MASARYKOVA UNIVERZITA
PŘÍRODOVĚDECKÁ FAKULTA
ÚSTAV MATEMATIKY A STATISTIKY

Diplomová práce

BRNO 2019 SIMONA KOLENČÍKOVÁ

MASARYKOVA
UNIVERZITA
PŘÍRODOVĚDECKÁ FAKULTA
ÚSTAV MATEMATIKY A STATISTIKY

Matice a jejich použití ve statistice

Diplomová práce

Simona Kolenčíková

Vedoucí práce: Mgr. Jiří Zelinka, Dr. **Brno 2019**

Bibliografický záznam

Autor:	Bc. Simona Kolenčíková Přírodovědecká fakulta, Masarykova univerzita Ústav matematiky a statistiky
Název práce:	Matice a jejich použití ve statistice
Studijní program:	Matematika
Studijní obor:	Statistika a analýza dat
Vedoucí práce:	Mgr. Jiří Zelinka, Dr.
Akademický rok:	2018/2019
Počet stran:	xi + 47
Klíčová slova:	singulární rozklad, varianční matice, analýza hlavních komponent, lineární regrese, klasifikace, lineární diskriminační analýza

Bibliografický záznam

Autor:	Bc. Simona Kolenčíková Prírodovedecká fakulta, Masarykova univerzita Ústav matematiky a štatistiky
Názov práce:	Matice a ich použitie v štatistike
Študijný program:	Matematika
Študijný odbor:	Štatistika a analýza dát
Vedúci práce:	Mgr. Jiří Zelinka, Dr.
Akademický rok:	2018/2019
Počet strán:	xi + 47
Kľúčové slová:	singulárny rozklad, variančná matica, analýza hlavných komponent, lineárna regresia, klasifikácia, lineárna diskriminačná analýza

Bibliographic Entry

Author:	Bc. Simona Kolenčíková Faculty of Science, Masaryk University Department of Mathematics and Statistics

Title of Thesis:	Matrices and their application in statistics
Degree Programme:	Mathematics
Field of Study:	Statistics and Data Analysis
Supervisor:	Mgr. Jiří Zelinka, Dr.
Academic Year:	2018/2019
Number of Pages:	xi + 47
Keywords:	singular value decomposition, covariance matrix, principal component analysis, linear regression, classification, linear discriminant analysis

Abstrakt

V této diplomové práci se věnujeme maticím a jejich použití ve statistice. V první části je čtenář seznámen s diagonalizací a singulárnim rozkladem matic. Následně je ukázaná aplikace této teorie v metodě hlavních komponent a v praxi na reálných datech. V druhé části je představena metoda nejmenších čtverců a její zobecnění – hřebenová regrese a metoda Lasso, kterou lze odvodit v kontextu robustní optimalizace. Následně je ukázána souvislost mezi těmito zobecněními a metodou nejmenších čtverců. Teoretické výsledky doplňuje simulační studie. Třetí část se zabývá problémem klasifikace a metodami jejího řešení. V závěru se dočteme o lineární diskriminační analýze a její odlišné geometrické interpretaci vůči analýze hlavních komponent.

Abstrakt

V tejto diplomovej práci sa venujeme maticiam a ich používaniu v štatistike. V prvej časti je čitateľ zoznámený s diagonalizáciou a singulárnym rozkladom. Následne je ukázaná táto teória v metóde hlavných komponent a v praxi na reálnych dátach. V druhej časti je predstavná metóda najmenších štvorcov a jej zovšeobecnenie – hrebeňová regresia a metóda Lasso, ktoré je možné odvodiť v kontexte robustnej optimalizácie. Následne je ukázaná súvislosť medzi týmito zovšeobecneniami a metódou najmenších štvorcov. Teoretické výsledky doplňuje simulačná štúdia. Tretia časť sa zaoberá problémom klasifikácie a metódami jej riešenia. V závere sa dočítame o lineárnej diskriminačnej analýze a jej odlišnej geometrickej interpretácii voči analýze hlavných komponent.

Abstract

In this thesis we study matrices and their application in statistics. In the first chapter is the reader familiarised with terms such as matrix diagonalization and singular value decomposition. This theory is pointed out in principal component analysis and applied on real data. The second chapter introduces least square method for regression and its generalisation – ridge regression and Lasso, which can be derived in context of robust optimalization. Then the context between the generalisation and initial method is shown. Theory is complemented with simulation study. The third chapter is dedicated to classification and related methods. Lastly, we read about linear discriminant analysis and its geometrical approach compared to principal component analysis.

MASARYKOVA UNIVERZITA
Přírodovědecká fakulta

ZADÁNÍ DIPLOMOVÉ PRÁCE

Akademický rok: 2018/2019

Ústav: Ústav matematiky a statistiky

Studentka: Bc. Simona Kolenčíková

Program: Matematika

Obor: Statistika a analýza dat

Ředitel *Ústavu matematiky a statistiky* PřF MU Vám ve smyslu Studijního a zkušebního řádu MU určuje diplomovou práci s názvem:

Název práce: Matice a jejich použití ve statistice

Název práce anglicky: Matrices and their application in statistics

Oficiální zadání:
V práci se zaměřte na maticové metody používané ve statistice. Pozornost věnujte zejména speciálním typům matic, se kterými se při statistickém zpracování dat setkáváme. Práce by měla obsahovat příklady použití prezentovaných metod na simulovaných i reálných datech.

Literatura:
SEARLE, Shayle R. *Matrix algebra useful for statistics*. New York: John Wiley & Sons, 1982. 438 s. ISBN 0471866814.

FIEDLER, Miroslav. *Speciální matice a jejich použití v numerické matematice*. 1. vyd. Praha: SNTL - Nakladatelství technické literatury, 1981. 272 s.

James R. Schott, Matrix Analysis for Statistics, John Wiley & Sons, 2016, 552 s., ISBN: 978-1-119-09248-3

Jazyk závěrečné práce: slovenština

Vedoucí práce: Mgr. Jiří Zelinka, Dr.

Datum zadání práce: 2. 10. 2017

V Brně dne: 25. 4. 2019

Souhlasím se zadáním (podpis, datum):

.....	.....	.....
Bc. Simona Kolenčíková	Mgr. Jiří Zelinka, Dr.	prof. RNDr. Jan Slovák, DrSc.
studentka	**vedoucí práce**	**ředitel Ústavu matematiky a statistiky**

Poďakovanie

Na tomto mieste by som sa chcela veľmi pekne poďakovať svojmu vedúcemu práce Mgr. Jiřímu Zelinkovi, Dr. za jeho pomoc a usmernenie pri písaní. Ďalej za jeho čas na pravidelné stretnutia, cenné rady, informácie a v neposlednom rade za trpezlivosť a ochotu.

Prehlásenie

Prehlasujem, že som svoju diplomovú prácu vypracovala samostatne s využitím informačných zdrojov, ktoré sú v práci citované.

Brno 25. apríla 2019

. . . . . .

Simona Kolenčíková

Obsah

Úvod **1**

Kapitola 1. Rozklady matíc **2**
1.1 Diagonalizácia matíc 2
1.2 Singulárny rozklad 3
1.3 Analýza hlavných komponent 5
1.3.1 Aplikácia v praxi 6

Kapitola 2. Lineárna regresia **14**
2.1 Pojmy a značenie 14
2.2 Metóda nejmenších štvorcov 15
2.3 Upravená regresia 16
2.3.1 Multikolinearita 16
2.3.2 Hrebeňová regresia 17
2.3.3 Metóda Lasso 22
2.4 Porovnanie metód 24
2.4.1 Simulačná štúdia s odľahlím pozorovaním 25
2.4.2 Simulačná štúdia v prítomnosti chyby merania 27
2.4.3 Simulačná štúdia v prítomnosti multikolinearity (1) 29
2.4.4 Simulačná štúdia v prítomnosti multikolinearity (2) 31

Kapitola 3. Klasifikačná analýza **34**
3.1 Vstupné dáta 34
3.1.1 Aplikácia v praxi 35
3.2 Proces klasifikácie 35
3.2.1 Test správnosti 36
3.3 Klasifikácia pomocou klasickej LDA 37
3.4 Vysokodimenzionálna klasifikácia 39
3.4.1 Pseudoinverzné matice 39
3.4.2 Klasifikácia pomocou LDA s pseudoinverziou 40
3.5 LDA verzus PCA 41
3.5.1 Aplikácia na reálnych dátach 42

Závěr **45**

Seznam použité literatury **46**

Úvod

Mnohorozmerná štatistika sa zaoberá analýzou javou, ktoré je možné vystihnúť množstvom premenných. Vzťahy vysvetľujúcich premenných sa zaznamenávajú napríklad do tzv. variančnej matice, s ktorou sa ďalej pracuje. Problém nastáva, ak je množstvo premenných príliš veľké a prevyšuje počet vzoriek skúmaného javu. Tieto prípady vedú k singularite použitých matíc, čo má za následok nepriaznivé numerické vlastnosti. Úloh vysokodimenzionálnej štatistiky je mnoho. Patrí sem lineárna a nelineárna regresia, logistická regresia, zhluková analýza, analýza hlavných komponent, klasifikácia a iné. Prehľad je k nahliadnutí v [7].

Redukcia dimenzie predstavuje dôležitý krok štatistickej analýzy mnohorozmerných dát. Pôvodný problém sa aproximuje na priestor nižšej dimenzie za účelom zníženia pamäťových nákladov s čo najmenšou stratou pôvodných informácií. Základom celej rady štatistických metód pre redukciu dimenzie je singulárny rozklad. Najpoužívanejšou metódou je analýza hlavných komponent, ktorej sa v práci venujeme.

Na základe analýzy dát sa vyvodzujú závery a rozhodnutia. Problém s reálnymi dátami je ten, že sú často postihnuté chybami merania alebo inými nepresnosťami. Zanedbanie hoci drobných nepresností môže viesť k odlišnej interpretácii a chybným rozhodnutiam. V regresnej analýze sa preto vyvinuli metódy, ktoré nájdu optimálne riešenie a sú odolné voči malým zmenám na vstupných dátach. Takéto optimálne riešenia nazveme robustné. V matematicke sa pojem robustnosť používa často a v rôznych disciplínach môže mať odlišný význam.

Práca je rozdelená do troch kapitol. V prvej kapitole sa venujeme rozkladu matíc, jeho významu a vlastnostiam. Pracujeme so všeobecným rozmerom reálnej matice, ktorá môže mať aj zlé numerické vlastnosti. Podrobne popísaný je singulárny rozklad, ktorý má široké využitie. Mnohé metódy čerpajú z dekompozície matíc. Teoreticky i v praktických príkladoch s reálnymi dátami uvedieme analýzu hlavných komponent.

V druhej kapitole sa ukazuje súvislosť medzi metódou najmenších štvorcov a robustnými metódami jej zovšeobecnenia, kam patrí hrebeňová regresia a metóda Lasso. Okrem teoretických výsledkov sa demonštrujú metódy aj na náhodne generovaných a reálnych dátach.

V tretej kapitole vysvetlíme podstatu klasifikácie, jej využitie a základné postupy. Bližšie sa zoznámime s konkrétnou metódou, ktorou je lineárna diskriminačná analýza. Táto metóda je populárna, relatívne jednoduchá a účinná v mnohých aplikáciách. V záverečnej časti sa porovnávajú výsledky aplikácie lineárnej diskriminačnej analýzy a metódy hlavných komponent v klasifikačných problémoch pri odlišnom rozsahu dát.

Hlavný prínos práce je založený na manipulácii s variančnými maticami a metódami regresnej analýzy.

Kapitola 1

Rozklady matíc

Pri práci s maticami využívame ich rôzne vlastnosti a charakteristiky. V tejto kapitole sa bližšie pozrieme na vyjadrenie danej matice pomocou súčinu špeciálnych matíc. Výsledok takejto dekompozície má široké využitie napríklad vďaka možným lepším numerickým vlastnostiam. Rozklady sú užitočné ako v teoretickej rovine, tak aj pri praktických výpočtoch.

Poznámka 1. *V texte sa pracuje s maticami nad $\mathbb{R}$. V komplexnom obore je výpočet analogický až na to, že sa miesto transponovanej matice používa hermitovsky združená matica a rolu ortogonálnych matíc hrajú unitárne.*

Poznámka 2. *V texte sa pracuje so štandardným skalárnym súčinom, ktorý je definovaný pre ľubovolné vektory $u, v \subseteq \mathbb{R}^n$ ako*

$$\langle u, v \rangle = v^T u = \sum_{i=1}^{n} u_i v_i$$

1.1 Diagonalizácia matíc

Princípom diagonalizácie je nájdenie podobnej matice D so štvorcovou maticou A, pričom D je diagonálna. Z definície podobnosti matíc vieme, že hľadáme matice D a Q, pomocou ktorých vieme maticu A vyjadriť nasledovne

$$A = QDQ^{-1}, \tag{1.1}$$

Tento predpis sa nazýva spektrálny rozklad matice A. Vlastné čísla matice A sú na diagonále matice D a Q je tvorená vlastnými vektormi prislúchajúcimi vlastným číslam v poradí, ako sú na diagonále.

Maticu je možné diagonalizovať, ak ku jej každému vlastnému číslu existuje toľko lineárne nezávislých vlastných vektorov, koľko je jeho násobnosť ako koreňa charakteristického polynómu. Toto tvrdenie je podložené nasledujúcou vetou

Veta 1. *Predpokladajme, že vlastné čísla $\lambda_1, \ldots, \lambda_m$ matice A typu $m \times m$ pozostávajú z h-odlišných hodnôt $\mu_1, \ldots, \mu_h$ násobnosti $r_1, \ldots, r_h$ pričom platí $r_1 + \cdots + r_h = m$. Potom má matica A množinu m-lineárne nezávislých vlastných vektorov a je diagonalizovateľná vtedy a len vtedy ak $rank(A - \mu_i I_m) = m - r_i$ pre $i = 1, \ldots, h$.*

Dôkaz. viz [16].

Ľubovoľná štvorcová matica sa všeobecne diagonalizovať nedá. Je známe, že symetrická štvorcová matica je diagonalizovateľná vždy. Pre symetrické matice tiež platí, že ich vlastné čísla sú reálne a príslušné vlastné vektory sú tiež reálne. Ak je symetrická matica pozitívne semidefinitná, všetky vlastné čísla sú nezáporné. Ďalej pre symetrické matice platí, že vlastné vektory pre rôzne vlastné čísla sú ortogonálne. Preto sa dá matica Q nájsť ako ortogonálna. S týmito typmi matíc sa v štatistike stretávame veľmi často.

