literaturu [5].

3.1.2 Výběr prahového parametru a počtu komponent

Ačkoliv vztah (3.3) čítá hned čtyři parametry ($\kappa, \lambda_1, \lambda_2$ a K), ve skutečnosti SPLS regrese zahrnuje pouze dva klíčové parametry, které je potřeba určitým způsobem nastavit. Jedná se o tzv. mezní, prahový či hraniční parametr λ_1 a počet skrytých komponent K. Jelikož algoritmus není výpočetně náročný, nejeví se jako problém zkusit hned několik hodnot κ, které společně s měnícími se hodnotami λ_1 mají vliv na počáteční volbu počtu komponent. Dále parametr λ_2 blížící se k nekonečnu vyprodukuje mezní odhad závislý pouze na parametru λ_1. Nejprve popíšeme tzv. jemný mezní vlastní vektor $\tilde{\boldsymbol{p}}$:

$$\tilde{\boldsymbol{p}} = (|\hat{\boldsymbol{p}}| - \eta \max_{1\leq i\leq p} |\hat{\boldsymbol{p}}_i|) \odot \boldsymbol{I}(|\hat{\boldsymbol{w}}| \geq \eta \max_{1\leq i\leq p} |\hat{\boldsymbol{p}}_i|) \odot \operatorname{sgn}(\hat{\boldsymbol{p}}),$$

kde $0 \leq \eta \leq 1$ a $\boldsymbol{I}(.)$ hraje roli indikátoru příslušné funkce. Symbol $\odot$ představuje tzv. Hadamardův součin [21], který je obecně definován jako

$$\boldsymbol{A} \odot \boldsymbol{B} = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ a_{21} & \cdots & a_{2n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} \odot \begin{pmatrix} b_{11} & \cdots & b_{1n} \\ b_{22} & \cdots & b_{2n} \\ \vdots & \ddots & \vdots \\ b_{n2} & \cdots & b_{nn} \end{pmatrix} = \begin{pmatrix} a_{11} \cdot b_{11} & \cdots & a_{1n} \cdot b_{1n} \\ a_{21} \cdot b_{22} & \cdots & a_{2n} \cdot b_{2n} \\ \vdots & \ddots & \vdots \\ a_{n1} \cdot b_{n2} & \cdots & a_{nn} \cdot b_{nn} \end{pmatrix} \in R^{m\times n}. \tag{3.7}$$

Dále parametr η vyjadřuje řídkost stěžejního parametru λ_1. Ladící parametr η je nastaven prostřednictvím křížové validace pro všechny vlastní vektory.

Další přístup se intuitivně nazývá tzv. hrubé prahování skrz kontrolu falešné míry. SPLS volí proměnné, které jsou charakteristické vysokou korelací se závisle proměnnou y v prvním kroku a přidává dodatečné proměnné s vysokou parciální korelací v následujících krocích.

3.2 Řídká PLS - robustní regrese

Jak již bylo zmíněno, robustní statistika hraje v regresní analýze důležitou roli. Z důvodu eliminace citlivosti na odlehlé hodnoty je proto na místě zavést též robustní obdobu pro SPLS. Můžeme se však na implementaci robustního přístupu dívat dvěma směry. Pvním z nich je aplikace řídké metody na robustní PLS regresi, druhý směr uvažuje na metodu SPLS implementovat robustní M-regresi. Oba přístupy vedou k odlišným výsledkům a uplatňují se při jiných situacích. Ke konci kapitoly budou robustní metody srovnány, a tak budeme mít lepší představu o výhodách a nevýhodách obou přístupů.

V problematice o PLS jsme si již zavedli maximalizační kritérium, omezení na skóry a zátěže, můžeme tedy přistoupit k centrování vysvětlované a vysvětlujících proměnných a následné maximalizaci

$$\mathrm{cov}^2(\boldsymbol{X}\boldsymbol{p}, \boldsymbol{y}) = \frac{1}{(n-1)^2}\boldsymbol{p}^T\boldsymbol{X}^T\boldsymbol{y}\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{p} = \frac{1}{(n-1)^2}\boldsymbol{p}^T\boldsymbol{M}\boldsymbol{p}, \tag{3.8}$$

kde $\boldsymbol{M} = \boldsymbol{X}^T\boldsymbol{y}\boldsymbol{y}^T\boldsymbol{X}$. Regrese závisle proměnné na příslušné skóry poté poskytne klasický vztah pro odhad parametrů. Stěžejní vztahy pro získání robustní regrese jsou detailně popsány ve druhé kapitole.

3.2.1 Řídká robustní PLS metoda

V této podkapitole si představíme první možný směr pro získání řídké a robustní regrese, a to aplikaci řídkosti na již robustní verzi metody nejmenších dílčích čtverců. Pro jednoduchost zavedeme označení tohoto přístupu jako PRM. Předpokládejme, že máme již aplikovanou robustní PLS metodu popsanou v kapitole 2.3. Vzhledem k dalšímu použití je však zapotřebí dalšího označení pro váhy příslušné matici $\boldsymbol{X}$ a vektoru $\boldsymbol{y}$

$$\tilde{\boldsymbol{X}} = \boldsymbol{\Omega}\boldsymbol{X} \text{ a } \tilde{\boldsymbol{y}} = \boldsymbol{\Omega}\boldsymbol{y}, \tag{3.9}$$

kde $\boldsymbol{\Omega}$ představuje diagonální matici s prvky $w_i \in [0, 1]$ pro $i = 1, \ldots, n$. Odlehlé hodnoty logicky obdrží hodnotu váhy nižší než 1. V kapitole 2.3 jsme definovali

váhy w_i jednoduše jako kombinaci $w_i = w_i^r w_i^t$. V literatuře se ovšem můžeme setkat např. také s

$$w_i^2 = w^r\left(\frac{r_i}{\hat{\sigma}}\right) w^t\left(\frac{\|\boldsymbol{t}_i - \mathrm{med}_j(\boldsymbol{t}_j)\|}{\mathrm{med}_i\|\boldsymbol{t}_i - \mathrm{med}_j(\boldsymbol{t}_j)\|}\right). \tag{3.10}$$

V tomto okamžiku můžeme přejít k samotnému řídkému odhadu, jehož dosáhneme nastavením vhodné volby penalizačního parametru L_1 vzhledem k zátěžovým vektorům. Příslušný odhad dostaneme vyřešením následující úlohy, která nám právem připomíná rovnici (3.3)

$$\min_{\boldsymbol{p},\boldsymbol{c}} -\kappa \boldsymbol{p}^T \tilde{\boldsymbol{M}}^T \tilde{\boldsymbol{M}} \boldsymbol{p} + (1-\kappa)(\boldsymbol{c}-\boldsymbol{p})^T \tilde{\boldsymbol{M}}^T \tilde{\boldsymbol{M}}(\boldsymbol{c}-\boldsymbol{p}) + \lambda_1 \|\boldsymbol{c}\|_1, \tag{3.11}$$

kde $\tilde{\boldsymbol{M}} = \tilde{\boldsymbol{y}}^T \tilde{\boldsymbol{X}}$ a opět jsme použili náhradní zátěžový vektor $\boldsymbol{c}$. Rovnice je zatížena podmínkami, a to konkrétně

$$\|\boldsymbol{p}_j\| = 1 \text{ a } \boldsymbol{p}_j^T \tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}} \boldsymbol{p}_i = 0 \text{ pro } 1 \leq i \leq j.$$

Konečný odhad formulujeme jako normovaný podíl $\boldsymbol{p}_j = \frac{\hat{\boldsymbol{c}}}{\|\hat{\boldsymbol{c}}\|}$ tzv. náhradního vektoru $\hat{\boldsymbol{c}}$, který navíc minimalizuje vztah (3.11). Tímto postupem obdržíme řídkou matici robustně odhadnutých zátěží $\boldsymbol{P}$ a skórů $\boldsymbol{T} = \boldsymbol{X}\boldsymbol{P}$. Po provedení regrese vysvětlované proměnné v závislosti na minimalizaci součtu $\sum_{i=1}^n \rho(y_i - \boldsymbol{t}_i^T \boldsymbol{g})$, která byla zmíněna již v kapitole 2.3, dostaneme výsledný řídký dílčí robustní M-odhad. Nakonec zdůrazníme, že řídkost odhadnutých zátěží je uskutečněna právě díky regresním koeficientům.

3.2.2 Robustní SPLS metoda

Nyní uvedeme druhý možný přístup, a to aplikaci robustnosti na již řídkou metodu nejmenších dílčích čtverců (SPRM). Označme $\boldsymbol{p}_j$ jako klasický, neřídký PLS vektor zátěží deflované matice $\boldsymbol{X}$, potom SPLS řešení nacházíme prostřednictvím vztahu

$$\boldsymbol{w}_j = (|\boldsymbol{p}_j| - \lambda_1/2) \odot \boldsymbol{I}(|\boldsymbol{p}_j| - \lambda_1/2 > 0) \odot \mathrm{sgn}(\boldsymbol{p}_j), \tag{3.12}$$

kde $\boldsymbol{I}(.)$ značí indikátor funkce přiřazující prvky zátěžového vektoru tak, aby se jeho prvky rovnaly 1 v případě, že argument je pravdivý, a 0 jinak.

Ve výše uvedené rovnici (3.12) $|\boldsymbol{p}_j|$ znázorňuje vektor absolutních hodnot zátěžového vektoru $\boldsymbol{p}_j$ náležející j komponentám. V poslední části vztahu (3.12) se objevuje $\text{sgn}(\boldsymbol{p}_j)$, což v matematickém jazyku vyjadřuje vektor znamének jednotlivých komponent. Uvažujeme-li matici $\boldsymbol{W}$, tak její sloupce jsou tvořeny vektory $\boldsymbol{w}_j$, pro $j = 1, \ldots, a$. Konečně můžeme přistoupit k odvození řídkých vektorů zátěží ve smyslu původních nedeflovaných proměnných, jež jsou dány vztahem $\boldsymbol{P} = \boldsymbol{W}(\boldsymbol{W}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{W})^{-1}$.