1.2 Singulárny rozklad

Štvorcový rozmer matice nie je nutnou podmienkou diagonalizovnosti. Pre akékoľvek matice, nie len štvorcové, je možné použiť singulárny rozklad, ktorý sa často využíva pri rôznych výpočtoch s maticami.

Existencia rozkladu pre ľubovoľnú maticu A je daná nasledujúcou vetou.

Veta 2. *Nech A je matica $m \times n$ s hodnosťou $r > 0$, potom existujú ortogonálne matice P typu $m \times m$, Q typu $n \times n$ a D typu $m \times n$ také, že platí $A = PDQ^T$. Pričom má matica D tvar*

$$\begin{bmatrix} \Delta & (0) \\ (0) & (0) \end{bmatrix},$$

kde Δ je diagonálna matica $r \times r$ s kladnými prvkami na diagonále a (0) označuje nulový blok príslušných rozmerov. Diagonálne hodnoty Δ sú odmocniny kladných vlastných čísel matice $A^T A$.

Poznámka 3. *Špeciálne prípady rozmerov matice A vedú k redukovanému tvaru matice D*

(1) $D = \begin{bmatrix} \Delta \end{bmatrix}$ *ak* $r = m = n$

(2) $D = \begin{bmatrix} \Delta & (0) \end{bmatrix}$ *ak* $r = m \leq n$

(3) $D = \begin{bmatrix} \Delta \\ (0) \end{bmatrix}$ *ak* $r = n \leq m$

Dôkaz. Nech A je matica typu $m \times n$ s hodnosťou $rank(A) = r$. Matica $A^T A$ je typu $n \times n$, symetrická a pozitívne semidefinitná. Diagonalizáciou tejto matice dostaneme

$$A^T A = QDQ^T \Longleftrightarrow Q^T A^T A Q = \tilde{D} = \begin{bmatrix} \lambda_1 & & & & & \\ & \ddots & & & & \\ & & \lambda_r & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{bmatrix},$$

pre nejakú ortogonálnu maticu Q. Hodnoty $\lambda_i > 0, i = 1,\dots,r$ sú kladné vlastné čísla matice A^TA. Nenulový blok diagonálnej matice D označme ako Δ^2. Označme $\lambda_i = \mu_i^2$, potom

$$\Delta^2 = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_r \end{bmatrix}, \quad \Delta = \begin{bmatrix} \mu_1 & & \\ & \ddots & \\ & & \mu_r \end{bmatrix}, \quad \Delta^{-1} = \begin{bmatrix} \frac{1}{\mu_1} & & \\ & \ddots & \\ & & \frac{1}{\mu_r} \end{bmatrix}.$$

Diagonalizácia v novom značení je nasledovná

$$Q^TA^TAQ = \begin{bmatrix} \Delta^2 & (0) \\ (0) & (0) \end{bmatrix}$$

Pomocou blokového zápisu $Q = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix}$, kde Q_1 je $n \times r$, dostávame

$$\begin{aligned} Q_1^TQ_1 &= I_r, & Q_1^TA^TAQ_1 &= \Delta^2, \\ Q_2^TQ_2 &= I_{n-r}, & Q_2^TA^TAQ_2 &= (0), \\ Q_2^TQ_1 &= (0), & AQ_2 &= (0). \\ Q_1^TQ_2 &= (0), \end{aligned} \tag{1.2}$$

Nech P je $n \times n$ matica s blokovým zápisom $P = \begin{bmatrix} P_1 & P_2 \end{bmatrix}$, kde $P_1 = AQ_1\Delta^{-1}$ typu $m \times r$ a P_2 je ortonormálna báza jadra matice A^T. Z toho vyplývajú nasledujúce vzťahy

$$\begin{aligned} A^TP_2 &= (0) \\ P_2^TA &= (0) \\ P_2^TAQ_1 &= (0) \\ P_2^TP_2 &= I_n \\ P_1^TP_1 &= \Delta Q_1^TA^TAQ_1\Delta^{-1} = \Delta^{-1}\Delta^2\Delta^{-1} = I_n \\ P_2^TP_1 &= P_2^TAQ_1\Delta^{-1} = (0) \end{aligned} \tag{1.3}$$

Z toho vyplýva, že P je ortogonálna matica. Použitím (1.2) a (1.3) vypočítame

$$\begin{aligned} P^TAQ &= \begin{bmatrix} P_1^TAQ_1 & P_1^TAQ_2 \\ P_2^TAQ_1 & P_2^TAQ_2 \end{bmatrix} = \begin{bmatrix} \Delta^{-1}Q_1^TA^TAQ_1 & \Delta^{-1}Q_1^TA^TAQ_2 \\ P_2^TAQ_1 & P_2^TAQ_2 \end{bmatrix} \\ &= \begin{bmatrix} \Delta & (0) \\ (0) & (0) \end{bmatrix} \end{aligned} \tag{1.4}$$

□

Poznámka 4. *Z konštrukcie dôkazu vidíme, že diagonálne prvky Δ sú kladné odmocniny kladných vlastných čísel A^TA. Tieto hodnoty sa nazývajú singulárnymi hodnotami matice A. Stĺpce matice Q tvoria ortonormálnu množinu vlastných vektorov A^TA a platí $A^TA = QDQ^T$.*

Nech A je symetrická matica s nenulovými vlastnými číslami $\mu_1,\dots,\mu_r$. Potom nenulové vlastné čísla matice $A^TA = AA^T = A^2$ sú $\mu_1^2,\dots,\mu_r^2$ (ozn. $\lambda_1,\dots,\lambda_r$). Singulárnymi hodnotami matice A sú odmocniny vlastných čísel A^2 teda $|\mu_1|,\dots,|\mu_r|$. Pozitívne semidefinitná matica má všetky vlástné čísla nezáporné, preto platí $|\mu_k| = \mu_k = \sqrt{\lambda_k}$ pre $k \in \{1,\dots,r\}$.

Dôsledok 1. *Matica AA^T má rovnaké nenulové vlastné čísla ako A^TA, pre ľubovolnú maticu A.*

Dôsledok 2. *Singulárny rozklad symetrickej pozitívne (semi)definitnej matice je v podstate jej diaginalizácia, teda spektrálny rozklad.*

Pri diagonalizovaní matice je nutné zistiť vlastné čísla, čo je pre veľké matice analyticky príliš náročné. Napríklad pre rozmer 10×10 je konštrukcia charakterického polynómu a hľadanie jeho koreňov veľmi zložitá. Singulárny rozklad má v tomto smere veľké numerické výhody. V štatistike pracujeme často s variančnými maticami, ktoré sú pozitívne semidefinitné, preto na ne pri diagonalizácii môžme použiť singulárny rozklad.

1.3 Analýza hlavných komponent

Analýza hlavných komponentov (PCA) je štatistická procedúra využívajúca maticovú algebru. Je to najčastejšie používaná metóda pri redikcii dimenzionality. Základnou myšlienkou PCA je transformovať pôvodné dáta tak, aby sme na novom súbore pozorovali menej veličín, než v pôvodnom súbore. Nové pozorované veličiny sú nazvané hlavné komponenty. V pôvodnej množine dát je prípustná korelácia, no novú menšiu množinu tvoria nekorelované dáta. Pričom pri znižovaní dimenzie je cieľom čo najmenšia strata informácií o pôvodnej množine. Hlavné komponenty volíme ako lineárne kombinácie pôvodných veličín tak, aby boli vzájomne nekorelované, a aby niekoľko prvých hlavných komponent vysvetlovalo čo najväčšiu časť variability obsiahnutej v pôvodonom súbore dát. Ďalej túto metódu detailnejšie popíšeme a pritom vychádzame z knihy J. Anděla [1].
Predpokladáme, že máme dáta

$$\begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix},$$

ktoré sú realizáciou p-rozmerného náhodného vektoru $\mathbf{X}$, pričom $X_i = (x_{1i}, \ldots, x_{ni})^T$ sú realizácie jeho i-tej zložky. Jeho variančnú maticu označíme $\Sigma_{\mathbf{X}}$, je symetrická a pozitívne semidefinitná. Budeme predpokladať, že vlastné čísla variančnej matice sú navzájom rôzne. Označíme ich $\lambda_1 > \lambda_2 > \cdots > \lambda_p > 0$ a príslušné vlastné vektory $v_1, v_2, \ldots, v_p$. Pre prvú komponentu hľadáme vektor koeficientov lineárnej kombinácie $\mathbf{c_1}$, pričom $\mathbf{c_1}'\mathbf{c_1} = 1$ tak, aby mala náhodná veličina $\mathbf{c_1}'\mathbf{X}$ maximálny rozptyl, tj.

$$\operatorname{var}(\mathbf{c_1}'\mathbf{X}) = \max_{\mathbf{c},\mathbf{c}'\mathbf{c}=1} \{\operatorname{var}(\mathbf{c}'\mathbf{X})\}. \tag{1.5}$$

Podmienku (1.5) splňuje $\mathbf{c_1} = v_1$ (viz [1], kap. XVII.2). Nájde sa tak smer, v ktorom sú dáta najvariabilnejšie. Prvá hlavná komponenta je teda $Z_1 = v_1'\mathbf{X}$, pre ktorú platí

$$\operatorname{var}(Z_1) = \operatorname{var}(v_1'\mathbf{X}) = \operatorname{var}(v_1'\Sigma v_1) = \lambda_1$$

Ďalšie komponenty $Z_i, i = 2, 3, \ldots, p$ hľadáme postupne ako lineárne kombinácie $\mathbf{c_i}'\mathbf{X}$, pričom $\mathbf{c_i}'\mathbf{c_i} = 1$ tak, aby boli nekorelované s doteraz nájdenými komponentami $Z_i, \ldots, Z_{i-1}$ a zároveň aby mali medzi týmito možnými kombináciami maximálny rozptyl, tj.

$$\operatorname{var}(\mathbf{c_i}'\mathbf{X}) = \max_{\mathbf{c},\mathbf{c}'\mathbf{c}=1,\mathbf{c}\perp\mathbf{c}_1,\ldots,\mathbf{c}\perp\mathbf{c}_{i-1}} \{\operatorname{var}(\mathbf{c}'\mathbf{X})\}.$$

Tieto podmienky sú splnené pre voľbu $\mathbf{c_i} = v_i$ a pre i-tu komponentu $Z_i = v_i'\mathbf{X}$ platí $\text{var}(Z_i) = \lambda_i$ (viz [1], kap. XVII.2). Celkovú variabilitu obsiahnutú vo vektore $\mathbf{X}$ označíme $\sigma^2 = \sum_{i=1}^{p} \text{var}\mathbf{X}_i$. Pre ďalšiu analýzu dát uvažujeme iba prvých q komponent, ktoré vysvetľujú dostatočnú variabilitu dát. Využijeme na to fakt, že zároveň platí $\sigma^2 = \sum_{i=1}^{p} \lambda_i$ a $\lambda_i > \lambda_{i+1}, i = 1,2,\ldots,p-1$. Väčšinou q volíme tak, aby

$$\frac{\sum_{i=1}^{q} \lambda_i}{\sigma^2} \geq \alpha, \tag{1.6}$$

kde za α volíme $0.9, 0.95$ alebo 0.99.

Poznámka 5 (Maticová interpretácia PCA)**.** *V prvom kroku aplikujeme lineárnu transformáciu vektorov $X_1,\ldots,X_p$, ktoré tak prevedieme na vzájomne nekorelované vektory pomocou vektorov $\mathbf{c}_i$. Lineárnu transformáciu ilustruje* (1.7).

$$\begin{aligned} Z_1 &= c_{11}X_1 + c_{12}X_2 + \ldots c_{1p}X_p = \mathbf{c_1}^T\mathbf{X} \\ Z_2 &= c_{21}X_1 + c_{22}X_2 + \ldots c_{2p}X_p = \mathbf{c_2}^T\mathbf{X} \\ &\vdots \\ Z_p &= c_{p1}X_1 + c_{p2}X_2 + \ldots c_{pp}X_p = \mathbf{c_p}^T\mathbf{X} \end{aligned} \quad \Longrightarrow \quad \mathbf{Z} = \mathbf{C}\mathbf{X} \tag{1.7}$$

Dôsledkom je nekorelovanosť premenných $Z_1,\ldots,Z_p$, zároveň chceme, aby mali premenné $Z_1,\ldots,Z_p$ po poradí čo najväčší rozptyl. Z nekorelovanosti vyplýva, že variančná matica $\text{var}(\mathbf{Z})$*, ozn. $\Sigma_\mathbf{Z}$, je diagonálna. Odvodením získame tvar*

$$\Sigma_\mathbf{Z} = \text{var}(\mathbf{y}) = \text{cov}(\mathbf{X}C^T) = \mathbf{C}\Sigma_\mathbf{X}\mathbf{C}^T$$

Vidíme, že sa jedná o diagonalizáciu matice $\Sigma_\mathbf{X}$. Tá je pozitívne (semi)definitná, a teda podľa teórie zo sekcie 1.2 je možné aplikovať pri diagonalizácii singulárny rozklad.

Algoritmy používané pre singulárny rozklad naviac zoraďujú singulárne čísla podľa veľkosti, čo uľahčuje výber hlavných komponent.

1.3.1 Aplikácia v praxi

V tejto sekcii budeme aplikovať analýzu hlavných komponent v dvoch odlišných príkladoch. V prvom príklade chceme identifikovať charakteristiky hlavných komponent pre dátový súbor pochádzajúci z R. V druhom príklade je úlohou kompresia obrázku. Pre spracovanie dát a implementáciu analýzy hlavných komponent používame štatistický program R. Pre výpočet hlavných komponent sa aplikuje sa funkcia *prcomp*.

Príklad 1 (Súbor *cars*[1])**.**

Súbor *cars* obsahuje 11 údajov o 32 modeloch áut. Nižšie je uvedená ukážka dátového súboru (viz Tabuľka 1.1) a popis jednotlivých parametrov

- **lkm**: spotreba paliva [počet litrov na 100km]. Výkonnejšie a ťažšie autá majú zvyčajne vyššiu spotrebu.
- **cyl**: počet valcov. Viac valcov zvyšuje výkon auta.
- **disp**: súčet objemu všetkých valcov.
- **kW**: výkon auta [kW].
- **drat**: pomer zadnej nápravy. Popisuje ako otočenie hnacieho hriadeľa zodpovedá otáčaniu kolies. Vyššie hodnoty znižujú spotrebu paliva.
- **wt**: váha [100kg]
- **qsec**: zrýchlenie. Minimálny čas [s] potrebný na prejdenie vzdialenosti 1 km z nulovej rýchlosti.
- **vs**: tvar motoru. Binárna premenná vypovedajúca, či je tvar motoru do V (1) alebo má bežný tvar (0).
- **am**: typ prevodovky. Binárna premenná vypovedajúca, či sa jedná o manuálnu (1) alebo automatickú (0) prevodovku.
- **gear**: počet prevodových stupňov. Športové autá majú zvyčajne viac stupňov prevodu.
- **carb**: počet karburátorov. Veličina úzko spätá s výkonom auta.

	lkm	cyl	disp	kW	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	11.2	6	160	82	3.90	11.9	40.9	0	1	4	4
Mazda RX4 Wag	11.2	6	160	82	3.90	13.0	42.3	0	1	4	4
Datsun 710	10.3	4	108	69	3.85	10.5	46.3	1	1	4	1
Hornet 4 Drive	11.0	6	258	82	3.08	14.6	48.3	1	0	3	1
Hornet Sportabout	12.6	8	360	130	3.15	15.6	42.3	0	0	3	2
Valiant	13.0	6	225	78	2.76	15.7	50.3	1	0	3	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Tabuľka 1.1: Ukážka súboru *cars* použitý pre praktickú aplikáciu analýzy hlavných komponent.