Nyní se vrátíme k rovnici (3.12) a přeformulujeme ji tak, aby v sobě zahrnovala i ladící parametr η

$$\boldsymbol{w}_j = (|\boldsymbol{p}_j| - \eta \max_i |p_{ij}|) \odot \boldsymbol{I}(|\boldsymbol{p}_j| - \eta \max_i |p_{ij}| > 0) \odot \text{sgn}(\boldsymbol{p}_j), \tag{3.13}$$

kde prvky p_{ij} náleží i-té složce zátěžového vektoru $\boldsymbol{p}_j$. Parametr $\eta \in [0, 1)$ vyjadřuje velikost hranice, pod kterou jsou již všechny prvky vektoru $\boldsymbol{w}_a$ rovny právě nule. Díky znalosti intervalu, ve kterém se pohybuje hodnota hraničního parametru, lze použít křížovou validaci pro vhodnou volbu ladícího parametru.

Přichází na řadu definovat detailněji váhy aplikované v robustní metodě PLS, konkrétně v sekci 2.3; pro připomenutí - máme na mysli regresi pomocí M-odhadů. V této chvíli se zaměříme na vhodnou volbu vah ve výrazu daném vztahem (2.15). K tomuto účelu použijeme tzv. klesající Hampelovu váženou funkci [6], jež poskytuje kompromis mezi robustností a také eficiencí, tzn., kdy na úkor vysoké robustnosti nepřicházíme zároveň o přesnost výsledných odhadů. Tato funkce je dána vztahem

$$\omega(x) = \begin{cases} 1 & |x| \leq a \\ \frac{a}{|x|} & a < |x| \leq b \\ \frac{q-x}{q-b}\frac{a}{|x|} & b < |x| \leq q \\ 0 & q < |x|, \end{cases} \tag{3.14}$$

kde ladící konstanty a, b a q jsou kvantily daných rozdělení. Vezmeme-li váhy reziduí w^r, uvažujeme kvantily 0,95, 0,975 a 0,999 normálního rozdělení, pro váhy w^t pak chí-kvadrát rozdělení s příslušnými kvantily.

Algoritmus SPRM

Finální odstavec teoretické práce bude věnovaný právě algoritmu, jehož kroky vedou k získání řídkého odhadu robustní SPLS metody.

1. Nejprve si opět inicializujeme matici $\boldsymbol{X}$ a také vektor vysvětlované proměnné $\boldsymbol{y}$ - tyto startovací hodnoty jsou robustně centrovány odečtením mediánu.

2. Následuje výpočet počátečních vah,

   (a) začneme výpočtem vzdáleností jak pro $\boldsymbol{x}_i$, tak pro y_i v tomto pořadí

   $$d_i = \frac{\|\boldsymbol{x}_i\|}{\text{med}_j\|\boldsymbol{x}_j\|},$$
   $$r_i = \frac{|y_i|}{c\ \text{med}_j|y_j|},$$

   kde $i = 1, \ldots, n$ a $c = 1{,}4829$ za účelem docílení konzistence mediánové absolutní chyby (MAD),

   $$\text{MAD} = \text{median}_i|x_i - \text{median}(\boldsymbol{x})|,$$

   (b) přichází na řadu definovat počáteční váhy $w_i = \sqrt{w^t(d_i)w^r(r_i)}$, tyto prvky dostanou využití v diagonální matici $\boldsymbol{\Omega}$.

3. Iteračním procesem získáme postupně další váhy,

   (a) Data zatížená váhami a jejich označení:

   $$\boldsymbol{X}_w = \boldsymbol{\Omega X},$$
   $$\boldsymbol{y}_w = \boldsymbol{\Omega y}.$$

   (b) Pokračujeme aplikováním řídké verze algoritmu NIPALS, přičemž bereme v úvahu váženou variantu matic $\boldsymbol{X}_w$ a $\boldsymbol{y}_w$, získáváme tak matici skórů $\boldsymbol{T}_w$, zátěží $\boldsymbol{P}_w$, odhad vektoru neznámých koeficientů $\boldsymbol{b}_w$ a v neposlední řadě vektor vyrovnaných hodnot $\hat{\boldsymbol{y}}_w$.

(c) V této části algoritmu vypočítáme váhy pro skóry a vysvětlovanou proměnnou tak, že vycentrujeme $\mathrm{diag}(1/w_1, \ldots, 1/w_n)\boldsymbol{T}_w$, čímž obdržíme matici $\tilde{\boldsymbol{T}}$. Dále vyčíslíme vzdálenosti pro $\tilde{\boldsymbol{t}}_i$ a pro robustně centrovaná a standardizovaná rezidua r_i

$$d_i = \frac{\|\tilde{\boldsymbol{t}}_i\|_2}{\mathrm{med}_j \|\tilde{\boldsymbol{t}}_j\|_2},$$
$$r_i = \frac{|y_{w,i} - \hat{y}_{w,i} - \mathrm{med}_k(y_{w,k} - \hat{y}_{w,k})|}{c\ \mathrm{med}_j |y_{w,i} - \hat{y}_{w,i} - \mathrm{med}_k(y_{w,k} - \hat{y}_{w,k})|},$$

iterační proces končí aktualizováním vah $w_i = \sqrt{w^t(d_i)w^r(r_i)}$.

Postup jednotlivých bodů ve třetím kroku opakujeme dokud nezískáme konvergenci odhadu váženého vektoru neznámých parametrů $\boldsymbol{b}_w$.

4. Závěrečná iterace ukončí algoritmus získáním výsledných odhadů $\boldsymbol{b}$, matice zátěží $\boldsymbol{P}$ a nakonec matice skórů $\boldsymbol{T} = \boldsymbol{X}\boldsymbol{P}$.

4 Praktická část

V této části budou předchozí teoretické poznatky s využitím statistického softwaru R aplikovány na konkrétních datech, přičemž datové soubory budou hned dvojího typu. Nejprve si vygenerujeme hodnoty datové sady náhodně, druhý z představených příkladů bude vycházet z reálných dat. Naším cílem pak bude srovnat všechny metody z pohledu jejich predikční schopnosti a také výběru vhodného počtu vysvětlujících (umělých) proměnných. V obou příkladech budou vloženy spolu s grafickou interpretací i ukázky jednotlivých příkazů z prostředí softwaru R. V praktické části byly použity především zdroje [11], [12], [15], [20], [23] a [24].

4.1 Simulační studie

Myšlenka spočívá nejprve ve vygenerování náhodných čísel z mnohorozměrného normálního rozdělení, následně provedeme regresní analýzu užitím výše zmíněných metod a vyhodnotíme kvalitu příslušných modelů. Dále provedeme tutéž analýzu při zařazení odlehlých pozorování a nakonec předvedeme výsledné srovnání.

Jelikož regresní metody, kterými jsme se v této práci zabývali, jsou zaměřené především na data s vysokým počtem proměnných a nízkým počtem pozorování, nasimulujeme si případ, kdy máme 600 proměnných a pouze 20 pozorování. Kvůli velkému rozměru matice plánu $\boldsymbol{X}$ jsou uvedeny pouze počáteční a koncové hodnoty.

4.1.1 Náhodná data bez odlehlých hodnot

```
> library(MASS)
> Sigma<-matrix(0.1,600,600)
> diag(Sigma)<-1    # matice sigma - na hlavní diagonále jsou 1,
                    # mimo diagonálu 0,1
```

```
> X<-mvrnorm(n=20,rep(0,600),Sigma)

             [,1]       [,2]       [,3]        [,4]        [,5]       [,6]
[1,]  0.889573517  0.9861291  2.3107508  0.71023694 -0.69152228 -0.5196984
[2,]  0.621027248  0.3967060 -0.4382004 -0.49107972 -0.04885482 -0.8602748
[3,] -0.371554734 -0.9985105  0.9317513  0.85026329  1.10078609 -0.8609645
[4,] -1.160166200  0.1856903 -0.6415714 -0.07383014  0.44405979  0.8999677
[5,] -0.008151431 -0.9381964 -2.5717257 -1.65190683 -0.84618167  0.1748377
[6,]  1.144342699  1.0233284 -0.2942201 -0.59791453 -1.19048104  0.9987625

           [,595]     [,596]     [,597]     [,598]     [,599]      [,600]
[15,] -0.10000348 -1.0355081 -0.6427298 -0.4184698 -0.4423897 -0.25841548
[16,] -1.14433404 -0.4298289 -0.9480865  0.8772079  0.4552718 -0.36730961
[17,]  0.11253224  0.1693885  0.4472491 -0.3685243  0.5505415 -0.08747634
[18,] -0.06479084 -1.1071541 -1.5110764  0.7401946 -0.3608100  0.67334701
[19,] -0.80117322  1.7174968 -1.2609340  0.7983749 -0.3502025  1.04379195
[20,]  0.24019192  0.1509314  0.2981366 -0.5270245  0.3788008  1.24516923
```

Protože jsme se v práci zabývali zejména modelem PLS1, vygenerujeme si náhodně hodnoty vysvětlované proměnné rozměru $n \times 1$. Součástí příslušných knihoven v softwaru R jsou ovšem také funkce navržené pro model PLS2.

```
> y<-mvrnorm(n=20,rep(0,1),Sigma))

 [1,] 0.2922245 -0.1654144 0.3795482 0.2602218 0.3446628 0.4834142 -0.9068365
 [8,]-0.5878771 -0.2269022 -1.152216 -0.641974 0.0689810 -1.644741 -0.2291005
[15,] 1.2643211 -0.1787650 0.8867992 0.3094683 0.9960373 -1.4879324
```

PLS metoda

Software R nabízí hned několik knihoven zahrnující funkce, které můžeme implementovat na metodu nejmenších dílčích čtverců a její modifikace. Začneme modelováním metody PLS a příslušnou knihovnou `pls` a jejími funkcemi `mvr`, `pslr` a další.

```
> simpls<-mvr(y~X,ncomp=5,method="simpls")      # SIMPLS algoritmus
> opls<-mvr(y~X,ncomp=5,method="oscorespls")    # O-PLS algoritmus
> kernel<-mvr(y~X,ncomp=5,method="kernelpls")   # Jádrový algoritmus
> eigen<-pls_eigen(y~X,a=5)                     # alg. vlast. vektorů
  # nefunguje, lze provést pouze pro model PLS2
> fitpls<-plsr(y~X,method=pls.options("kernelpls")$plsralg, ncomp=8)
> summary(fitpls)
```

Necháme-li si zobrazit shrnutí této regrese, vidíme, že 15 zvolených komponent vysvětluje téměř 80 % variability, které nám nahrazují původních 600 proměnných. Následující informace jsou výstupem příkazů výše.