Realizáciou náhodného vektora $\mathbf{X}$ sú jednotlivé parametre áut. PCA najlepšie funguje pre numerické premenné, preto vyradíme binárne premenné (**vs** a **am**). Pracujeme ďalej s dátovou maticou rozmerov 32×9. Súbor vycentrujeme a preškálujeme, kvôli odlišným jednotkám parametrov. Variančná matica $\Sigma_{\mathbf{X}}$ má tvar

[1] Zdroj: súbor mtcars R

$$\Sigma_{\mathbf{X}} = \begin{pmatrix} 1.00 & 0.82 & 0.88 & 0.76 & -0.64 & 0.89 & -0.39 & -0.48 & 0.53 \\ 0.82 & 1.00 & 0.90 & 0.83 & -0.70 & 0.78 & -0.59 & -0.49 & 0.53 \\ 0.88 & 0.90 & 1.00 & 0.79 & -0.71 & 0.89 & -0.43 & -0.56 & 0.39 \\ 0.76 & 0.83 & 0.79 & 1.00 & -0.45 & 0.66 & -0.71 & -0.13 & 0.75 \\ -0.64 & -0.70 & -0.71 & -0.45 & 1.00 & -0.71 & 0.09 & 0.70 & -0.09 \\ 0.89 & 0.78 & 0.89 & 0.66 & -0.71 & 1.00 & -0.17 & -0.58 & 0.43 \\ -0.39 & -0.59 & -0.43 & -0.71 & 0.09 & -0.17 & 1.00 & -0.21 & -0.66 \\ -0.48 & -0.49 & -0.56 & -0.13 & 0.70 & -0.58 & -0.21 & 1.00 & 0.27 \\ 0.53 & 0.53 & 0.39 & 0.75 & -0.09 & 0.43 & -0.66 & 0.27 & 1.00 \end{pmatrix}$$

Variančná matica je symetrická a test *is.positive.definite* v R, ktorý skúma vlastné čísla matice, potvrdil pozitívnu definitnosť. Vďaka dôsledku 2 môžme $\Sigma_{\mathbf{X}}$ diagonalizovať pomocou SVD rozkladu ako

$$\Sigma_{\mathbf{X}} = QDQ^T, \tag{1.8}$$

pričom použitím zápisu (1.7) platí $Q = C$ a $\Sigma_Z = C\Sigma_X C^T$.

Matica $\Sigma_{\mathbf{Z}}$ je diagonálna, po poradí tvorená vlastnými číslami matice $\Sigma_{\mathbf{X}}$

$$\begin{aligned} \lambda_1 &= 5.64 \quad \lambda_4 = 0.28 \quad \lambda_7 = 0.10 \\ \lambda_2 &= 2.08 \quad \lambda_5 = 0.18 \quad \lambda_8 = 0.05 \\ \lambda_3 &= 0.52 \quad \lambda_6 = 0.13 \quad \lambda_9 = 0.02 \end{aligned} \tag{1.9}$$

Matica $\mathbf{C}$ lineárne transformuje pôvodný súbor. To znamená, že hovorí o vzťahoch medzi pôvodnými parametrami. Má tvar

$$\mathbf{C} = \begin{pmatrix} -0.39 & 0.04 & 0.29 & 0.25 & 0.08 & 0.63 & -0.45 & 0.31 & 0.00 \\ -0.40 & -0.02 & -0.27 & -0.00 & -0.21 & -0.23 & 0.31 & 0.74 & -0.17 \\ -0.40 & 0.09 & -0.06 & 0.29 & 0.41 & -0.15 & 0.26 & -0.12 & 0.68 \\ -0.37 & -0.27 & -0.02 & 0.02 & 0.20 & -0.57 & -0.58 & -0.15 & -0.24 \\ 0.31 & -0.34 & 0.20 & 0.82 & -0.18 & -0.17 & 0.08 & 0.09 & -0.06 \\ -0.38 & 0.17 & 0.44 & 0.08 & 0.09 & 0.01 & 0.47 & -0.33 & -0.53 \\ 0.22 & 0.49 & 0.61 & -0.12 & 0.06 & -0.40 & -0.14 & 0.34 & 0.16 \\ 0.21 & -0.55 & 0.21 & -0.26 & 0.64 & 0.09 & 0.20 & 0.28 & -0.07 \\ -0.24 & -0.49 & 0.42 & -0.31 & -0.53 & -0.03 & 0.08 & -0.10 & 0.36 \end{pmatrix}.$$

Tabuľka 1.2 obsahuje podiel rozptylu každej komponenty z rozptylu celého súboru a kumulatívny rozptyl. Z tabuľky vidíme, ako každá komponenta prispieva svojím rozptylom do celkového rozptylu. Prvé 2 komponenty obsahujú vyše 85% rozptylu pôvodného súboru a prvé 3 vyše 90%. Použitím všetkých komponent by sme dostali rovnaký rozptyl ako pri pôvodnom súbore, avšak tým by sa cieľ redukcie dimenzie vôbec nenaplnil.

Prvá komponenta (PC_1) vysvetľuje až 63% variability celého súboru, t.j. obsahuje skoro $2/3$ celkovej informácie o dátach s 32 pozorovaniami a 9 premennými. Druhá komeponenta (PC_2) vysvetľuje 23% variability celého súboru. A tretia skoro 6%. Použitím prvej, druhej a tretej komponenty dostaneme dosť presný obraz a popis analyzovaného súboru. Dokopy komponenty vysvetľujú skoro 92% jeho variability.

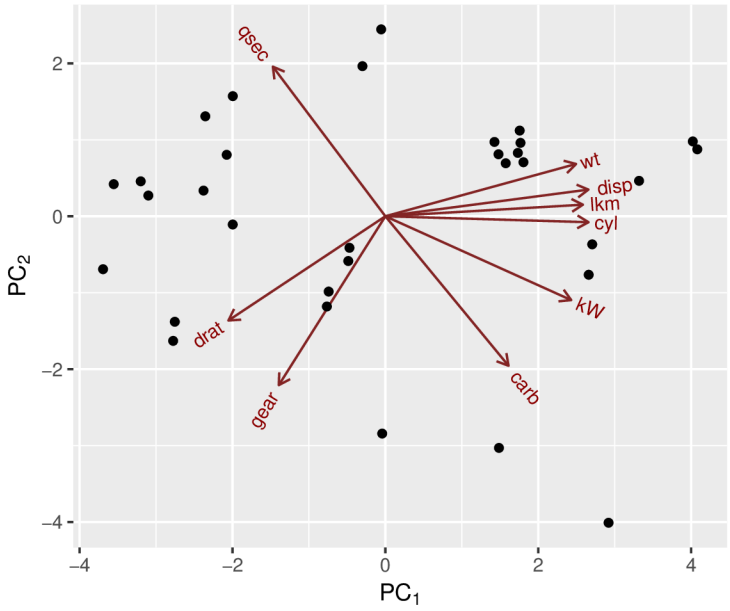
	PC_1	PC_2	PC_3	PC_4	PC_5	PC_6	PC_7	PC_8	PC_9
proporcionálny rozptyl	0.6264	0.2316	0.0580	0.0310	0.0195	0.0143	0.0109	0.0057	0.0027
kumulatívny rozptyl	0.6264	0.8580	0.9160	0.9470	0.9665	0.9807	0.9917	0.9973	1.0000

Tabuľka 1.2: Výsledok aplikácie metódy PCA na súbor *cars*.

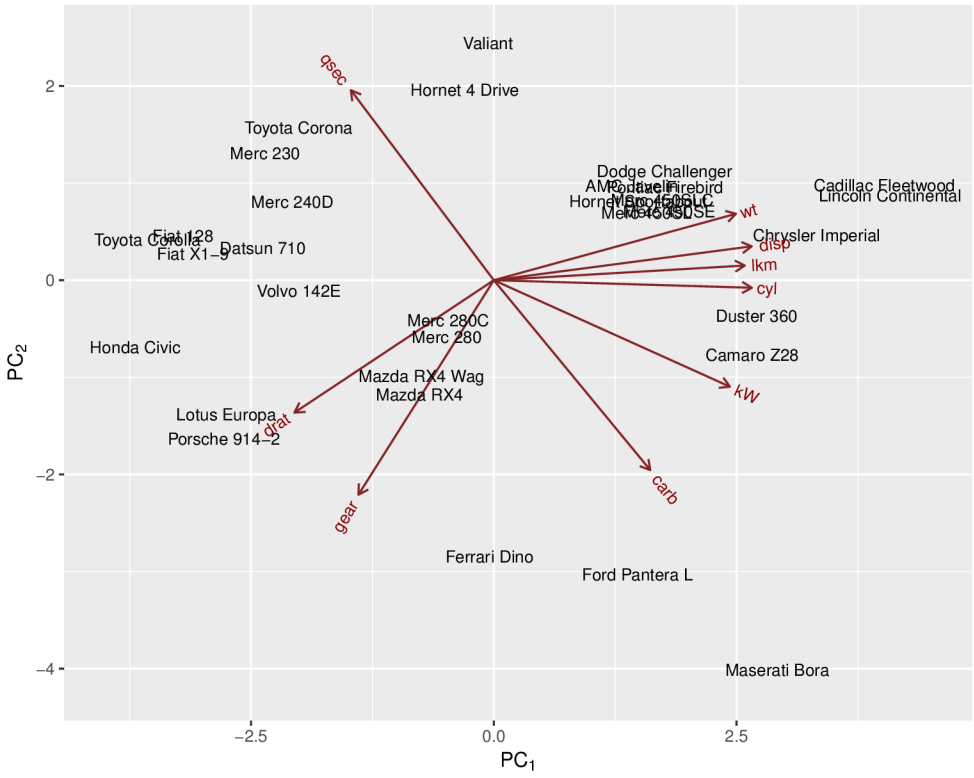
Pozrime sa na výber počtu koponent cez iný prístup. Súčet vlastných čísel je 18.8. Využitím (1.6) a zvolením $\alpha = 0.9$ dostávame, že počet vybraných hlavných komponent, ktoré obsahujú 90% variability pôvodného súboru je 3

$$\frac{5.64+2.08}{18.8} = 0.86 < \alpha$$
$$\frac{5.64+2.08+0.52}{18.8} = 0.92 > \alpha.$$

Na obrázku 1.1 je zobrazená pozícia každého pozorovania (čierne bodky) a osy (červené šípky) prvej a druhej kompenenty. Osy vychádzajú z počiatku, pretože sme súbor vycentrovali a preškálovali. Vidíme, že premenné $wt, disp, lkm, cyl$ a kW prispievajú ku komponente PC_1, pretože sú v rovnakom smere ako jej os PC_1. To znamená, že prvá komponenta je dobre popísateľná týmito parametrami. Pomocou týchto parametrov sa vysvetlí až vyše 62% všetkej variability pôvodného súboru. Preto, ak chceme opísať nejakým typ auta, veľa nám povie jeho váha, spotreba paliva, výkon, počet a objem valcov. Obrázok 1.2 je rovnaký, iba namiesto čiernych bodiek sú názvy jednotlivých áut. Analýzou značiek áut dospejeme k záveru, že vpravo sa nachýdzajú väčšie typy ako sedany a limuzíny a vpravo skôr menšie, 4- a 2-dverové autá. Toto rozdelenie pozdĺž osy PC_1 zapadá do vyššie popísaného vysvetlenia prvej komponenty.



Obr. 1.1: PCA aplikovaná na súbor *cars*.



Obr. 1.2: PCA aplikovaná na súbor *cars*.

Príklad 2 (Obrázok *zebry*)**.**

V tomto príklade budeme analyzovať obrázok *zebry* (viz obr 1.3f) zo zdroju [24]. Pôvodný súbor má veľkosť 313 kB a je reprezentovaný tromi 1499×1000 maticami, ich kombinácia tvorí RGB popis obrázku. Matica R, reprezentuje červenú farebnú škálu, matica G zelenú a matica B modrú.

V tomto príklade sa zameriavame na kompresiu a nie popis či interpretáciu premenných. Cieľom je z vysokodimenzinálnych matíc farieb dostať menšie, popri čo najmenšej strate kvality výsledného obrazu. Vďaka kompresii sa šetrí úložná pamäť.

Metódu PCA aplikujeme na každú maticu. Výsledkom sú nové matice farieb, ktoré majú nižšiu dimenziu. Nové hodnoty pixlov sú projektované do nižších dimenzí a vytvárajú príslušný obrázok.

Pri voľbe počtu hlavných komponet postupujeme rovnako ako v predchádzajúcej úlohe. Pre aplikáciu pravidla (1.6) potrebujeme pracovať s vlastnými číslami. Pre každú maticu (R, G, B) vypočítame jej variančnú maticu a pomocou SVD rozkladu zistíme priamo jej vlastné čísla. Počet vlastných čísel, vrátane ich násobností, každej matice je 1499.

Prvých 10 vlastných čísel, zoradených zostupne, a súčet všetkých pre každú variančnú maticu sú vypísané v tabuľke 1.3. Tabuľka 1.4 obsahuje počet hlavných komponent nutný na vysvetlenie daného percenta variability matice R, G, alebo B.