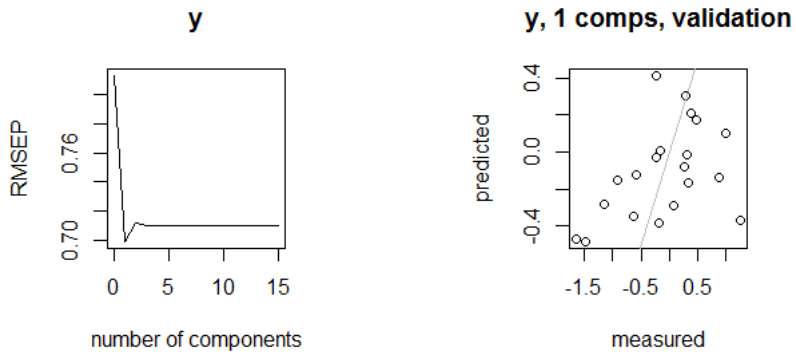
```
Data:  X dimension: 20 600
Y dimension: 20 1
Fit method: kernelpls
Number of components considered: 15

VALIDATION: RMSEP
Cross-validated using 10 random segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
CV          0.8241   0.7875   0.7911   0.7909   0.7910   0.7910   0.7910
adjCV       0.8241   0.7497   0.7506   0.7503   0.7504   0.7504   0.7504
       7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
CV      0.7910   0.7910   0.7910    0.7910    0.7910    0.7910    0.7910
adjCV   0.7504   0.7504   0.7504    0.7504    0.7504    0.7504    0.7504
       14 comps  15 comps
CV       0.7910    0.7910
adjCV    0.7504    0.7504

TRAINING: % variance explained
   1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps
X    6.031    11.43    16.86    22.39    27.49    32.74    38.12    43.53
y   97.029    99.94   100.00   100.00   100.00   100.00   100.00   100.00
   9 comps  10 comps  11 comps  12 comps  13 comps  14 comps  15 comps
X     48.8      53.9     58.78     63.91     69.29     74.14     79.24
y    100.0     100.0    100.00    100.00    100.00    100.00    100.00
```

V kapitole 1.5 jsme si představili, jakým způsobem budeme posuzovat kvalitu predikce, proto si následně pomocí softwaru R tyto charakteristiky vyčíslíme. Výsledné hodnoty jednotlivých modelů a charakteristik bude uvedena na konci této kapitoly.

Obrázek 4.1 ukazuje na prvním grafu vhodný počet komponent, a to dokonce pouze jednu. Nicméně druhý z grafů nevykazuje příliš vysokou kvalitu predikce, což můžeme vidět pouhým okem, že minimální počet pozorovaných dat leží na vyznačené přímce.

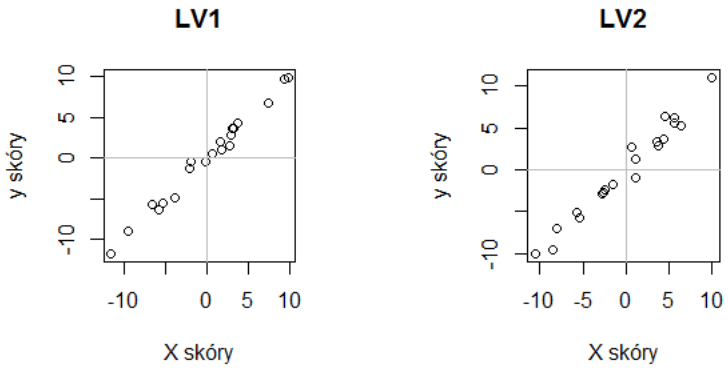
```
> RMSEP(fitpls,ncomp = 2,intercept = FALSE)
       (Intercept)  2 comps
```

```
CV           0.813    0.7247
adjCV        0.813    0.6881
> MSEP(fitpls,ncomp = 2,intercept = FALSE)
(Intercept)      2 comps
   0.596493     0.002948
> R2(fitpls,ncomp=2,intercept = FALSE)    [1]  0.9951
(Intercept)      2 comps
    -0.1080       0.1196
```



Obrázek 4.1: Křížová validace a kvalita predikce

Na dalším obr. 4.2 jsou vykresleny vztahy mezi x-ovými a y-ovými skóry. Můžeme říci, že oba vztahy vykazují lineární závislost.

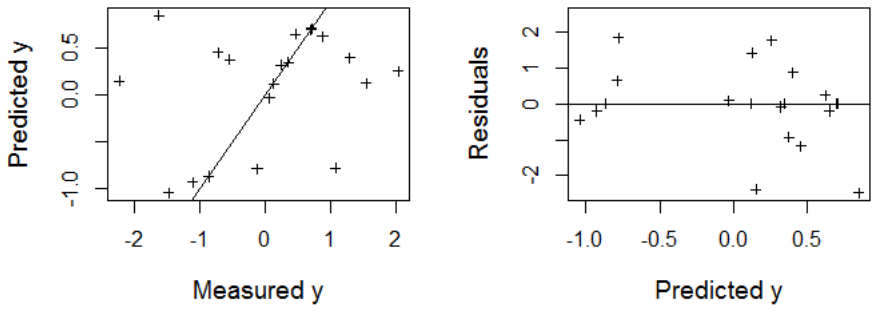


Obrázek 4.2: Vztah mezi x-ovými a y-skóry pro 1. a 2. komponentu

Robustní metoda PLS

Nyní si představíme robustní PLS regresi, která slibuje menší citlivost na odlehlá pozorování, očekáváme tedy vyšší kvalitu predikce. K této metodě byla využita knihovna `chemometrics` s těmito příkazy,

```
> robust.cv <- prm_cv(X,y,a=15,segments=4,plot.opt=TRUE)
> prm(X, y, a = robust.cv$optcomp, opt = "l1m", usesvd = TRUE)
> plotprm(robust.cv, y)
```



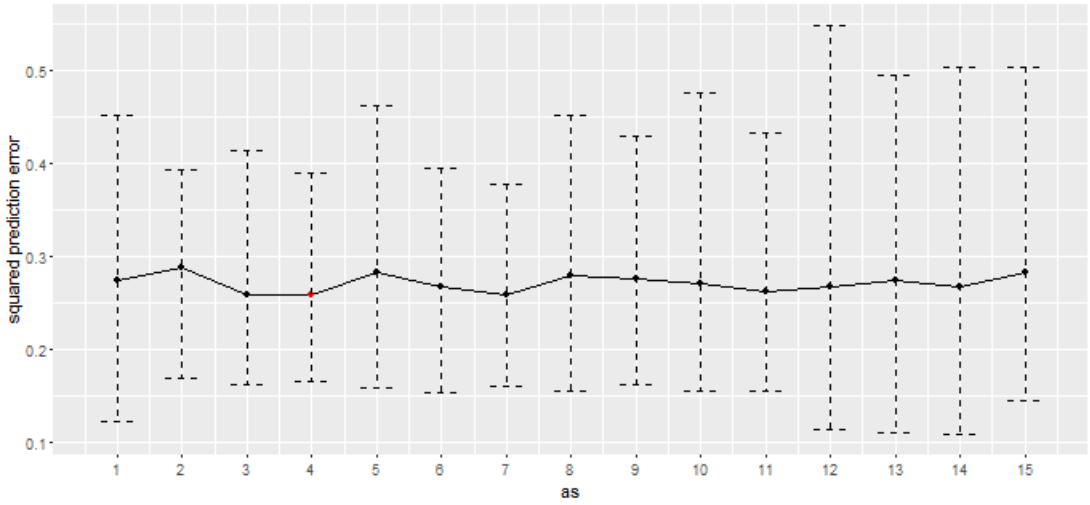
Obrázek 4.3: Rezidua a predikované hodnoty

Obrázek 4.3 v levé části zobrazuje naměřené hodnoty versus vyrovnané hodnoty. Ve srovnání s nerobustní verzí PLS dostáváme rozhodně lepší hodnotu RMSEP. Napravo pak vidíme vykreslené vyrovnané hodnoty versus rezidua. Detailnější analýzu jsem se rozhodla provést pomocí implementace balíčku `sprm`, jelikož nabízí hned několik funkcí, které využijeme k analýze modelů jak pro robustní, tak pro řídkou metodu PLS.

```
> library(sprm)
> d<-as.data.frame(X)
> d$y<-y
> robustPLS<-prms(y~X,data=d,a=15,fun="Hampel")
> robustPLS$fitted.values
```

Příkaz `summary(robustPLS)` nám vytiskne vysvětlenou variabilitu původních dat pomocí nových latentních proměnných. Ukazuje se, že 15 komponent vysvětluje 81,78 %, tedy o dva procentní body více než metoda PLS.

Při 4 zvolených hlavních komponentách nabývá charakteristika SEP nejnižší hodnotu, již za použití knihovny `sprm`. Třebaže tyto 4 komponenty vysvětlují pouze 23,1 % variability původních vysvětlujících proměnných.