	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}	$\sum_{i=1}^{1499} \lambda_i$
var (R)	10.00	6.95	5.82	4.94	3.53	2.54	2.37	1.75	1.53	1.33	58.53
var (G)	8.43	6.31	5.03	4.13	2.73	2.39	2.21	1.86	1.52	1.17	53.53
var (B)	22.52	7.83	5.92	4.70	3.04	2.52	2.28	2.00	1.52	1.50	72.31

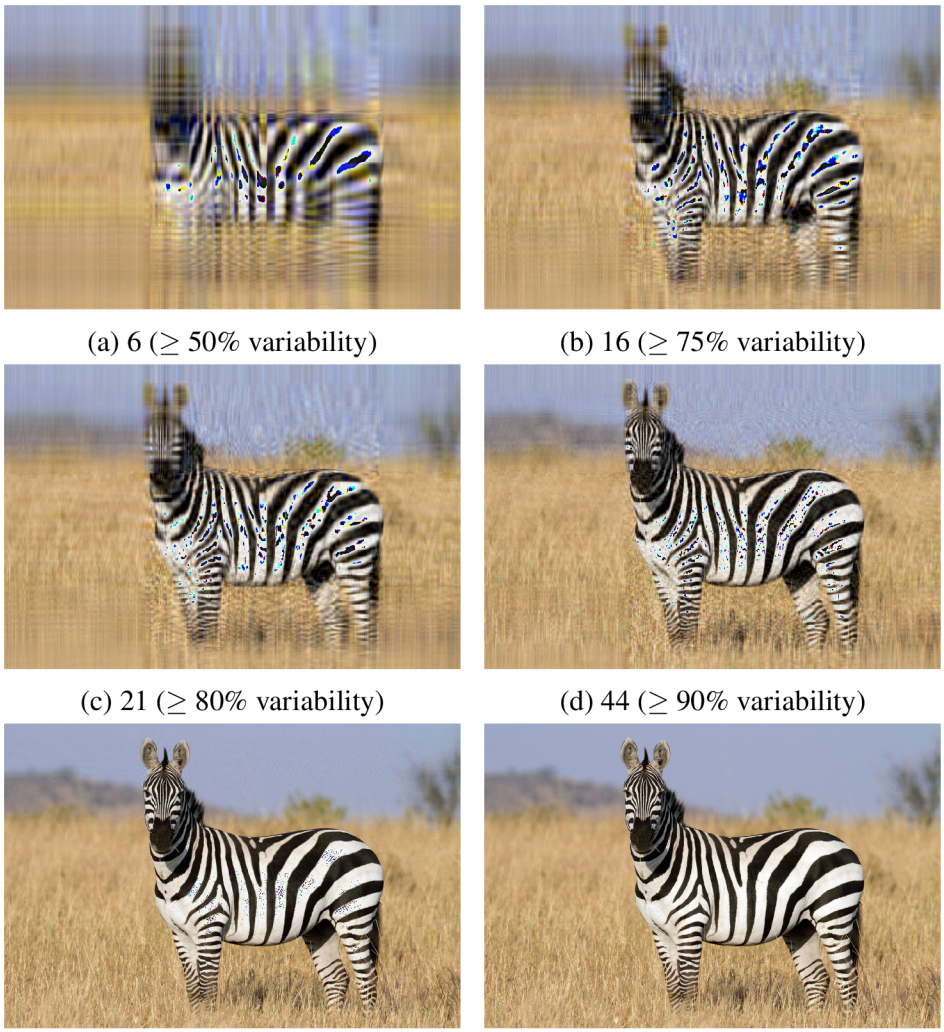
Tabuľka 1.3: Prvých 10 vlastných čísel a súčet všetkých vlastných čísel pre každú variančnú maticu pôvodného súboru.

variabilita/ variančná matica	$\geq 50\%$	$\geq 75\%$	$\geq 80\%$	$\geq 90\%$	$\geq 99\%$
var (R)	5	14	18	40	168
var (G)	6	16	21	44	176
var (B)	3	11	15	34	159

Tabuľka 1.4: Minimálny počet vlasných čísel zaisťujúcich dané percento variability pre konktrétnu variančnú maticu.

Aby bolo zaručené percento variability z každej matice, zvolíme maximum z počtu hlavných komponent pre všetky prípady. Výberom vznikajú nové obrázky (viz obrázky 1.3a-1.3e). Vidíme, že čím je počet komponent vyšší, je reprezentácia pôvodného obrázku presnejšia, jasnejšia a ostrejšia. Prvých 6 hlavných komponent (viz obrázok 1.3a) obsiahnu 50% variability pôvodného súboru, obrázok má však dosť zlú kvalitu. Na druhej strane už od 16 hlavných komponent, z celkových 1000, vieme obrázok dobre rozornať. Pri obsiahnutí 99% variability, teda použitím 176 komponent, nevidíme voľným okom rozdiel v kvalite od originálu. Veľkosť súbotu je pritom zmenšená o 38%.

Tabuľka 1.5 obsahuje počet použitých hlavných komponent, aktuálnu veľkosť a precentuálne zmenšenie veľkosti oproti originálu pre každý obrázok 1.3a-1.3f.



(a) 6 ($\geq$ 50% variability)

(b) 16 ($\geq$ 75% variability)

(c) 21 ($\geq$ 80% variability)

(d) 44 ($\geq$ 90% variability)

(e) 176 ($\geq$ 99% variability)

(f) 1000 (pôvodný obrázok)

Obr. 1.3: Aplikácia PCA na obrázok zebry s počtom použitých hlavných komponent a percentom obsiahnutej variability pôvodného súboru.

obrázok	počet hlavných komponent	veľkosť obrázku [kB]	percentuálna redukcia
1.3a	6	106	- 66%
1.3b	16	141	- 55%
1.3c	21	149	- 52%
1.3d	44	178	- 43%
1.3e	176	194	- 38%
1.3f	1000	313	0%

Tabuľka 1.5: Redukcia veľkosti obrázku zebry metódou PCA.

Prvé komponenty obsahujú najvyšiu variabilitu, preto je signifikatné zlepšenie kvality pri prechode od 6 ku 16. Pri ďalšom náraste variability už nie je rozdieľ medzi obrázkami taký očividný. Kopresia obrázkov je pomocou PCA jednoduchá a jasná. Obrázok predstavujeme ako združenie matíc $(n \times p)$ alebo $(n \times n)$ tvorené z farebných hodnôt pixelov.

Kapitola 2

Lineárna regresia

Pojem regresie prvýkrát použil už koncom 19. storočia britský učenc Francis Galton, viz [11] a [12]. Skúmal závislosť priemernej výšky detí na výške rodičov. Jedným zo zistení bol fakt, že synovia vysokých rodičov mali síce v priemere výšku vyššiu než priemer populácie, ale individuálne nedosahovali extrémnych hodnôt predchádzajúcej generácie. Ako keby sa jednotlivci "vracali k priemeru". Podobne to tak je aj s inými vlastnosťami, nie len u ľudí.

Postupne sa pojem regresie ustálil a zastrešoval akékoľvek skúmanie súvislostí medzi náhodnými veličinami. Dnes sa pod termínom regresná analýza rozumie analýza závislostí medzi závisle premennou náhodnou veličinou (niekedy definovaná ako odozva alebo vysvetľovaná premenná) a jednou alebo viacerými nezávislými (náhodnými) veličinami (nazývané tiež prediktory, regresory alebo vysvetľujúce premenné). Kapitola čerpá hlavne z [1] a [22].

2.1 Pojmy a značenie

V tejto sekcii uvedieme pojmy a značenia používané v nasledujúcich sekciách kapitoly. Zavedenie regresného modelu je prevzaté z [1].

Máme náhodné veličiny $Y_1,\dots,Y_n$ a maticu daných čísel $\mathbf{X} = (x_{i,j})$ typu $n \times k$, kde $k < n$. Predpokladáme, že pre náhodný vektor $\mathbf{Y} = (Y_1,\dots,Y_n)^T$ platí

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

kde $\beta = (\beta_1,\dots,\beta_k)^T$ je vektor neznámych parametrov a $\varepsilon = (\varepsilon_1,\dots,\varepsilon_n)^T$ je náhodný vektor splňujúci podmienky $\mathrm{E}\,\varepsilon = 0$, $\mathrm{var}\,\varepsilon = \sigma^2\mathbf{I}$. Pritom $\sigma^2 > 0$ je taktiež neznámy parameter. Tento model nazývame regresný. Pretože $\mathbf{Y}$ závisí na β lineárne, hovoríme o *lineárnom regresnom modely*. Vektor $\mathbf{X}\beta$ je náhodný, preto platí $\mathrm{E}\,\mathbf{Y} = \mathbf{X}\beta$ a heteroskedasticita $\mathrm{var}\,\mathbf{Y} = \sigma^2\mathbf{I}$. Zbytočne sa tiež nezavádzajú nadbytočné vysvetľujúce premenné $x_{i,j}$, pretože sa ideálne predpokladá lineárna nezávislosť stĺpcov matice $\mathbf{X}$. Z predpokladu $k < n$ teda plynie $h(\mathbf{X}) = k$ a regularita matice $\mathbf{X}^T\mathbf{X}$. V praxi sa však bežne stretneme s porušením lineárnej nezávislosti a so singuláritou matice $\mathbf{X}^T\mathbf{X}$. K tomuto problému sa pristupuje *upravením regresie* a podrobnejšie je to rozobrané v nasledujúcich sekciách 2.3 a 2.4.

Ďalšie poznámky k lineárnemu regresnému modelmu:

- Vektor β je sa tiež nazýva vektor regresných koeficientov.
- Parameter σ^2 je rozptyl reziduí daného modelu.
- Keďže $\mathrm{E}\,\mathbf{Y} = \mathbf{X}\beta$ a $\mathrm{var}\,\mathbf{Y} = \sigma^2\mathbf{I}$, potom môžme značiť rozdelenie náhodného vektoru $\mathbf{Y} \sim (\mathbf{X}\beta, \sigma^2\mathbf{I})$.
- Ak namiesto vektoru závisle premenných $\mathbf{Y}$ bude matica typu $n \times m, m > 1$, hovoríme o viacrozmernej regresii.

2.2 Metóda nejmenších štvorcov

Metóda najmenších štvorcov, nazývaná OLS (z anglického "ordinary least squares"), sa využíva v regresnej analýze, ekonómii a v mnohých ďalších oblastiach. Metódu formulujeme s využitím sekcie 2.1 ako optimalizáciu úlohy

$$\min_{\beta \in \mathbb{R}^p} || \mathbf{Y} - \mathbf{X}\beta ||^2, \tag{2.1}$$

pričom $|| \cdot ||$ je euklidovská norma v $\mathbb{R}^n$. Geometrická interpretácia úlohy (2.1) je hľadanie vektoru β takého, aby súčet vzdialenosti bodov $\mathbf{x}_1, \ldots, \mathbf{x}_n$ od nadroviny definovanej predpisom $\beta^T\mathbf{w} - \mathbf{Y} = 0$, kde $\mathbf{w}$ je nezávisle premenná, bol minimálny.

Analytické riešenie úlohy (2.1) označíme $\hat{\beta}_{OLS}$. Úloha nemá žiadne obmädzujúce podmienky, takže riešenie vypočítame minimalizáciou funkcie

$$L(\beta) = || \mathbf{Y} - \mathbf{X}\beta ||^2 = (\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta) = \mathbf{Y}^T\mathbf{Y} - 2\mathbf{Y}^T\mathbf{X}\beta + \beta^T\mathbf{X}^T\mathbf{X}\beta.$$

Druhá derivácia podľa β funkcie $L(\beta)$ má tvar

$$\frac{\partial^2 L}{\partial\beta\partial\beta}(\beta) = 2\mathbf{X}^T\mathbf{X}.$$

Dostávame pozitívne definitnú maticu, takže pôvodná funkcia je konvexná a každé lokálne minimum funkcie $L(\beta)$ je jej globálnym minimom. Preto vieme, že pre nájdenie minima musí platiť

$$\frac{\partial L}{\partial\beta}(\beta) = 0,$$

a riešime sústavu lineárnych rovníc

$$\mathbf{X}^T\mathbf{X}\beta = \mathbf{X}^T\mathbf{Y}. \tag{2.2}$$

Riešením sústavy (2.2) je hľadaný vektor $\hat{\beta}_{OLS}$

$$\hat{\beta}_{OLS} = (\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{Y}. \tag{2.3}$$

Odhad $\hat{\beta}_{OLS}$ je najlepší nevychýlený odhad, tj. $\mathrm{E}\hat{\beta}_{OLS} = \beta$, s najnižším rozptylom, čo znamená $\mathrm{var}\,\hat{\beta}_{OLS} < \mathrm{var}\,\hat{\beta}$ (kde $\hat{\beta}$ je ľubovoľný odhad β), pre pôvodný vektor regresných parametrov β . Dôkaz je uvedený v [22] kapitola 3.2.

Matica $\mathbf{X}^T\mathbf{X}$ nemusí byť vždy regulárna, tj. nemusí vždy existovať jej inverzia, preto označenie $(\mathbf{X}^T\mathbf{X})^-$ značí jej pseudoinverziu. Jednou z najrozšírenejších pseudoinverzí je Mooreova-Penroseova pseudoinverzia, tá existuje pre ľubovoľnú reálnu maticu a je určená jednoznačne. Pseudoinverzii sa podrobnejšie venujeme v poslednej kapitole, v časti 3.4.1. Ak má matica regresorov $\mathbf{X}$ plnú hodnosť, je matica $\mathbf{X}^T\mathbf{X}$ regulárna, pozitívne definitná a dá sa invertovať.

Poznámka 6. *Niekedy sa sústava rovníc* (2.2) *nazýva aj sústava nomálnych rovníc, alebo iba normálne rovnice.*

2.3 Upravená regresia

V tejto sekcii sa venujeme zovšeobnenením metódy najmenších štvorcov.

Regularizovaná regresia je typicky implementovaná ako penalizovaná funkcia metódy najmenších štvortcov. Jej všeobecný tvar je

$$\mathrm{PLS}(\beta) = (\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta) + \lambda \cdot \mathrm{pen}(\beta) \to \min_{\beta}, \tag{2.4}$$

kde $\mathrm{pen}(\beta)$ je penalizovaný výraz merajúci komplexitu vektora regresných koeficientov a $\lambda \geq 0$ je vyhľadzujúci parameter riadiaci veľkosti penalty. V nasledujúcich podsekciách sa zoznámime s hrebeňovou regresiou a metódou Lasso. Obe metódy riešia penalizovanú funkciu ako optimalizačnú úlohu. Motiváciou používania iných metód je vyrovnanie sa s problémom multikolinearity.

2.3.1 Multikolinearita

Multikolinearita negatívne ovplyvňuje odhad koeficientov $\hat{\beta}_{OLS}$. Problém nastáva, ak sú stĺpce regresnej matice $\mathbf{X}$ lineárne závisle, tj. $\mathrm{rank}(\mathbf{X}) < p$. Častým prípadom je aj situácia, keď síce platí $\mathrm{rank}(\mathbf{X}) = p$, ale stĺpce matice sú takmer lineárne závislé.