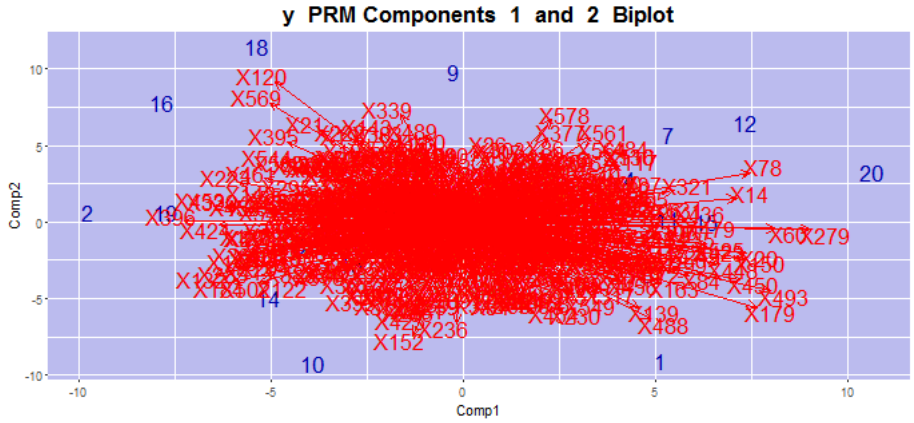
Obrázek 4.4: Robustní PLS - křížová validace

Dalším z výstupů je graf, který je v dnešní době hodně užívaným nástrojem v prostředí mnohorozměrných dat - biplot (4.5). Nutno uznat, že se nejedná kvůli velkému počtu proměnných o přehledný obrázek, nicméně interpretace je stále možná a je následující - biplot zobrazuje vliv původních proměnných na nové latentní proměnné a také jakým způsobem přispívají do modelu. Většinou se jedná o projekci dat na dvě hlavní komponenty.

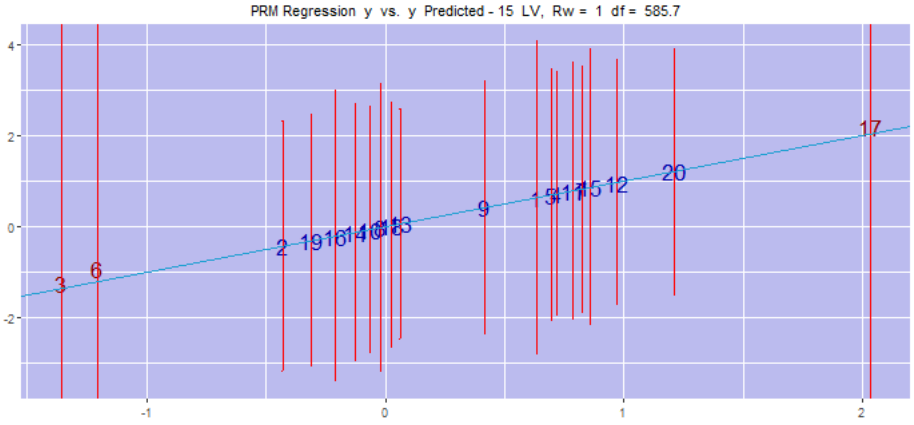
Obrázek 4.6 zobrazuje hodnoty vysvětlované proměnné a predikované hodnoty fitovaného modelu spolu s konfidenčními intervaly pro každé pozorování. Vidíme, že pro 3., 5. a 17. pozorovanou hodnotu jsou intervaly spolehlivosti dosti široké, což je ve statistické analýze nežádoucí. Tato pozorování jsou identifikovaná jako odlehlé hodnoty.

Řídká metoda PLS (SPLS)

V tomto okamžiku přecházíme k aplikaci řídké metody, která dokáže vybrat vhodný počet proměnných z datového souboru a také disponuje dobrými predikčními vlastnostmi. Z teoretické části víme, že u této metody je zapotřebí



Obrázek 4.5: Robustní PLS - biplot



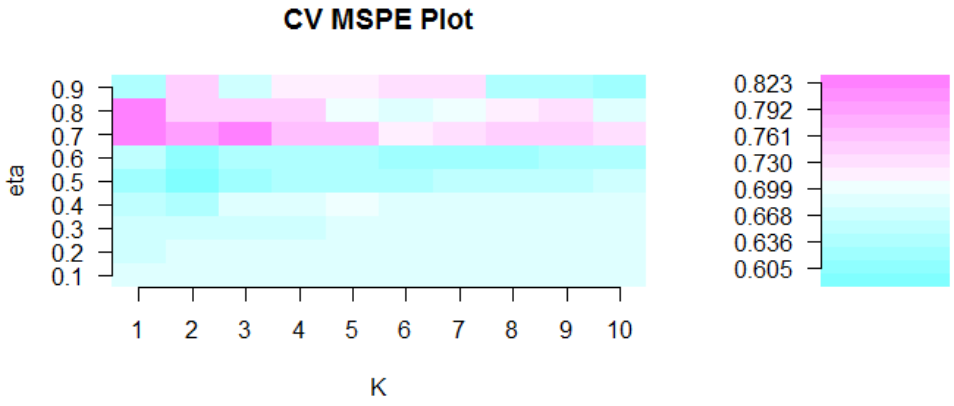
Obrázek 4.6: Robustní PLS - vyrovnané hodnoty

vhodně zvolit prahový parametr, který se objevuje ve funkcích knihovny `spls` pod názvem eta. V sekci 3.1.2 jej značíme symbolem η, také si připomeneme, že tento parametr může nabývat hodnot v intervalu [0,1]. Ladící parametr, jak uvidíme dále, nemusíme zadávat jako fixní hodnotu, naopak lze pomocí křížové validace zvolit takový, který zapříčiní nejnižší hodnotu MSEP.

```
> library(spls)
> fitspls<-spls(X, y, K=8, eta = 0.7)
> fitspls$fit
> print(fitspls)
> coef.f<-coef(fitspls)
> coefplot.spls(fitspls,xvar=c(1:4))
```

```
> pred.f<-predict(fitspls, type = "fit")
```

Pomocí příkazu `print(fitspls)` získáme výčet vybraných konkrétních nezávisle proměnných. Níže vidíme, že z 600 náhodně vygenerovaných vysvětlujících proměnných řídká metoda vybrala pouze 175. Jako postačující vybral tento algoritmus pouze dvě komponenty, tedy o jednu více než v modelu PLS. Na obrázku 4.7 pozorujeme, že pro vyšší hodnoty ladícího parametru dostáváme také vyšší hodnotu střední čtvercové chyby predikce. V našem případě optimální hodnota nabývá hodnoty 0,5.



Obrázek 4.7: SPLS - MSEP

```
Sparse Partial Least Squares for an univariate response
----
Parameters: eta = 0.7, K = 8
PLS algorithm:
pls2 for variable selection, simpls for model fitting

SPLS chose 175 variables among 600 variables

> cv<-cv.spls(X,y,K=c(1:10),eta=seq(0.1, 0.9, 0.1))
> cv$eta.opt
[1] 0.7
> cv$K.opt
> cv$mspemat[2]
```

Prozatím můžeme říci, že z pohledu charakteristiky RMSEP nerobustní řídká metoda PLS nabývá horších hodnot tohoto ukazatele než robustní PLS, což může být dáno větší citlivostí na odlehlé hodnoty. Naproti tomu, SPLS disponuje nižší hodnotou RMSEP než PLS.

Řídká metoda robustní PLS (SPRM)

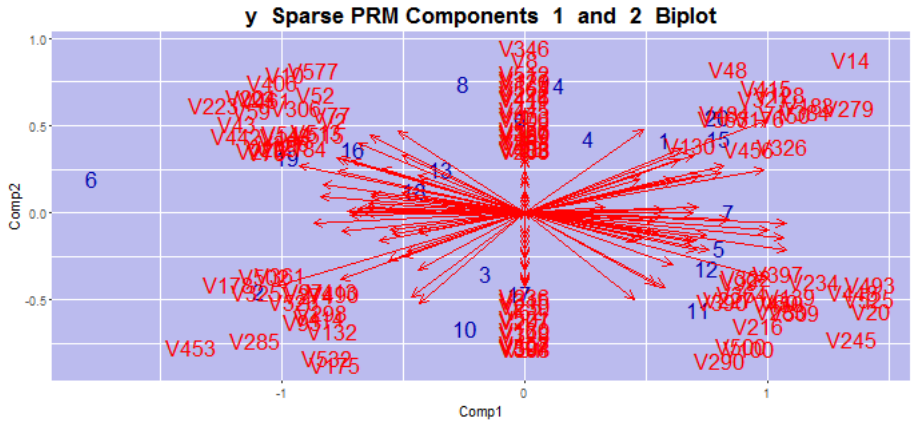
Nakonec přichází aplikace řídké metody na robustní PLS. K modelování byla použita opět knihovna `sprm`, ve které jsme tentokrát aplikovali stěžejní funkci pro tuto metodu `sprms`. Při využití níže uvedených příkazů dostáváme opět optimální počet komponent a odpovídající počet vysvětlujících proměnných.

```
> sprm <- sprms(y~., data=d, a=5, eta=0.5, fun="Hampel")
> y_pred <- sprm$fitted.values

Sparse partial M-robust regression
 Number of components:  2
 Sparsity parameter: eta = 0.6
 weight function:  Hampel with cutoff 0.95 0.975 0.999

Number of variables included in the model:
 With 1 component : 58
 With 2 components: 100

Percentage of explained variance
                              X        y
With  1  component(s): 3.146937 87.67472
With  2  component(s): 5.771702 87.72000
```

Zobrazené výsledky v softwaru R ukazují, že optimální počet komponent je opět pouze 2. Tedy pouze dvě skryté latentní proměnné nahrazují 600 původních proměnných, přičemž bylo tímto algoritmem bylo vybráno 100 relevantních proměnných.

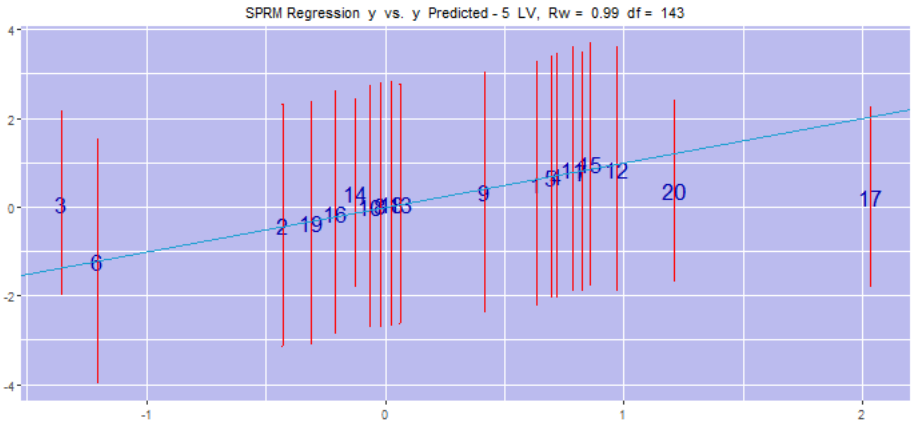
Na obrázku 4.8 vidíme, jak jednotlivé proměnné jsou mezi sebou korelovány a také rozdělení do shluků proměnných. Vertikálně jsou zřejmě zobrazeny proměnné, které do modelu nakonec nevstupují. Například šesté pozorování je vzhledem

k vysvětlované proměnné charakterizováno velkými hodnotami vysvětlujících proměnných v levé části biplotu.



Obrázek 4.8: SPRM - biplot

Také obrázek 4.9 známe již z analýzy předchozí metody. Jedná se o vyrovnané neboli predikované hodnoty závisle proměnné. Stejně jako v robustní PLS metodě jsou totožně odhalena odlehlá pozorování. Liší se však délkou konfidenčních intervalů, které jsou v tomto případě užší, tedy vykazují vyšší přesnost predikovaných hodnot.

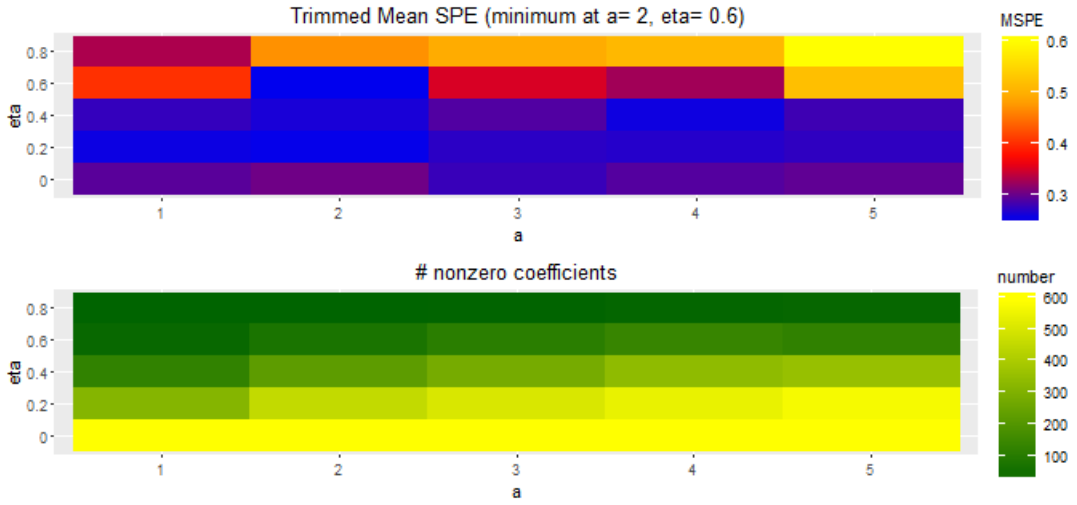


Obrázek 4.9: SPRM - vyrovnané hodnoty

Předposlední obrázek 4.10 patří mezi stěžejní, jelikož ukazuje právě vhodný počet komponent a také vhodnou volbu ladícího parametru; toto nastavení plyne z použiti křížové validace. Parametr řídkosti se liší oproti SPLS pouze o jednu

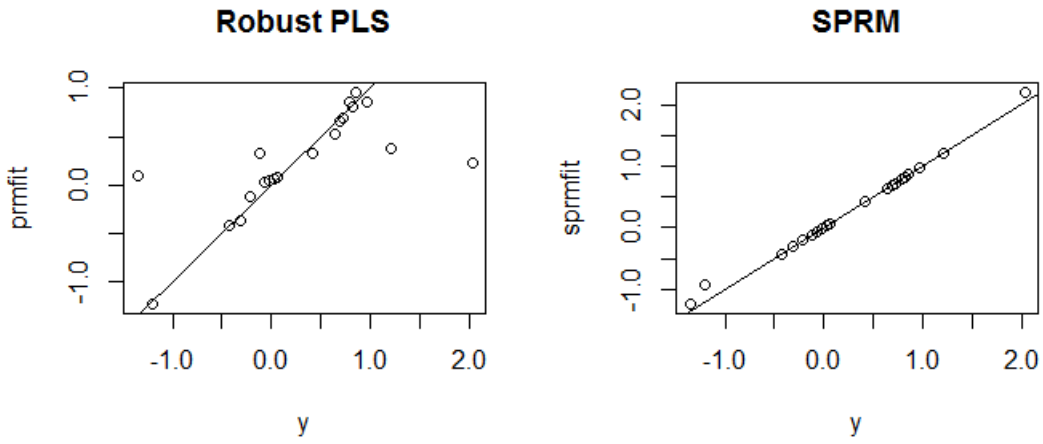
desetinu, kde optimální hodnota parametru nabývá 0,6, což nám také znázorňuje příšlušný obrázek 4.10. Můžeme vidět, že ladící parametr při dvou zvolených komponentách odpovídá nejnižší hodnotě charakteristiky MSEP. Dále lze vypozorovat na spodním obrázku 4.10, že čím nižší η parametr, tím více obdržíme v modelu nenulových parametrů.

```
> sprmcv <- sprmsCV(y~., data=d, as=1:5, etas=seq(0,0.9,0.2),
  nfold=5, fun="Hampel", prec=0.1, plot=TRUE)
> summary(sprmcv$opt.mod)
```



Obrázek 4.10: SPRM - MSEP

Na závěr srovnáme dvě metody, které patří mezi kandidáty s nejlepší predikční schopností, čímž je dle očekávání robustní PLS a SPRM. K tomuto srovnání pomůže jak obrázek 4.11, tak i tabulka 4.1. Vidíme, že vyrovnané hodnoty více přiléhají regresní přímce v modelu SPRM, čemuž odpovídá i menší hodnota RMSEP, než je tomu u robustní PLS metody. Proto můžeme závěrem říci, že metoda SPRM „vyhrává“ z pohledu této charakteristiky.



Obrázek 4.11: Srovnání robustní PLS a SPRM metody

Model	RMSEP	MSEP	komponenty
PLS	0.7875	0.6203	1
Robustní PLS	0.7703	0.5933	4
SPLS	0.7782	0.6056	2
SPRM	**0.6624**	**0.4388**	**2**

Tabulka 4.1: Srovnání modelů

.

4.1.2 Náhodná data s odlehlými hodnotami

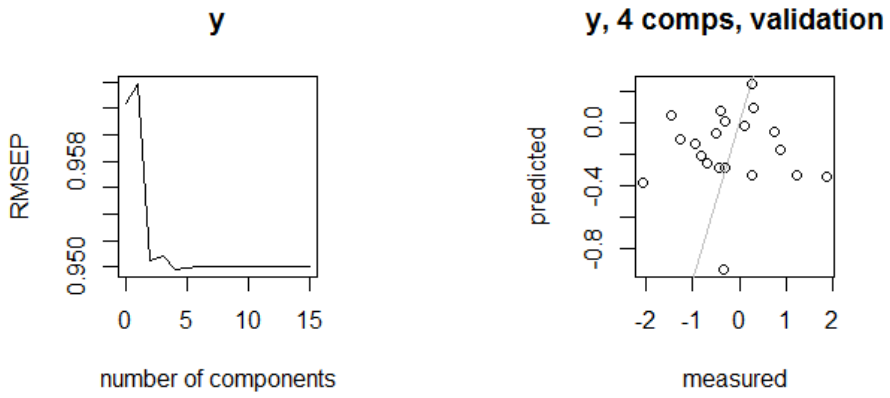
Nyní budeme zkoumat chování příslušných modelů v případě, kdy je zahrnuto v datech určité procento odlehlých hodnot. Nahradíme tedy např. 6 pozorování odlehlými hodnotami, což docílíme tak, že pro vygenerování náhodných hodnot z normálního rozdělení zvolíme odlišnou matici Sigma. Dostáváme tak nový datový soubor s 30 % odlehlých pozorování.

```
> Sigma<-matrix(0.5,600,600)
> diag(Sigma)<-2
> B<-mvrnorm(n=6,rep(0,600),Sigma)
> S<-2
> y1<-mvrnorm(n=6,rep(0,1),S)
#nahrazení v matici plánu
A<-X[-c(15:20),]
X<-rbind(A,B)
#nahrazení 6 pozorování ve vektoru vysvětlované proměnné
d <-y[c(1:14)]
> for (i in 1:length(d)){
  y1[length(y1)+1] <- d[i]  #přidání hodnot k vektoru d
  }
> y<-y1
```

Analýzu dat s odlehlými hodnotami provedeme analogicky jako v předchozí kapitole, proto budou uvedeny pouze výsledné charakteristiky modelů s příslušnými grafy. Opět začneme s klasickou metodou dílčích nejmenších čtverců, tedy bez aplikace robustnosti a řídkosti. Na obr. 4.12 vidíme, že predikce pro optimální počet komponent (4) zde nevykazuje vysokou kvalitu. Ve srovnání s daty bez výskytu odlehlých hodnot se tato metoda nevyznačuje stabilitou dat. Také index determinace je dle očekávání nepatrně nižší 0,9715 oproti původnímu modelu bez outlierů 0,9951.