V tejto podsekcii si ukážeme, ako stredná hodnota štvorca odhadu $\hat{\beta}_{OLS}$ závisí na rozptyle odhadov jednotlivých regresných koeficientov.

Veta 3. *Nech* $\mathbf{Y} \sim (\mathbf{X}\beta, \sigma^2\mathbf{I}_n), \mathrm{rank}\,\mathbf{X} = p$ *a* $\hat{\beta}_{OLS} = \left(\hat{\beta}_{OLS,1}, \ldots, \hat{\beta}_{OLS,p}\right)^T$ *je optimálnym riešením* (2.1)*. Potom pre strednú hodnotu štvorcového odhadu platí*

$$E \mid\mid \hat{\beta}_{OLS} \mid\mid^2 = \mid\mid \beta \mid\mid^2 + \sigma^2 tr\left\{(\mathbf{X}^T\mathbf{X})^{-1}\right\} = \mid\mid \beta \mid\mid^2 + \sum_{i=1}^{p} var\,\hat{\beta}_{OLS,i}. \tag{2.5}$$

Dôkaz. $E \mid\mid \hat{\beta}_{OLS} - \beta \mid\mid = tr\{var\,\hat{\beta}_{OLS}\} = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} = E \mid\mid \hat{\beta}_{OLS} \mid\mid - \mid\mid \beta \mid\mid$ □

Aplikáciou singulárneho rozkladu zo sekcie 1.2 máme nasledovné vyjadrenie matice regresorov

$$\mathbf{X} = \mathbf{PDQ}^T = \sum_{i=1}^{p} d_i \mathbf{p}_i \mathbf{q}_i^T, \tag{2.6}$$

kde je

$\mathbf{P} = (\mathbf{p}_1, \ldots, \mathbf{p}_p)$ ortogonálna matica $n \times n$,
$\mathbf{Q}^T = (\mathbf{q}_1^T, \ldots, \mathbf{q}_p^T)$ ortogonálna matica $p \times p$,
$\mathbf{D} = \begin{bmatrix} \Delta \\ 0 \end{bmatrix}$, $\Delta = diag(d_1, \ldots, d_p)$ diagonálna matica $p \times p$,
a $d_1 \geq \cdots \geq d_p > 0$ sú odmocniny kladných vlastných čísel matice $\mathbf{X}^T\mathbf{X}$.

Ak sú stĺpce matice $\mathbf{X}$ takmer lineárne závislé, je matica $\mathbf{X}^T\mathbf{X}$ skoro singulárna, a preto hodnota d_p^2 je veľmi blízko hodnote 0. Maticu $\mathbf{X}^T\mathbf{X}$ môžme vyjadriť pomocou (2.6) nasledovne

$$\mathbf{X}^T\mathbf{X} = \mathbf{Q}\mathbf{D}^T\mathbf{D}\mathbf{Q}^T = \sum_{i=1}^{p} d_i^2 \mathbf{q}_i \mathbf{q}_i^T \tag{2.7}$$

Vďaka ortogonalite platí

$$tr\{\mathbf{q}_i\mathbf{q}_i^T\} = tr\{\mathbf{q}_i^T\mathbf{q}_i\} = 1.$$

Potom stopu matice $\mathbf{X}^T\mathbf{X}$ môžme zapísať ako

$$tr\left\{(\mathbf{X}^T\mathbf{X})^{-1}\right\} = \sigma^2 \sum_{i=1}^{p} \frac{1}{d_i^2}. \tag{2.8}$$

Vďaka (2.5) a (2.8) odvodíme

$$E \mid\mid \hat{\beta}_{OLS} \mid\mid^2 - \mid\mid \beta \mid\mid^2 = \sum_{i=1}^{p} var\,\hat{\beta}_{OLS,i} = \sigma^2 tr\left\{(\mathbf{X}^T\mathbf{X})^{-1}\right\} = \sigma^2 \sum_{i=1}^{p} \frac{1}{d_i^2}. \tag{2.9}$$

Dôsledok 3. *Vyjadrením rozdielu* (2.9) *vidíme, že multikolinearita môže spôsobovať veľké rozdiely v odhadoch jednotlivých koeficientov. Príliš vysoké hodnoty* $var\,\hat{\beta}_{OLS_i}$ *spôsobujú nepoužiteľnosť odhadov v praxi.*

Najčastejšie používané kritérium na odhalenie dvojice regresorov zodpovedných za multikolinearitu je vysoká absolútna hodnota výberového korelačného koeficientu medzi danou dvojicou regresorov, alebo priamy štatistický test na multikolinearitu.

2.3.2 Hrebeňová regresia

Ako už bolo povedané na začiatku tejto sekcie, hrebeňová regresia (z angličtiny *Ridge regresion*) je zovšeobcnením metódy OLS. Prvotnou motiváciu bolo vyriešenie singularity pri úlohách s neinvertovateľnými maticami. Autormi myšlienky hrebeňovej regresie sú Hoerl a Kennard a počiatky siahajú do roku 1970. Formulujeme ju ako nasledujúcu optimalizačnú úlohu

$$\min_{\beta \in \mathbb{R}^p} \mid\mid \mathbf{Y} - \mathbf{X}\beta \mid\mid^2 + \lambda \mid\mid \beta \mid\mid^2 = \min_{\beta_1, \ldots, \beta_p \in \mathbb{R}} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p} x_{i,j}\beta_j\right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2, \tag{2.10}$$

kde $\lambda \geq 0$ je parameter. Postup a cieľ je podobný metóde OLS - minimalizácia súčtu štvorcov odchýliek $\mathbf{X}\beta$ od $\mathbf{Y}$, ale pod podmienkou penalizačného člena $\lambda \mid\mid \beta \mid\mid^2$. Parameter

λ hovorí o veľkosti penalizácie, ak je $\lambda = 0$ jedná sa o klasickú OLS metódu. Čím je penalizačný parameter väčší, tým sa parametre β_i zmenšujú až k nule. Ekvivalentný zápis úlohy (2.10) je podľa [13]

$$\begin{gathered} \min_{\beta \in \mathbb{R}^p} || \mathbf{Y} - \mathbf{X}\beta ||^2, \\ \text{za podmienky } || \beta ||^2 \leq t. \end{gathered} \tag{2.11}$$

Táto ekvivalencia zaručuje, že úlohu hrebeňovej regresie môžme riešiť ako minimalizáciu súčtu štvorcov odchýliek $\mathbf{X}\beta$ od $\mathbf{Y}$ pod podmienkou, že druhá mocnina dĺžky vektoru β má maximálnu hodnotu t.

Ak existuje v lineárnom regresnom modely veľa korelovaných premenných, odhady koeficientov môžu byť chybné a prejavujú sa vysokým rozptylom. Príliš veľký kladný koeficient jednej premennej môže byť rušený podobne veľkým záporným koeficientom na jej korelovanej premennej. Zavedením obmedzenia veľkosti všetkých koeficientov, ako je to v (2.11), sa tento problém zmierňuje.

Zahrnutie parametru λ do optimalizačnej úlohy vyrieši problém singularity aj ak je matica $\mathbf{X}^T\mathbf{X}$ z (2.3) neinvertovateľná.

Ako sme si vyššie všimli, optimalizačná úloha (2.10) sa skladá zo súčtu dvoch funkcií. V sekcii 2.2 sme dokázali, že je funkcia $|| \mathbf{Y} - \mathbf{X}\beta ||^2$ konvexná. Ďalej triviálne platí, že funkcia $\lambda || \beta ||^2 = \lambda \sum_{j=1}^{p} \beta_j^2$, kde $\lambda > 0$, je tiež konvexná a súčet konvexných funkcií je konvexná funkcia. Preto každé lokálne minimum úlohy (2.10) je aj globálnym minimom.

Optimalizačná úloha (2.10) neobsahuje žiadne obmädzujúce podmienky. Analytické riešenie vypočítame minimalizáciou funkcie $L(\beta)$.

$$\begin{aligned} L(\beta) &= || \mathbf{Y} - \mathbf{X}\beta ||^2 + \lambda || \beta ||^2 \\ &= (\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta) + \lambda\beta^T\beta \\ &= \mathbf{Y}^T\mathbf{Y} - 2\mathbf{Y}^T\mathbf{X}\beta + \beta^T\mathbf{X}^T\mathbf{X}\beta + \lambda\beta^T\beta. \end{aligned}$$

Pod podmienkou optimality musí platiť

$$\frac{\partial L}{\partial \beta}(\beta(\lambda)) = 0, \tag{2.12}$$

pričom zápisom $\beta(\lambda)$ sa zdôrazňuje závislosť riešenia na parametre λ. Nutnú podmienku (2.12) maticovo vyjadríme ako

$$\mathbf{X}^T\mathbf{X}\beta + \lambda\beta = \mathbf{X}^T\mathbf{Y}$$

a sústavou lineárnych rovníc ako

$$(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)\beta = \mathbf{X}^T\mathbf{Y}. \tag{2.13}$$

Matica $(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)$ je vždy regulárna (dokonca aj v prípade, ak je $\mathbf{X}^T\mathbf{X}$ singulárna), pretože k diagonálnym prvkom $\mathbf{X}^T\mathbf{X}$ pripočítavame kladnú konštantu λ. Preto môžme odhadovať koeficienty β pomocou hrebeňovej regresie ako

$$\hat{\beta}_{Ridge} = \hat{\beta}_{Ridge}(\lambda) = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-}\mathbf{X}^T\mathbf{Y}. \tag{2.14}$$

Odhad $\hat{\beta}_{Ridge}$ je (podobne ako odhad $\hat{\beta}_{OLS}$) lineárnou funkciou odozvy $\mathbf{Y}$.

Poznámka 7.

(i) Ak je hodnota parametru t z optimalizačnej úlohy (2.11) *väčšia než hodnota* $|| \hat{\beta}_{OLS} ||^2$*, potom je zjavne* $\hat{\beta}_{OLS}$ *optimálnym riešením.*

(ii) Penalizačný člen obsahuje druhú mocninu dĺžky vektoru β*. Niekedy predpokladáme, že jednotlivé regresné koeficienty príslušných regresorov sú namerané v rovnakých jednotkách. Ak to tak nie je, doporučuje sa štandardizovať ako závisle premenné tak aj regresory. To znamená, že uvažujeme vektor* $\mathbf{Y}^*$ *namiesto vektoru* $\mathbf{Y}$*,*

$$\mathbf{Y}^* = \left(\frac{Y_1 - \bar{Y}}{T_Y}, \ldots, \frac{Y_n - \bar{Y}}{T_Y}\right)^T = \frac{1}{T_Y}(\mathbf{Y} - \bar{Y}\mathbf{1}_n), \tag{2.15}$$

kde je

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \text{ výberový priemer,}$$

$$T_Y = || \mathbf{Y} - \bar{Y}\mathbf{1}_n || = \sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \text{ norma vycentrovanej odozvy,}$$

$$\mathbf{1}_n = (1, \ldots, 1)^T \text{ vektor dimenzie n.}$$

Ďalej sa namiesto matice $\mathbf{X}$ *uvažuje matica* $\mathbf{X}^*$

$$\mathbf{X}^* = \left(\frac{1}{T_1}(\mathbf{x}_1 - \bar{x}_1\mathbf{1}_n), \ldots, \frac{1}{T_p}(\mathbf{x}_p - \bar{x}_p\mathbf{1}_n)\right), \tag{2.16}$$

kde pre k-ty stĺpec matice $\mathbf{X}$ *(*$k \in \{1, \ldots, p\}$*) je*

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_{i,k} \text{ výberový priemer,}$$

$$T_k = || \mathbf{x}_k - \bar{x}_k\mathbf{1}_n || = \sqrt{\sum_{i=1}^{n}(x_{i,k} - \bar{x}_k)^2} \text{ norma vycentrovaného stĺpca.}$$

Po preškálovaní riešime optimalizačnú úlohu

$$\min_{\beta \in \mathbb{R}^p} || \mathbf{Y}^* - \mathbf{X}^*\beta^* ||^2 + \lambda || \beta^* ||^2,$$

kde $\lambda \geq 0$ *je parameter a optimálne riešenie označíme*

$$\hat{\beta}^*_{Ridge} = \left(\hat{\beta}^*_{Ridge,1}, \ldots, \hat{\beta}^*_{Ridge,p}\right)^T.$$

Odhad $\hat{\beta}_{Ridge}$ *vektoru regresných koeficientov* β *príslušného pôvodným nezávislým premenným* $\mathbf{x}_1, \ldots, \mathbf{x}_n$ *spočítame vďaka [20] podľa vzorcu*

$$\hat{\beta}_{Ridge,k} = \hat{\beta}^*_{Ridge,k}\frac{T_Y}{T_k} \quad k = 1, \ldots, p. \tag{2.17}$$

(iii) Hrebeňovú regresiu je možné použiť i v prípade lineárneho modelu s absolútnym členom. Matica plánu má tvar

$$(\mathbf{1}_n \mathbb{X}) = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,p} \\ \vdots & \vdots & \dots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,p} \end{pmatrix}.$$

V tomto prípade má optimalizačná úloha (2.10) *tvar*

$$\min_{\beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p} || \mathbf{Y} - (\beta_0 \mathbf{1}_n + \mathbb{X}\beta) ||^2 + \lambda || \beta ||^2 . \tag{2.18}$$

Parameter β_0, príslušný absolútnemu členu, nie je obsiahnutý v penalizovanom člene. Jeho odhad $\hat{\beta}_{Ridge,0}$ odvodíme nasledovne:

1) *Štandardizáciou závisle premennej $\mathbf{Y}$ a nezávisle premenných $\mathbf{x}_1, \dots, \mathbf{x}_p$ získame novú závislú $\mathbf{Y}^*$ a novú projekčnú maticu*

$$(\mathbf{1}_n, \mathbb{X}) = \left(\mathbf{1}_n, \frac{1}{T_1}(X_1 - \bar{x}_1 \mathbf{1}_n, \dots, \frac{1}{T_p}(X_p - \bar{x}_p \mathbf{1}_n) \right).$$

2) *Optimalizačná úloha* (2.18) *má s novými premennými tvar*

$$\min_{\beta^* \in \mathbb{R}^p} || \mathbf{Y}^* - (\mathbb{X}\beta^*) ||^2 + \lambda || \beta^* ||^2 .$$

3) *Vyriešením úlohy dostávame odhad $\hat{\beta}^*_{Ridge} = (\hat{\beta}^*_{Ridge,1}, \dots, \hat{\beta}^*_{Ridge,p})^T$ vektoru regresných parametrov $(\beta^*_1, \dots, \beta^*_p)^T$.*

4) *Avšak odhad vektoru regresných koeficientov $\beta = (\beta_0, \dots, \beta_p)^T$ je $\hat{\beta}_{Ridge} = (\hat{\beta}_{Ridge,1}, \dots, \hat{\beta}_{Ridge,p})^T$.*

5) *Pomocou* (2.17) *ohadneme $\hat{\beta}_{Ridge,1}, \dots, \hat{\beta}_{Ridge,p}$.*

6) *Odhad $\hat{\beta}_{Ridge,0}$ dopočítame podľa vzorcu*

$$\hat{\beta}_{Ridge,0} = \bar{Y} - \sum_{i=1}^{p} \hat{\beta}^*_{Ridge,i} \bar{x}_i.$$

Teraz nasleduje odpovedať na otázku, ako najlepšie zvoliť hodnotu parametru λ. Hoerl a Kennard v [21] doporučujú vyniesť do grafu komponenty $\hat{\beta}^*_{Ridge}$ voči parametru λ. Vyberie sa čo najmenšia hodnota parametru, pre ktorú sa hodnoty koeficientov ustália a ďalej sú približne konštantné. Táto metóda je dnes kritizovaná, kvôli svojej subjektivite. V praxi sa využíva na určenie λ prevažne krížová validácia.

Ďalej si ukážeme vychýlenosť odhadu $\hat{\beta}^*_{Ridge}$, porovnáme ho s odhadom $\hat{\beta}_{OLS}$ a ukážeme vzťah ich stredných kvadratických chýb.

$$\begin{aligned} \hat{\beta}_{Ridge} &= (\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}^T \mathbf{Y} \\ &= (\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}^T \mathbf{X} \Big((\mathbf{X}^T\mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} \Big) \\ &= (\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}^T \mathbf{X} \hat{\beta}_{OLS} \\ &= \Big(\mathbf{I}_p + \lambda (\mathbf{X}^T\mathbf{X})^{-1} \Big)^{-1} \hat{\beta}_{OLS} \end{aligned} \tag{2.19}$$

Odhad $\hat{\beta}_{OLS}$ je nestranným odhadom vektoru β

$$\mathrm{E}\,\hat{\beta}_{OLS} = \mathrm{E}\Big((\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I}_p)^{-}\mathbf{X}^T\mathbf{Y}\Big) = (\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I}_p)^{-}\mathbf{X}^T\mathbf{X}\beta = \beta$$

Z toho plynie, že $\hat{\beta}_{Ridge}$ je vychýleným odhadom β

$$\begin{aligned}\mathrm{E}\,\hat{\beta}_{Ridge} &= \mathrm{E}\Big\{\Big(\mathbf{I}_p+\lambda(\mathbf{X}^T\mathbf{X})^{-1}\Big)^{-1}\hat{\beta}_{OLS}\Big\}\\ &= \Big(\mathbf{I}_p+\lambda(\mathbf{X}^T\mathbf{X})^{-1}\Big)^{-1}\beta\\ &\neq \beta \quad \text{ak } \lambda \neq 0\end{aligned} \tag{2.20}$$

Na záver sekcie si ukážeme počítanie hrebeňovej regresie pomocou SVD rozkladu. Pri numerickom počítaní sa snažíme vyhnúť invertovaniu matice $\mathbf{X}^T\mathbf{X}$, kvôli výpočetnej náročnosti. Využíva sa singulárny rozklad matice $\mathbf{X}$, viz (2.6)

$$\mathbf{X} = \mathbf{PDQ}^T,$$

Výsledkom SVD rozkladu matice $\mathbf{X}^T\mathbf{X}$ je

$$\begin{aligned}\mathbf{X}^T\mathbf{X} &= \Big(\mathbf{PDQ}^T\Big)^T\mathbf{PDQ}^T\\ &= \mathbf{QD}^T\mathbf{P}^T\mathbf{PDQ}^T\\ &= \mathbf{QD}^2\mathbf{Q}^T.\end{aligned}$$

Potom má odhadovaný vektor $\hat{\beta}_{Ridge}$ tvar

$$\hat{\beta}_{Ridge} = (\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{P}\,\mathrm{diag}\Big\{\frac{d_j}{d_j^2+\lambda}\Big\}\mathbf{Q}^T\mathbf{Y}. \tag{2.21}$$

Dôsledok 4.

(i) Pomocou (2.21) *odvodíme*

$$\begin{aligned}\hat{\mathbf{Y}}_{Ridge} = \mathbf{X}\hat{\beta}_{Ridge} &= \mathbf{X}(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I}_p)^{-1}\mathbf{X}^\top\mathbf{y} = \mathbf{Q}\,\mathrm{diag}\left\{\frac{d_j^2}{d_j^2+\lambda}\right\}\mathbf{Q}^\top\mathbf{y}\\ &= \sum_{j=1}^{n}\Big(\mathbf{q}_j\frac{d_j^2}{d_j^2+\lambda}\mathbf{q}_j^T\Big)\mathbf{Y}.\end{aligned}$$

(ii) Hrebeňová regresia je vo vzťahu nie len s metódou najmenších štvorcov, ale aj s analýzou hlavných komponent.

Odhadnuté hodnoty $\mathbf{y}$ *pomocou metódy hlavných komponent označíme* $\hat{\mathbf{y}}_{\mathrm{OLS}}$.

$$\hat{\mathbf{y}}_{\mathrm{OLS}} = \mathbf{X}\beta_{\mathrm{OLS}} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} = \mathbf{QQ}^\top\mathbf{y}.$$

Odhadnuté hodnoty $\mathbf{y}$ *pomocou metódy hlavných komponent označíme* $\hat{\mathbf{y}}_{\mathrm{PCR}}$ *(PCR = Principal component regression).*

$$\hat{\mathbf{y}}_{\mathrm{PCR}} = \mathbf{X}_{\mathrm{PCA}}\hat{\beta}_{\mathrm{PCR}} = \mathbf{Q}\,\mathrm{diag}\,\{1,\ldots,1,0,\ldots 0\}\,\mathbf{Q}^\top\mathbf{y},$$

kde pri výbere k komponent, je na diagonále k jedničiek nasledovaných (n-k) nulami.

- *Ak* $\lambda = 0$ *potom* $\hat{\mathbf{Y}}_{Ridge} = \hat{\mathbf{Y}}_{OLS}$.
- *Ak* $\lambda > 0$ *potom je sú vyššie singulárne hodnoty* d_j *menej penalizované ako nižšie hodnoty* d_j.
- *Pri PCA sú vysoké singulárne hodnoty* d_j *ponechané a malé (od indexu k) sú úplne odstránené. Tento prístup korešponduje s voľbou* $\lambda = 0$ *pre prvých k prvkov a voľbou* $\lambda = \infty$ *pre zvyšných (n-k) prvkov.*

2.3.3 Metóda Lasso

Táto metóda bola prvýkrát publikovaná v roku 1996. Hlavným zdrojom sekcie je [22]. Názov metódy je akronynom anglického výrazu "Least Absolute Shrinkage and Selection Operator". Jej optimalizačná funkcia je formulovaná ako

$$\min_{\beta\in\mathbb{R}^p} || \mathbf{Y}-\mathbf{X}\beta || +\lambda || \beta ||= \min_{\beta_1,\ldots,\beta_p\in\mathbb{R}} \sum_{i=1}^{n}\left(Y_i-\sum_{j=1}^{p} x_{i,j}\beta_j\right)^2+\lambda\sum_{j=1}^{p} | \beta_j |, \tag{2.22}$$

kde $\lambda \geq 0$ je parameter. Vidíme podobnosť s optimalizačnou úlohou (2.10). Rovnako ako pri hrebeňovej regresie minimalizujeme súčet štvorcov odchýliek $\mathbf{X}\beta$ od $\mathbf{Y}$, ale penalizačný parameter má tvar $\lambda || \beta ||$. Parameter λ opäť udáva veľkosť penalizácie a v prípade, že je nulový sa jedná o metódu OLS. Optimalizačná úloha (2.22) je ekvivalentná optimalizačnej úlohe

$$\begin{gathered}\min_{\beta\in\mathbb{R}^p} || \mathbf{Y}-\mathbf{X}\beta ||^2, \\ \text{za podmienky } || \beta ||\leq t.\end{gathered} \tag{2.23}$$

Optimalizačné úlohy (2.22) a (2.23) sú úlohami kvadratického programovania.

Pre všeobecne zadanú maticu $\mathbb{X}$, na rozdiel od hrebeňovej regresie, neexistuje analytické riešenie. Existuje ale množstvo algoritmov, ktoré hľadajú optimálne riešenie iteratívne. Jedným z najefektívnejších je napríklad LARS (Least Angle Regression) algoritmus, o ktorom sa dočítame v [10]. V špecialnom prípade, ak je matica $\mathbb{X}$ ortogonálna (t.j. platí $\mathbb{X}^T\mathbb{X} = \mathbf{I}_p$), existuje analytické riešenie úlohy (2.22). Toto riešenie budeme značiť

$$\hat{\beta}_{Lasso} = \left(\hat{\beta}_{Lasso,1},\ldots,\hat{\beta}_{Lasso,p}\right)^T$$

a optimalizačná úloha (2.22) je ekvivalentná úlohe

$$\mathbf{Y}^T\mathbf{Y}+\min_{\beta\in\mathbb{R}^p}\{-2\mathbf{Y}^T\mathbf{X}\beta+\beta^T\beta+\lambda || \beta ||\}. \tag{2.24}$$

S využitím odhadu $\hat{\beta}_{OLS}$ v (2.3) sa dá (2.24) prepísať na tvar

$$\mathbf{Y}^T\mathbf{Y}+\min_{\beta_1,\ldots,\beta_p\in\mathbb{R}}\sum_{i=1}^{p}-2\hat{\beta}_{OLS,i}\beta_i+\beta_i^2+\lambda|\beta_i| \tag{2.25}$$

Definujme funkciu $f(\beta_i)$ pre $i \in \{1,\ldots,p\}$ pevné ako

$$f(\beta_i) = -2\hat{\beta}_{OLS,i}\beta_i+\beta_i^2+\lambda|\beta_i|.$$

Takže optimalizačnú úlohu riešime ako

$$\min_{\beta_i \in \mathbb{R}} f(\beta_i). \tag{2.26}$$

Odhad $\hat{\beta}_{OLS,i}$, s ktorým pracujeme, nadobúda dve hodnoty:

1. buď $\hat{\beta}_{OLS,i} > 0$,
2. alebo $\hat{\beta}_{OLS,i} \leq 0$.

Tieto situácie ovplyvňujú riešenie úlohy nasledovne

1. Minimum funkcie f stačí hľadať pre $\beta_i \in [0, \infty)$, pretože pre ľubovoľné pevné $x > 0$ je funkcia

   $$f(x) = -2\hat{\beta}_{OLS,i}x + x^2 + \lambda|x| \quad > \quad 2\hat{\beta}_{OLS,i}x + x^2 + \lambda|x| = f(-x).$$

   V tomto prípade je riešenie optimalizačnej úlohy (2.26)

   $$\frac{\partial f(\beta_i)}{\beta_i} = -2\hat{\beta}_{OLS,i} + 2\beta_i + \lambda = 0, \quad \text{pričom } \beta_i \geq 0.$$

   Takže

   $$\hat{\beta}_{Lasso,i} = \max\left\{\hat{\beta}_{OLS,i} - \frac{1}{2}\lambda, 0\right\}, \tag{2.27}$$

   kde $\hat{\beta}_{OLS,i} > 0$.

2. Minimum funkcie f stačí hľadať pre $\beta_i \in (-\infty, 0]$, pretože pre ľubovoľné pevné $x > 0$ je funkcia

   $$f(x) = -2\hat{\beta}_{OLS,i}x + x^2 + \lambda|x| \quad < \quad 2\hat{\beta}_{OLS,i}x + x^2 + \lambda|x| = f(-x).$$

   V tomto prípade je riešenie optimalizačnej úlohy (2.26)

   $$\frac{\partial f(\beta_i)}{\beta_i} = -2\hat{\beta}_{OLS,i} + 2\beta_i - \lambda = 0, \quad \text{pričom } \beta_i \leq 0.$$

   Takže

   $$\hat{\beta}_{Lasso,i} = \min\left\{\hat{\beta}_{OLS,i} + \frac{1}{2}\lambda, 0\right\} = -\max\left\{-\hat{\beta}_{OLS,i} - \frac{1}{2}\lambda, 0\right\}, \tag{2.28}$$

   kde $\hat{\beta}_{OLS,i} \leq 0$.

Kombináciou oboch situácií vieme odhady (2.27) a (2.28) spoločne vyjadriť pomocou funkcie signum $sgn(\cdot)$ ako

$$\hat{\beta}_{Lasso,i} = sgn\left(\hat{\beta}_{OLS,i}\right) \max\left\{|\hat{\beta}_{OLS,i}| - \frac{1}{2}\lambda, 0\right\}, \quad i = 1, \ldots, p. \tag{2.29}$$

Všeobecne sa dá povedať, že odhad regresných parametrov je pomocou metódy *Lasso* vychýlený. Rastúcou hodnotou parametru λ toto vychýlenie narastá, no zároveň na druhú stranu rozptyl odhadu klesá.

Poznámka 8.

(i) Voľbou $\lambda = 0$ *je optimálnym riešením úlohy* (2.22) *odhad* $\hat{\beta}_{OLS}$.

(ii) Vysoké hodnoty λ *spôsobujú nulovanenie niektorých regresných koeficientov a iné koeficienty majú v absolútnej hodnote veľmi malé hodnoty. Metóda tak môže slúžiť ako nástroj pre výber vhodných regresorov. Viac sa dočítate v [19], kde je tento prístup rozobraný podrobnejšie.*

(iii) Ak je hodnota parametru t z optimalizačnej úlohy (2.23) *väčšia než hodnota* $|| \hat{\beta}_{OLS} ||$, *potom je zjavne* $\hat{\beta}_{OLS}$ *optimálnym riešením (rovnaký scenár ako pri hrebeňovej regresii).*

(iv) Je rozumné predpokladať, aby boli premenné v modeli nezávislé a namerané v rovnakých jednotkách. V opačnom prípade je vhodné závisle aj nezávislé premenné štandardizovať. Nové premenné potom vstupujú do optimalizačnej úlohy (2.22). *Riešením úlohy je odhad regresných parametrov pre príslušnú štandardizáciu. Analógiou* (2.17) *vypočítame odhady jednotlivých zložiek vektoru regresných parametrov príslušného pôvodným premenným.*

(v) V prípade lineárneho modelu s absolútnym členom, je postup odhadu vektoru regresných koeficientov analogický ako v poznámke 7 pre hrebeňovú regresiu.