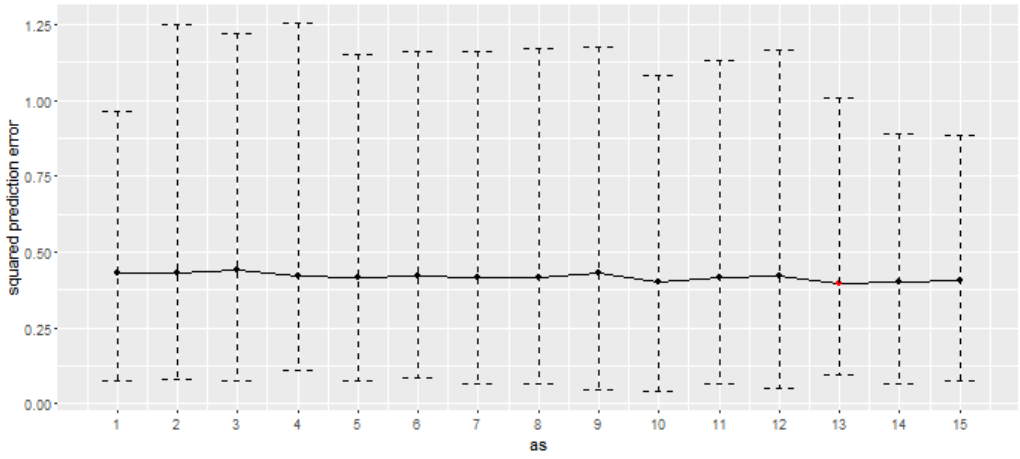
```
> RMSEP(pls30,ncomp = 4,intercept = FALSE)
> MSEP(pls30,ncomp = 4,intercept = FALSE)
> R2(pls30,ncomp=4,intercept = FALSE)
[1]  0.9715
```

Pokračujeme s robustní PLS metodou, od které očekáváme menší citlivost právě na odlehlá pozorování. V softwaru R jsme si opět vyčíslili optimální počet



Obrázek 4.12: Křížová validace a kvalita predikce

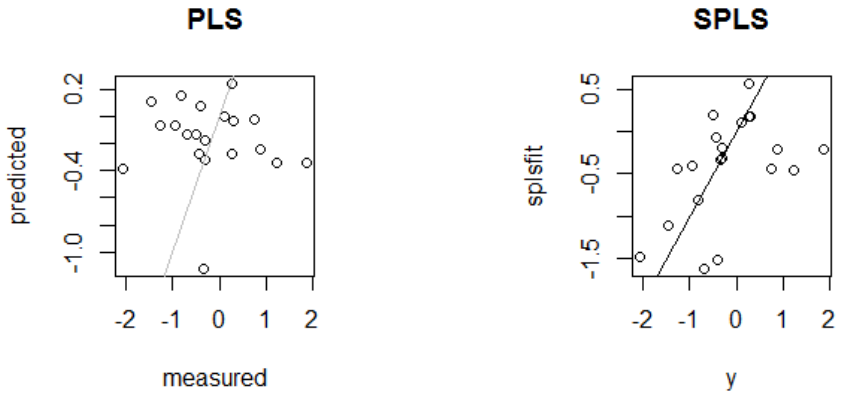
komponent - 13, který se v tomto případě dosti liší od původního robustního PLS modelu, kde pomocí křížové validace byly vybrány pouze 4 umělé proměnné. Nicméně hodnota charakteristiky SEP je vyšší v robustním PLS modelu obsahujícího odlehlá pozorování, a to 0,3796 oproti robustnímu PLS modelu bez odlehlých hodnot.



Obrázek 4.13: SEP pro různý počet komponent

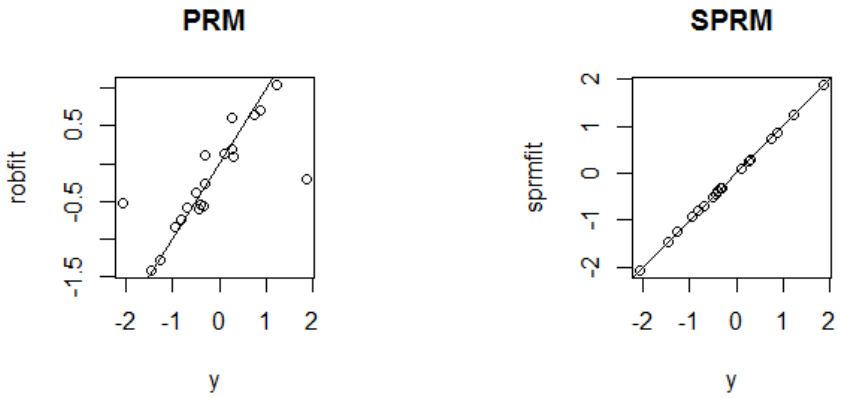
Další metodou v pořadí je řídká PLS metoda, neboli aplikace řídkosti na klasickou PLS metodu. SPLS metoda zde nevykazuje příliš vysokou kvalitu predikce, což je dáno právě zahrnutým velkým procentem odlehlých hodnot v datovém souboru; je zřejmé z obrázku 4.14.

```
> cv<-cv.spls(X,y,K=c(1:10),eta=seq(0.1, 0.9, 0.1))
> cv$eta.opt      # [1] 0.4
> cv$K.opt        # [1] 2
> cv$mspemat[5]   # [1] 1.03908
```



Obrázek 4.14: Srovnání PLS a SPLS modelů

Nakonec si představme poslední z metod, a to řídkou robustní metodu SPRM. Můžeme ji obecně považovat za nejvíce stabilní a s nejlepšími predikčními schopnostmi. Závěrem lze konstatovat, že se opět potvrdila hypotéza, že aplikace řídkosti spolu s robustifikací dosahují nejvyšší kvality modelu, což pěkně dokládá srovnání na obrázku 4.15.



Obrázek 4.15: Srovnání PRM a SPRM modelů

4.2 Aplikace na reálná data

Pro snadnější uchopení interpretace regresní analýzy pomocí PLS představíme reálná data poskytnutá Ústavem translační a molekulární medicíny Lékařské fakulty UP. Jedná se o data měřená pomocí necílené analýzy. K dispozici je tak 273 metabolitů, které ovšem nejsou identifikovány, proto nejsou označeny názvy, ale jen číselnými identifikátory. Dále jsou zde pouze dvě skupiny vzorků - 25 pacientů s jednou konkrétní nemocí a 25 provedených kontrol. Kvůli dodržení rozsahu práce a také citlivosti dat uvádím pouze vybrané metabolity pro prvních 6 kontrol (Control) a posledních 6 pacientů s konkrétním onemocněním (MCADD).

```
     sample X72.080966 X74.096124 X78.212489 X84.044983 X86.059806 X86.096225
1 Control01  0.6280626  0.2174143  1.0223077  0.1212418  0.7882556 0.77879643
2 Control02  0.8248765  0.2174143  1.1201980  1.1214116  0.7098177 0.92405315
3 Control03  0.0984793  0.9287449  0.5145475  0.1212418  1.4473355 0.00235655
4 Control04  0.6131866  1.1191734  0.9132367  0.5039025  0.7053937 0.78044438
5 Control05  0.8729618  0.2174143  1.0542507  0.8564464  1.3530839 0.96035162
6 Control06  0.5612812  1.6765328  0.9675868  0.4514951  1.8164352 0.72986949

     sample X72.080966 X74.096124 X78.212489 X84.044983 X86.059806 X86.096225
45 MCADD020  1.2129649  1.9135055  1.0050437  0.1141646  0.6103507  0.9129539
46 MCADD021  0.9898026  1.0786420  1.0002946  0.5780141  0.9479815  0.9987213
47 MCADD022  1.0504486  0.4099087  0.9465702  0.6229320  0.7855707  0.8935574
48 MCADD023  0.7385370  0.3579466  0.7984062  0.4537749  0.8581424  0.6819085
49 MCADD024  0.6901536  0.5913333  0.9574624  0.6745181  0.8730102  0.8011072
50 MCADD025  0.7704024  1.4157553  0.6599329  0.3251013  0.9149823  0.6645886
```

Pro další analýzu je vhodné data centrovat a také použít logpodílovou (angl. *logratio*) transformaci, kterou umožňuje funkce `clr` v knihovně `Hotelling`. Jelikož se v datech vyskytuje vysvětlovaná proměnná v klasifikační formě, využijeme zde pro metodu PLS a její modifikace varianu diskriminační analýzy, kterou jsme si nastínili již v teoretické části.

PLS-DA

Pro tuto implementaci se nabízí knihovna `DiscriMiner`; konkrétně použijeme funkci `plsDA`. Možností, jak nastavit výsledný model, je hned několik. Například si můžeme zvolit typ křížové validace, jež vybírá vhodný počet komponent.

```
# PLS-DA s konkrétním počtem hlavních komponent
```

```
> plsda1= plsDA(skupiny[,2:274], skupiny$sample, autosel=FALSE, comps=4)
> plsda1$error_rate
> plsda1$error_rate
[1] 0
```

V tomto příkladě budeme výhradně posuzovat kvalitu predikce dle indexu determinace R^2 nejprve si vyzkoušíme, jaké hodnoty této charakteristiky dosahuje model se 4 zvolenými komponentami. Vidíme, že $R^2 = 0{,}9657$. Máme zde však i možnost automatického výběru vhodného počtu hlavních komponent, přičemž pro 37 vybraných latentních proměnných dostáváme vyšší hodnotu indexu determinace $R^2 = 0{,}9711$. Připomeňme si, že preferujeme vyšší hodnotu před nižší.

```
#vysvětlení variability, index determinace
> plsda1$R2
          R2X    R2Xcum        R2Y    R2Ycum
t1 0.28769208 0.2876921 0.70806113 0.7080611
t2 0.16132245 0.4490145 0.18521233 0.8932735
t3 0.08381520 0.5328297 0.03937027 0.9326437
t4 0.04794809 0.5807778 0.03311391 0.9657576

> plsda1$Q2
   Q2.Control   Q2.MCADD   Q2.global
t1 0.65882087 0.65882087  0.65882087
t2 0.55350845 0.55350845  0.55350845
t3 0.06589265 0.06589265  0.06589265
t4 0.21834608 0.21834608 -0.53754882

# PLS-DA s automatickým výběrem hlavních komponent
> plsda2= plsDA(skupiny[,2:274], skupiny$sample, autosel=TRUE)
> my_pls2$R2
            R2X    R2Xcum          R2Y    R2Ycum
t1  0.287692076 0.2876921 7.080611e-01 0.7080611
t2  0.161322447 0.4490145 1.852123e-01 0.8932735
t3  0.083815199 0.5328297 3.937027e-02 0.9326437

t34 0.003233589 0.8036401 4.001196e-24 0.9711774
t35 0.002420401 0.8060605 5.363412e-25 0.9711774
t36 0.002361267 0.8084218 3.316017e-26 0.9711774
t37 0.002281301 0.8107031 6.133955e-28 0.9711774

# PLS-DA s validací za použití testovací a trénovací množiny
> learning = c(1:15, 26:40)
```

```
> testing = c(16:25,41:50)
> plsda3 = plsDA(skupiny[,2:274], skupiny$sample, validation="learntest",
                learn=learning, test=testing)
> plsda3$error_rate
[1] 0.15
```

Vyzkoušíme-li odlišný typ křížové validace, konkrétně LKO, obdržíme 36 hlavních komponent s příslušným koeficientem $R^2 = 0{,}9811$, což je vyšší než u předchozí validace.

```
# diskriminační analýza s leave-K fold-Out validací
> plsda4= plsDA(skupiny[,2:274], skupiny$sample, validation = "NULL",
  cv ="LKO")
> plsda4$R2
            R2X    R2Xcum          R2Y    R2Ycum
t1  0.287692076 0.2876921 7.080611e-01 0.7080611
t36 0.002281301 0.8261901 6.133955e-28 0.9811026
```

Robustní metoda PLS-DA

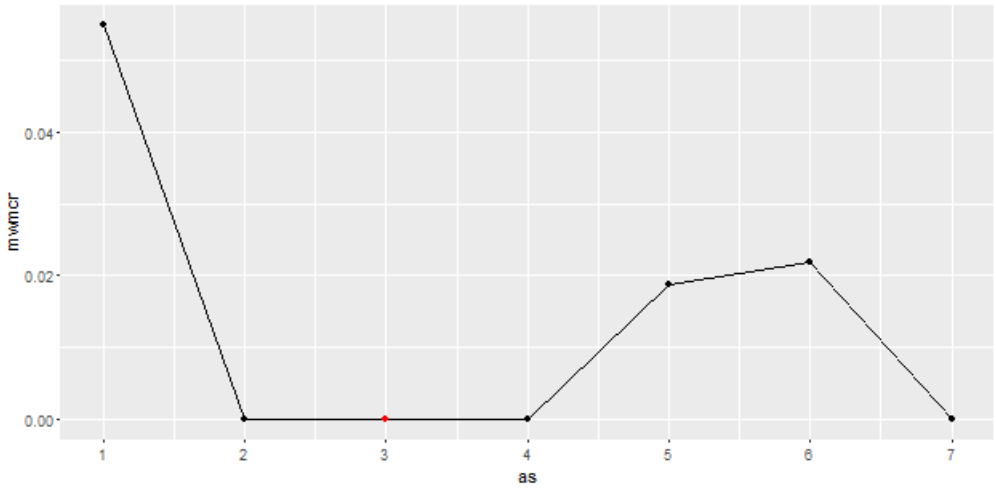
Nyní využijeme již známé knihovny `sprm`, jež jsme použili v prvním příkladu na simulovaných datech. Knihovna nabízí analogii funkcí právě i pro diskriminační analýzu.

```
> rplsda<-prmda(sample~.,skupiny,a=4, fun = "Hampel",class = "regfit")
> summary(rplsda)
Partial M-robust regression
 Number of components:  4
 weight function:  Hampel with cutoff 0.95 0.975 0.999
```

Aplikujeme-li robustnost na metodu PLS při zvolených 4 komponentách, oproti klasické PLS-DA metodě dostáváme nižší index determinace. Pomocí křížové validace opět vygenerujeme vhodný počet komponent, v tomto případě nám obrázek 4.16 ukazuje právě počet 3.

```
Percentage of explained variance
                              X        y
With  1  component(s): 33.29739 83.09572
With  2  component(s): 48.36772 91.76054
```

```
With  3  component(s): 52.90090 95.30175
With  4  component(s): 59.28112 96.88999

> plot(rplsda,type = "weights")
> plot(rplsda,type = "yyp")
> plot(rplsda, type="dd")
> predik_rplsda<-table(skupiny$sample,predict(rplsda))
> plot(predik_rplsda)

# křížová validace pro robustní PLS-DA metodu
> rplsda_cv <- prmdaCV(sample~.,skupiny, as=1:5, fun="Hampel",
  class="regfit", numit=10, prec=0.1)
> rplsda_cv$opt.mod
> mod$opt.mod
```



Obrázek 4.16: Robustní PLS-DA - křížová validace

Řídká metoda SPLS-DA

Nyní si představíme aplikaci řídkosti na model PLS-DA. Využijeme následujících příkazů v softwaru R. Ve výstupu níže uvedených příkazů vidíme, že model SPLS-DA vybral 30 relevantních proměnných z 273 možných.

```
> modelspls<-splsda(x,skup_x$sample,K=3, eta=0.8, scale.x=FALSE)
> print(modelspls)
```

```
Sparse Partial Least Squares Discriminant Analysis
----
Parameters: eta = 0.8, K = 3
Classifier: Linear Discriminant Analysis (LDA)

SPLSDA chose 30 variables among 273 variables

Selected variables:
x.35  x.71  x.75  x.78  x.79
x.80  x.93  x.146 x.148 x.150
x.151 x.153 x.154 x.155 x.156
x.157 x.188 x.208 x.210 x.216
x.219 x.222 x.231 x.238 x.250
x.252 x.256 x.259 x.262 x.264
```

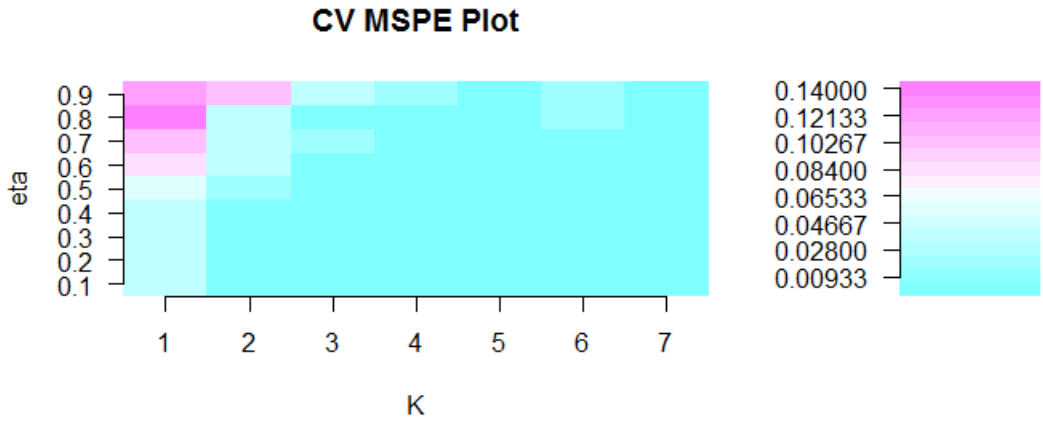
Dalším postupem v analýze je opět využít křížové validace k určení vhodné volby ladícího parametru η a také vhodného počtu komponent. Na základě těchto optimálních parametrů je následně spočítána charakteristika kvality predikce MSEP, což nám dokládá obr. 4.17.

```
Optimal parameters: eta = 0.9, K = 5
> cv$err.mat
NULL
> cv$eta.opt
[1] 0.9
> cv$K.opt
[1] 5
```

Řídká robustní metoda SPRM-DA

Na závěr se dostáváme k aplikaci řídkosti na robustní PLS-DA metodu. Kromě vhodného výběru počtu komponent, vygeneruje také relevantní proměnné, které nejvíce přispívají modelu. Na obr. 4.18 vidíme biplot, který znázorňuje, že proměnné 33, 75, 133 a 175 mají největší význam. Dále konkrétně proměnná x.157 má největší vliv na 32. pozorování.

```
> srplsda<-sprmda(sample~.,skslovy,a=4,eta= 0.7,class = "regfit")
> biplot.sprmda(srplsda)
> summary(srplsda)
```

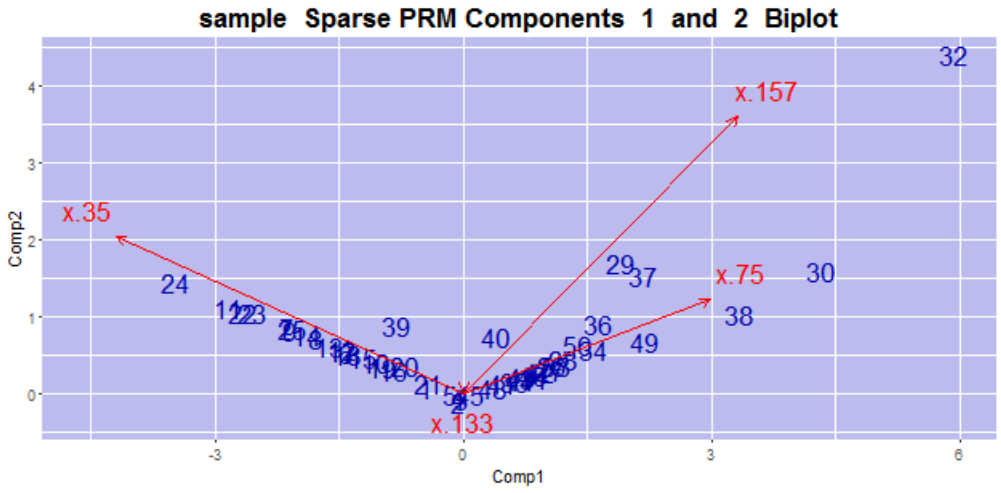


Obrázek 4.17: SPLS-DA - křížová validace

```
Sparse partial M-robust regression
 Number of components:  4
 Sparsity parameter: eta = 0.7
 weight function:  Hampel with cutoff 0.95 0.975 0.999

Number of variables included in the model:
 With 1 component : 3
 With 2 components: 4
 With 3 components: 4
 With 4 components: 6

Percentage of explained variance
                              X        y
With  1  component(s): 12.69110 70.78669
With  2  component(s): 21.55533 71.24784
With  3  component(s): 26.10395 75.53153
With  4  component(s): 31.86951 76.91002
```

```
> srplsda_cv<- sprmdaCV(sample~.,skslovy, as=1:5, etas=seq(0.1,0.9,0.1),
                nfold=5,class="regfit", numit=10, prec=0.1)
```