2.4 Porovnanie metód

V tejto sekcii sa bude ilustrovať a porovnávať robustnosť odhadov vektoru regresných parametrov metódou najmenších štvorcov (OLS), hrebeňovou regresiou (Ridge) a metódou Lasso.

V častiach 2.4.1- 2.4.4 sa pracuje s náhodne generovanými dátami. Pre generovanie a spracovávanie dát, taktiež pre implementáciu jednotlivých odhadov, používame štatistický program R. V tabuľke 2.1 sú názvy balíčkov a funkcií použitých na výpočet odhadov vektoru regresných parametrov.

Metóda	R balíček	Funkcia
OLS	stats	lm
Ridge	mass	lm.ridge
Lasso	lars	lars

Tabuľka 2.1: Prehľad použitých balíčkov a funkcií potrebných k odhadom vektoru regresných parametrov.

V prípade hrebeňovej regresie riešime optimalizačnú úlohu (2.10) a optimálnu hodnotu penalizačného parametru λ hľadáme pomocou krížovej validácie. V simulácii s prítomnosťou chyby merania použijeme zovšeobecnenú krížovú validáciu, ktorá je komentovaná v článku [15].

Odhad vektoru regresných koeficientov metódou Lasso vychádza z optimalizačnej úlohy

$$\begin{gathered}\min_{\beta\in\mathbb{R}^p} || \mathbf{Y}-\mathbf{X}\beta ||^2, \\ \text{za podmienky } || \beta ||\leq s || \hat{\beta}_{\mathbf{OLS}} ||,\end{gathered} \tag{2.30}$$

kde $s \in [0,1]$ je parameter. Pôvodná optimalizačná úloha definovaná v (2.23) je ekvivalentná s optimalizačnou úlohou (2.30), ak položíme $t = s || \hat{\beta}_{\mathbf{OLS}} ||$. Optimálnu hodnotu penalizujúceho parametru s hľadáme pomocou krížovej validácie. Vidieť, že v prípade voľby $s = 1$ je optimálnym riešením úlohy (2.30) odhad $\hat{\beta}_{OLS}$.

2.4.1 Simulačná štúdia s odľahlím pozorovaním

Pomocou metód v tabuľke 2.1 budeme skúmať robustnosť jednotlivých odhadov voči prítomnosti odľahlého pozorovania. To znamená, že porovnávame odhady vektoru regresných koeficientov na dátach s odhalím pozorovaním a po odstránení odľahlého pozorovania.

Vygenerované dáta majú charakteristiky:

- počet pozorovaní $n = 30$,
- počet regresorov $p = 2$,
- skutočná hodnota regresorov $\beta = (\beta_0, \beta_1)^T = (-1, 2)^T$,
- $x_{1,0} = \cdots = x_{0,n} = 1$,
- $X_1 \sim N(0;1)$ s realizáciami $x_{1,1}, \ldots, x_{n-1,1}$,
- $\varepsilon \sim N(0; 0,3^2)$ s realizáciami $\varepsilon_1, \ldots, \varepsilon_{n-1}$,
- $Y_i = \beta_0 x_{i,0} + \beta_1 x_{i,1} + \varepsilon_i$, kde $i = 1, \ldots, n-1$,
- odľahlé pozorovanie $((x_{n,0}, x_{n,1}), Y_n) := ((0, 2 * \max\{x_{1,1}, \ldots, x_{n-1,1}\}), 0)$.

V maticovom zápise

$$\mathbb{X}^{\text{op}}_{30\times 2} = \begin{pmatrix} 1 & x_{1,1} \\ \vdots & \vdots \\ 1 & x_{29,1} \\ 0 & x_{30,1} \end{pmatrix}, \quad \mathbf{Y}^{\text{op}}_{30\times 1} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_{29} \\ 0 \end{pmatrix}, \quad \beta^{\text{op}}_{2\times 1} = \begin{pmatrix} \beta_0^{\text{op}} \\ \beta_1^{\text{op}} \end{pmatrix} = \begin{pmatrix} -1 \\ 2 \end{pmatrix},$$

$$\mathbb{X}^{\text{bop}}_{29\times 2} = \begin{pmatrix} 1 & x_{1,1} \\ \vdots & \vdots \\ 1 & x_{29,1} \end{pmatrix}, \quad \mathbf{Y}^{\text{bop}}_{29\times 1} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_{29} \end{pmatrix}, \quad \beta^{\text{bop}}_{2\times 1} = \begin{pmatrix} \beta_0^{\text{bop}} \\ \beta_1^{\text{bop}} \end{pmatrix} = \begin{pmatrix} -1 \\ 2 \end{pmatrix},$$

kde *op* značí prítomnosť odľahlého pozorovania a *bop* je tvar po odstránení odľahlého pozorovania.

V tabuľke 2.2 sa nachádzajú odhady všetkých regresných koeficientov prislúchajúce uvažovaným regresným metódam. Koeficienty β_0^{op} a β_0^{op} sú označené ako $\hat{\beta}_0^{\text{op}}$ a $\hat{\beta}_0^{\text{op}}$. Pomocou krížovej validácie sa v hrebeňovej regresii určila hodnota penalizujúci parameter λ rovná 7.06 a parameter s pre metódu Lasso má hodnotu 0.79.

Koeficienty β_0^{bop} a β_0^{bop} sú označené ako $\hat{\beta}_0^{\text{bop}}$ a $\hat{\beta}_0^{\text{bop}}$. Pomocou krížovej validácie sa v hrebeňovej regresii určila hodnota penalizujúci parameter $\lambda = 0$. Parameter s, použitý v metóde Lasso, má hodnotu 1. To znamená, že že odhady $\hat{\beta}_0^{\text{bop}}, \hat{\beta}_1^{\text{bop}}$ sú pre Ridge a Lasso totožné s odhadom metódou najmenších štvorcov.

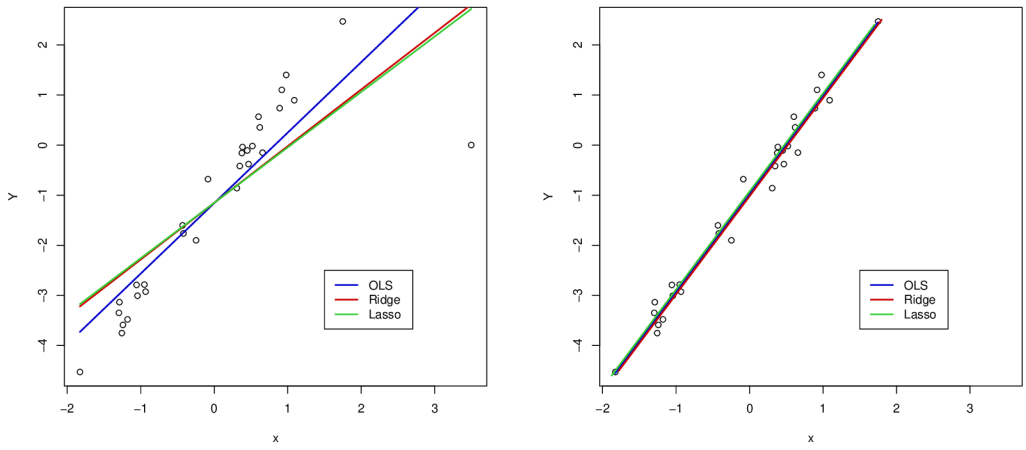
V tabuľke 2.2 nájdeme aj absolútne rozdiely odhadov $\hat{\beta}_0^{\text{op}} - \hat{\beta}_0^{\text{bop}}$ a $\hat{\beta}_1^{\text{op}} - \hat{\beta}_1^{\text{bop}}$. Najväčšieho rozdielu nadobúdajú odhady metódou Lasso a najmenšie metódou hrebeňovej regresie. Vidíme, že odhady pre dáta bez odľahlého pozorovania sú zhodné.

regresná metóda	$\hat{\beta}_0^{\text{op}}$	$\hat{\beta}_1^{\text{op}}$	$\hat{\beta}_0^{\text{bop}}$	$\hat{\beta}_1^{\text{bop}}$	$\lvert \hat{\beta}_0^{\text{op}} - \hat{\beta}_0^{\text{bop}} \rvert$	$\lvert \hat{\beta}_1^{\text{op}} - \hat{\beta}_1^{\text{bop}} \rvert$
OLS	-1.159	1.407	-0.973	1.959	0.1851	0.5515
Ridge	-1.153	1.132	-0.973	1.959	0.1770	0.8270
Lasso	-1.153	1.107	-0.973	1.959	0.1792	0.8521

Tabuľka 2.2: Odhady vektoru regresných koeficientov pre dáta s odľahlým pozorovaním a pre rovnaké dáta bez odľahlého pozorovania.

Pre geometrickú predstavu sú tieto odhady ilustrované na obrázkoch 2.1a-2.1e. Regresné priamky sú spočítané na základe odhadnutých parametrov. Keďže sú parametre $\hat{\beta}_0^{\text{bop}}, \hat{\beta}_1^{\text{bop}}$ totožné, tak sú aj krivky všetkých metód zhodné, viz obrázok 2.1b. Na obrázkoch 2.1c-2.1e sú červené priamky zkonštruované na základe $\hat{\beta}_0^{\text{bop}}, \hat{\beta}_1^{\text{bop}}$ a modré priamky na základe $\hat{\beta}_0^{\text{op}}, \hat{\beta}_1^{\text{op}}$.

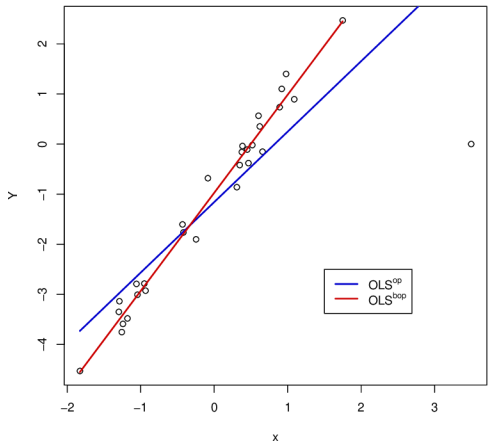
Záver úlohy je, že pri prítomnosti odľahlého pozorovania sú odhady Ridge a Lasso menej robustné než metóda najmenších štvorcov. Takže pokiaľ máme v našej vzorke dát podozrenie na prítomnosť odlahlých pozorovaní, nie je všeobecne lepšie preferovať upravenú regresiu nad klasickou metódou najmenších štvorcov.



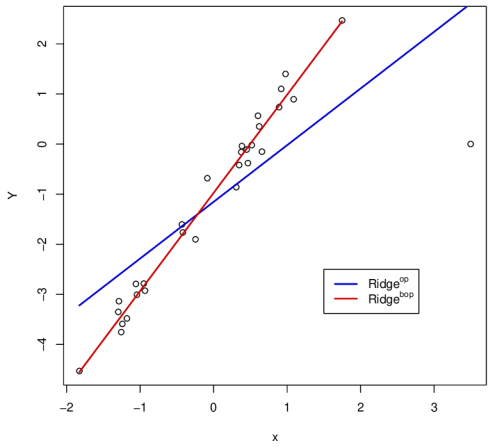
(a) Regresné priamky s prítomnosťou odľahlého pozorovania.

(b) Regresné priamky po odstránení odľahlého pozorovania.

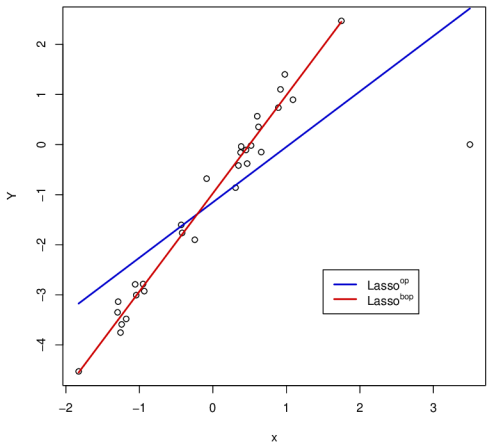
Obr. 2.1: Odhady regresných priamok pre všetky spomínané metódy na dátach s/bez prítomnosti odľahlého pozorovania.



(c) Regresné priamky na základe odhadu metódy OLS.



(d) Regresné priamky na základe odhadu hrebeňovej regresie.



(e) Regresné priamky na základe odhadu metódy Lasso.

Obr. 2.1: Odhady regresných priamok pre všetky spomínané metódy na dátach s/bez prítomnosti odľahlého pozorovania.

2.4.2 Simulačná štúdia v prítomnosti chyby merania

V celej práci pracujeme s lineárnym regresným modelom, pre ktorý platí

$$\mathrm{E}\mathbf{Y} = \mathbb{X}\beta. \tag{2.31}$$

Pri reálnych dátach sa často môžu vyskytnúť nejaké nepresnosti. Preto nás zaujíma ako jednotlivé regresné metódy reagujú v prítomnosti chýb merania vysvetľujúcich premenných. Najprv budeme simulovať odhady regresných parametrov podľa rovnice (2.31). Potom zohľadníme možnosť, že sú vysvetľujúce premenné zaťažené chybou merania, a preto pred-

pokladáme platnosť nasledujúceho modelu

$$\mathrm{E}Y_i = (\mathbf{x}_i + \Delta\mathbf{x}_i)^T \beta + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2.32}$$

ktorý je možné ekvivalentne zapísať ako

$$\mathrm{E}Y_i = \beta_1(x_{i,1} + \Delta x_{i,1}) + \cdots + \beta_p(x_{i,p} + \Delta x_{i,p}), \quad i = 1, \ldots, n, \tag{2.33}$$

pričom p je počet parametrov modelu a $\Delta x_{1,j}, \ldots, \Delta x_{n,j}, j = 1, \ldots, p$ sú realizácie náhodnej veličiny ΔX_j s rozdelením $N(0, \delta_j^2)$, kde $\delta_j \geq 0$ je známe. Takže chyby merania sú vo forme matice $\Delta\mathbb{X} = (\Delta x_{i,j})_{i,j=1}^{n,p}$

Vygenerované dáta majú nasledovné charakteristiky:
- počet pozorovaní $n = 30$,
- počet regresorov $p = 3$,
- skutočná hodnota regresorov $\beta = (\beta_0, \beta_1, \beta_2)^T = (-3, 2, -1)^T$,
- $x_{1,0} = \cdots = x_{0,n} = 1$,
- $X_1 \sim N(5; 1)$ s realizáciami $x_{1,1}, \ldots, x_{n,1}$,
- $X_2 \sim N(0; 2^2)$ s realizáciami $x_{1,2}, \ldots, x_{n,2}$,
- $\Delta X_1 \sim N(0; 0,001)$ s realizáciami $\Delta x_{1,1}, \ldots, \Delta x_{n,1}$,
- $\Delta X_2 \sim N(0; 0,05^2)$ s realizáciami $\Delta x_{1,2}, \ldots, \Delta x_{n,2}$,
- $\varepsilon \sim N(0; 0,2^2)$ s realizáciami $\varepsilon_1, \ldots, \varepsilon_n$,
- $Y_i = \beta_0 x_{i,0} + \beta_1(x_{i,1} + \Delta x_{i,1}) + \beta_2(x_{i,2} + \Delta x_{i,2}) + \varepsilon_i$, kde $i = 1, \ldots, n$.