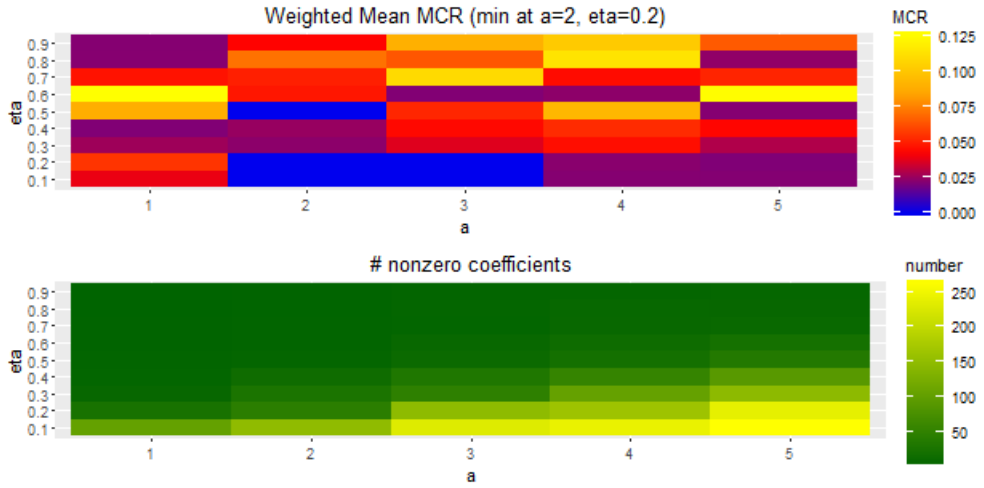
Poslední z grafů (obr. 4.19) ukazuje vhodnou volbu ladícího parametru. Díky použití křížové validace dostáváme eta = 0,2 a při zvolených dvou komponentách.



Obrázek 4.18: SPRM-DA - biplot

Níže položený graf nám ukazuje, jak souvisí výběr ladícího parametru η s výběrem počtu hlavních komponent na počet nenulových regresních parametrů. Použijeme-li hodnotu eta = 0,2 a 6 hlavních komponent, budeme mít v modelu přibližně 200 nenulových regresních koeficientů.

Na závěr můžeme říci, že aplikace zároveň robustní i řídké metody na metodu PLS (ve variantě klasifikační analýzy) dává nejlepší predikční vlastnosti, kde tomu tak bylo i při aplikaci na simulovaná data. Řídká metoda je mimo jiné výhodná z pohledu výběru relevantních proměnných.



Obrázek 4.19: SPRM-DA - křížová validace

Závěr

V práci jsem se zabývala regresní analýzou pomocí metody nejmenších dílčích čtverců. Jedná se o metodu, která se především využívá k analýze v situacích, kdy datový soubor obsahuje více proměnných než pozorování. S ohledem na dosud publikovanou literaturu o tomto tématu, bylo cílem práce poskytnout ucelený přehled této metody a jejich modifikací s následným využitím aplikačního potenciálu na konkrétních datech. V neposlední řadě bylo záměrem práce též srovnání jednotlivých modelů z pohledu kvality predikce.

V první kapitole teoretické části jsem představila metodu nejmenších dílčích čtverců, princip, na kterém pracuje, její přednosti a také jednotlivé algoritmy. V následujících dvou kapitolách byla metoda dílčích nejmenších čtverců robustifikována a bylo použito možností zjednodušení interpretace regresních parametrů zahrnutím podmínky řídkosti. V poslední kapitole mé diplomové práce byly zmíněné metody demonstrovány na simulovaných i reálných datech pomocí statistického softwaru R. Praktická část především sloužila ke srovnání modelů a k jejich následnému vyhodnocení.

Jsem přesvědčena, že se nejednalo o jednoduchou problematiku, a tak značným přínosem práce je především komplexní popsání jednotlivých metod a poskytnutí přehledu o jejich využití. Tyto metody jsou i v nynější době stále ve vývoji, nebylo tedy vždy snadné najít jednotný přístup k popisu jejich vlastností. Čtenář má také možnost nahlédnout do prostředí softwaru R společně s detailními popisy použitých funkcí a příkazů pro praktické použití. Metoda dílčích nejmenších čtverců má velký aplikační potenciál, proto jsem ráda za příležitost využít ji k diagnostikování metabolitů ovlivňující konkrétní onemocnění.

Při psaní této práce jsem nabyla hodně zkušeností, především díky zdokonalení orientace v cizojazyčné literatuře a prohloubení si dovedností programování v softwaru R. Práce představovala v jistých okamžicích i úskalí, a to konkrétně v kapitole o aplikaci řídkosti, o které je minimum literatury, a tak bylo obtížné

najít české ekvivalenty k těm anglickým. Přesto doufám, že se hlavní cíl práce, podat čtenářům ucelený přehled o možných přístupech k regresnímu modelování pomocí metody dílčích nejmenších čtverců, podařilo naplnit.

Literatura

[1] CHUN, H., KELES, S., *Sparse Partial Least Squares Regression for Simultaneous Dimension Reduction and Variable Selection*, Department of Statistics, University of Wisconsin, Madison, USA, 2007.

[2] EVERITT, B., HOTHORN, T., *An Introduction to Applied Multivariate Analysis with R*, Springer New York Dordrecht Heidelberg London, 2011.

[3] FILZMOSER, P. *Robust Statistics: Theoretical and Practical Considerations.* Department of Statistics and Probability Theory, University of Technology, Austria, 2013.

[4] GARCIA, H., FILZMOSER, P. *Multivariate Statistical Analysis using the R package chemometrics.* Department of Statistics and Probability Theory, University of Technology, Austria, 2016.

[5] GILL, P., MURRAY, W., WRIGHT, M. *Practical Optimization.* New York: Academic Press, 1981.

[6] HAMPEL, F., RONCHETTI, E., ROUSSEEUW, P., STAHEL, W. *Robust Statistics: the approach based on influence functions.* Wiley, 1986.

[7] HUBER, P. J. Robust Estimation of a Location Paramete. *Annals of Statistics*, 73–101, 1964.

[8] HOFFMANN, I., SERNEELS, S., FILZMOSER, P. *Sparse partial robust M regression.* Faculty of Economics and Business, KU Leuven, 2015.

[9] JUREČKOVÁ, J.: *Robustní statistické metody.* Nakladatelství: Karolinum, Praha 2001.

[10] LEE, D., LEE, W., LEE, Y., PAWITAN, Y. Sparse partial least-squares regression and its applications to high-throughput data analysis. *Journal of Chemometrics and Intelligent Laboratory Systems 109*, 1-8, 2011.

[11] NAES, T., ISAKSSON, T., FEARN, T., DAVIES, T. *A user-friendly guide to Multivariate Calibration and Classification.* NIR Publications, Chichester, UK, 2004

[12] NAJDEKR, L., GARDLO, A., MÁDROVÁ, L., FRIEDECKÝ, D., JANEČKOVÁ, H., CORREA, E. S., GOODACRE, R., ADAM T. Oxidizen phosphatidylcholines suggest oxidative stress in patients with medium-chain acyl-CoA dehydrogenase deficiency. *Journal of Talanta 139*, 62-66, 2015.

[13] POLAT, E., GUNAY, S. The comparison of Partial Least Squares regression, principal component regression and ridge regression. *Journal of Data Science 13*, 663-692, 2015.

[14] ROSIPAL, R., KRÄMER, N. *Overview and recent advances in partial least squares.*Austrian Research Institute for Artificial Intelligence, Vienna, Austria, 34-51, 2006.

[15] TRYGG, J., LUNDSTEDT, T., MADSEN, R. Chemometrics in metabolomics. *Journal of Proteome Research 6*, 469-479, 2007.

[16] VARMUZA, K., FILZMOSER, P. *Introduction to multivariate statistical analysis in chemometrics.* Boca Raton: CRC Press, 2009.

[17] WEHRENS, R. *Chemometrics with R.* Springer Heidelberg Dordrecht London New York, 2011.

[18] WOLD, H., BLALOCK H. M., AGANBEGIAN, A., BORODKIN, F. M., BOUDON, R. and V. CAPPECCHI. *PLS path models with latent variables: The nipals approach.* New York: Academic Press, 1975.

[19] WOLD, S., MARTENS, H., WOLD, H. *The multivariate calibration problem in chemistry solved by the PLS method.* Institute of Chemmistry, Umea University, Sweden, 1983.

[20] Classification with O-PLS-DA. [online]. [2013] [cit. 2016-10-15]. Dostupné z: https://www.r-bloggers.com/classification-with-o-pls-da/

[21] Hadamard product (matrices). [online]. [2016] [cit. 2016-10-16]. Dostupné z: https://en.wikipedia.org/wiki/Hadamard_product_(matrices)

[22] k-násobná křížová validace. [online]. [2016] [cit.2016-11-15]. Dostupné z: http://portal.matematickabiologie.cz/index.php?pg=analyza-a-hodnoceni-biologickych-dat–vicerozmerne-metody-pro-analyzu-dat–klasifikace–hodnoceni-uspesnosti-klasifikace–k-nasobna-krizova-validace

[23] Predictive modeling, O-PLS. [online]. [2015] [cit. 2016-10-22]. Dostupné z: https://github.com/dgrapov/TeachingDemos/tree/master/Demos/Predictive%20Modeling/O-PLS

[24] The R Project for Statistical Computing. [online]. [2016] [cit.2016-06-04]. Dostupné z: https://www.r-project.org/