V maticovom zápise

$$\mathbb{X}^0_{30\times 3} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} \\ \vdots & \vdots & \vdots \\ 1 & x_{30,1} & x_{30,2} \end{pmatrix}, \quad \mathbf{Y}^0_{30\times 1} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_{30} \end{pmatrix} = \mathbf{Y}^\Delta_{30\times 1},$$

$$\beta^0_{3\times 1} = \begin{pmatrix} \beta_0^0 \\ \beta_1^0 \\ \beta_2^0 \end{pmatrix} = \begin{pmatrix} -3 \\ 2 \\ -1 \end{pmatrix} = \begin{pmatrix} \beta_0^\Delta \\ \beta_1^\Delta \\ \beta_2^\Delta \end{pmatrix} = \beta^\Delta_{3\times 1},$$

$$\mathbb{X}^\Delta_{30\times 3} = \mathbb{X}^0_{30\times 3} + \mathbb{X}^\Delta_{30\times 3} =$$

$$= \begin{pmatrix} 1 & x_{1,1} & x_{1,2} \\ \vdots & \vdots & \vdots \\ 1 & x_{30,1} & x_{30,2} \end{pmatrix} + \begin{pmatrix} 0 & \Delta x_{1,1} & \Delta x_{1,2} \\ \vdots & \vdots & \vdots \\ 0 & \Delta x_{30,1} & \Delta x_{30,2} \end{pmatrix} =$$

$$= \begin{pmatrix} 1 & x_{1,1} + \Delta x_{1,1} & x_{1,2} + \Delta x_{1,2} \\ \vdots & \vdots & \vdots \\ 1 & x_{30,1} + \Delta x_{30,1} & x_{30,2} + \Delta x_{30,2} \end{pmatrix},$$

kde *0* značí dáta, ktoré nezohľadňujú chybu merania a Δ značí dáta, ktoré zohľadňujú chybu meriania. V tabuľke 2.3 sa nachádzajú odhady všetkých regresných koeficientov prislúchajúce uvažovaným regresným metódam. V prípade ignorovania chýb merania sú odhady koeficientov $\beta_i^0, i = 0, 1, 2$, označené ako $\hat{\beta}_i^0, i = 0, 1, 2$. Pomocou krížovej validácie

sa v hrebeňovej regresii určila hodnota penalizujúci parameter λ rovná 0.15 a parameter s pre metódu Lasso má hodnotu 0.9942.

Na druhej strane, pri zohľadnení chýb merania, označujeme odhady koeficientov β_i^{Δ}, $i = 0,1,2$, ako $\hat{\beta}_i^{\Delta}$, $i = 0,1,2$. Pomocou krížovej validácie sa v hrebeňovej regresii určila hodnota penalizujúci parameter $\lambda = 0.14$. Parameter s, použitý v metóde Lasso, má hodnotu 0.9948. V tabuľke 2.4 nájdeme absolútne rozdiely jednotlivých odhadov. Z uvažovaných regresných metód sa zdá byť metóda Lasso najviac robustná voči prítomnosti chýb merania. Ďalej nasleduje hrebeňová regresia a v tesnom závese metóda najmenších štvorcov, ktorá je najviac citlivá na výskyt chyby merania.

regresná metóda	$\hat{\beta}_0^0$	$\hat{\beta}_1^0$	$\hat{\beta}_2^0$	RSS^0	$\hat{\beta}_0^{\Delta}$	$\hat{\beta}_1^{\Delta}$	$\hat{\beta}_2^{\Delta}$	RSS^{Δ}
OLS	-3.1742	2.0258	-1.0463	1.1730	-3.0727	2.0076	-1.0345	1.0857
Ridge	-3.1697	2.0250	-1.0458	1.1730	-3.0686	2.0068	-1.0345	1.0857
Lasso	-3.0930	2.0107	-1.0413	1.1796	-3.0003	1.9941	-1.0302	1.0910

Tabuľka 2.3: Odhady vektoru regresných koeficientov a príslušnej odchýlky pre dáta s ohľadom a bez ohľadu na chyby merania vysvetľujúcich parametrov.

regresná metóda	$\lvert \hat{\beta}_0^0 - \hat{\beta}_0^{\Delta} \rvert$	$\lvert \hat{\beta}_1^0 - \hat{\beta}_1^{\Delta} \rvert$	$\lvert \hat{\beta}_2^0 - \hat{\beta}_2^{\Delta} \rvert$
OLS	0.10163	0.01824	0.01175
Ridge	0.10117	0.01817	0.01172
Lasso	0.09271	0.01659	0.01118

Tabuľka 2.4: Absolútne hodnoty rozdielov odhadov regresných parametrov pre dáta s ohľadom a bez ohľadu na chyby merania vysvetľujúcich parametrov.

2.4.3 Simulačná štúdia v prítomnosti multikolinearity (1)

Vygenerujeme si dáta s nižšie popísanými charakteristikami a testujeme odhad koeficientov regresnými metódami.

- počet pozorovaní $n = 30$,
- počet regresorov $p = 3$,
- skutočná hodnota regresorov $\beta = (\beta_0, \beta_1, \beta_2)^T = (-5, 2, -2)^T$,
- $x_{1,0} = \cdots = x_{0,n} = 1$,
- $X_1 \sim N(5; 2^2)$ s realizáciami $x_{1,1}, \ldots, x_{n,1}$,
- $Z \sim N(0; 0, 7^2)$ s realizáciami $z_1, \ldots, z_n$,
- $X_2 = X_1 + Z$ s realizáciami $x_{1,2} = x_{1,1} + z_1, \ldots, x_{n,2} = x_{n,1} + z_n$,
- $\Delta X_1 \sim N(0; 0, 009^2)$ s realizáciami $\Delta x_{1,1}, \ldots, \Delta x_{n,1}$,
- $\Delta X_2 \sim N(0; 0, 006^2)$ s realizáciami $\Delta x_{1,2}, \ldots, \Delta x_{n,2}$,
- $\varepsilon \sim N(0; 0, 2^2)$ s realizáciami $\varepsilon_1, \ldots, \varepsilon_n$,
- $Y_i = \beta_0 x_{i,0} + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \varepsilon_i$, kde $i = 1, \ldots, n$.

V maticovom zápise

$$\mathbb{X}^{m,0}_{30\times 3} = \begin{pmatrix} 1 & x_{1,1} & x_{1,1}+z_1 \\ \vdots & \vdots & \vdots \\ 1 & x_{30,1} & x_{30,1}+z_{30} \end{pmatrix}, \mathbf{Y}^{m,0}_{30\times 1} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_{30} \end{pmatrix} = \mathbf{Y}^{m,\Delta}_{30\times 1},$$

$$\beta^{m,0}_{3\times 1} = \begin{pmatrix} \beta^{m,0}_0 \\ \beta^{m,0}_1 \\ \beta^{m,0}_2 \end{pmatrix} = \begin{pmatrix} -5 \\ 2 \\ -2 \end{pmatrix} = \begin{pmatrix} \beta^{m,\Delta}_0 \\ \beta^{m,\Delta}_1 \\ \beta^{m,\Delta}_2 \end{pmatrix} = \beta^{m,\Delta}_{3\times 1},$$

$$\mathbb{X}^{m,\Delta}_{30\times 3} = \mathbb{X}^{0}_{30\times 3} + \mathbb{X}^{\Delta}_{30\times 3} =$$

$$= \begin{pmatrix} 1 & x_{1,1} & x_{1,1}+z_1 \\ \vdots & \vdots & \vdots \\ 1 & x_{30,1} & x_{30,1}+z_{30} \end{pmatrix} + \begin{pmatrix} 0 & \Delta x_{1,1} & \Delta x_{1,2} \\ \vdots & \vdots & \vdots \\ 0 & \Delta x_{30,1} & \Delta x_{30,2} \end{pmatrix} =$$

$$= \begin{pmatrix} 1 & x_{1,1}+\Delta x_{1,1} & x_{1,2}+z_1+\Delta x_{1,2} \\ \vdots & \vdots & \vdots \\ 1 & x_{30,1}+\Delta x_{30,1} & x_{30,2}+z_{30}+\Delta x_{30,2} \end{pmatrix},$$

kde *m,0* označuje dáta, ktoré nezohľadňujú chybu merania a m,Δ označuje dáta, ktoré zohľadňujú chybu meriania.

Výberový korelačný koeficient medzi X_1 a X_2 je 0.87, takže tieto regresory môžu spôsobovať multikolinearitu. Pri hrozbe multikolinearity treba skontrolovať aj podmienenosť matice. V tomto prípade je matica $\mathbb{X}$ dobre podmienená a disponuje nízkym číslom podmienenosti $\kappa = 17.6$. Situáciu so zle podmienenou maticou rozoberáme v simulačnom príklade 2.4.4.

V tabuľke 2.5 sa nachádzajú odhady všetkých regresných koeficientov prislúchajúce uvažovaným regresným metódam. V prípade ignorovania chýb merania sú odhady koeficientov $\beta^{m,0}_i, i = 0,1,2$, označené ako $\hat{\beta}m,0_i, i = 0,1,2$. Pomocou krížovej validácie sa v hrebeňovej regresii určila hodnota penalizujúci parameter λ rovná 0.15 a parameter s pre metódu Lasso má hodnotu 0.9942.

Na druhej strane, pri zohľadnení chýb merania označujeme odhady koeficientov $\beta^{m,\Delta}_i$, $i = 0,1,2$, ako $\hat{\beta}^{m,\Delta}_i$, $i = 0,1,2$. Pomocou krížovej validácie sa v hrebeňovej regresii určila hodnota penalizujúci parameter $\lambda = 0.14$. Parameter s, použitý v metóde Lasso, má hodnotu 0.9948. V tabuľke 2.6 nájdeme absolútne rozdiely jednotlivých odhadov. Z uvažovaných regresných metód sa zdá byť metóda Lasso najviac robustná voči prítomnosti chýb merania a multikolinearity. Ďalej nasleduje hrebeňová regresia a v tesnom závese metóda najmenších štvorcov, ktorá je najviac citlivá na výskyt chyby merania a multikolinearity.

regresná metóda	$\hat{\beta}^{m,0}_0$	$\hat{\beta}^{m,0}_1$	$\hat{\beta}^{m,0}_2$	$RSS^{m,0}$	$\hat{\beta}^{m,\Delta}_0$	$\hat{\beta}^{m,\Delta}_1$	$\hat{\beta}^{m,\Delta}_2$	$RSS^{m,\Delta}$
OLS	-4.8655	2.0118	-2.0318	1.1864	-4.8595	2.0132	-2.0351	1.1032
Ridge	-4.8648	2.0102	-2.0303	1.1865	-4.8589	2.0116	-2.0336	1.1032
Lasso	-4.8625	2.0047	-2.0251	1.1871	-4.8562	2.0051	-2.0275	1.1041

Tabuľka 2.5: Odhady vektoru regresných koeficientov a príslušnej odchýlky pre dáta s ohľadom a bez ohľadu na chyby merania vysvetľujúcich parametrov.

regresná metóda	$\lvert \hat{\beta}_0^{m,0} - \hat{\beta}_0^{m,\Delta} \rvert$	$\lvert \hat{\beta}_1^{m,0} - \hat{\beta}_1^{m,\Delta} \rvert$	$\lvert \hat{\beta}_2^{m,0} - \hat{\beta}_2^{m,\Delta} \rvert$
OLS	0.00598	0.00145	0.00336
Ridge	0.00596	0.00145	0.00336
Lasso	0.00635	0.00040	0.00237

Tabuľka 2.6: Absolútne hodnoty rozdielov odhadov regresných parametrov pre dáta s ohľadom a bez ohľadu na chyby merania vysvetľujúcich parametrov.

2.4.4 Simulačná štúdia v prítomnosti multikolinearity (2)

V tejto simulačnej štúdii máme rovnaké zadanie ako v príklade 2.4.3, je prítomná multikolinearita a chyby meriania. Zaujíma nás robustnosť jednotlivých metód pri zle podmienenej matici plánu. Vygenerujeme si dáta s nasledovnými charakteristikami

- počet pozorovaní $n = 30$,
- počet regresorov $p = 5$,
- skutočná hodnota regresorov $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)^T = (3, -1, 1, 4)^T$,
- $x_{1,0} = \cdots = x_{n,0} = 1$,
- $X_1 : x_{1,1} = 1/30, x_{2,1} = 2/30, \ldots, = x_{n,1} = 1$,
- $Z^1 \sim N(0; 0,002^2)$ s realizáciami $z_1^1, \ldots, z_n^1$,
- $X_2 = X_1 + Z^1$ s realizáciami $x_{1,2} = x_{1,1} + z_1^1, \ldots, x_{n,2} = x_{n,1} + z_n^1$,
- $Z^2 \sim N(0; 0,001^2)$ s realizáciami $z_1^2, \ldots, z_n^2$,
- $X_3 = 2X_1 + Z^2$ s realizáciami $x_{1,3} = 2x_{1,1} + z_1^2, \ldots, x_{n,3} = 2x_{n,1} + z_n^2$,
- $\Delta X_1 \sim N(0; 0,009^2)$ s realizáciami $\Delta x_{1,1}, \ldots, \Delta x_{n,1}$,
- $\Delta X_2 \sim N(0; 0,006^2)$ s realizáciami $\Delta x_{1,2}, \ldots, \Delta x_{n,2}$,
- $\Delta X_3 \sim N(0; 0,002^2)$ s realizáciami $\Delta x_{1,3}, \ldots, \Delta x_{n,3}$,
- $\varepsilon \sim N(0; 0, 2^2)$ s realizáciami $\varepsilon_1, \ldots, \varepsilon_n$,
- $Y_i = \beta_0 x_{i,0} + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \varepsilon_i$, kde $i = 1, \ldots, n$.

V maticovom zápise

$$\mathbb{X}_{30\times 4}^{c,0} = \begin{pmatrix} 1 & x_{1,1} & x_{1,1}+z_1^1 & 2x_{1,1}+z_1^2 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{30,1} & x_{30,1}+z_{30}^1 & 2x_{1,1}+z_{30}^2 \end{pmatrix}, \mathbf{Y}_{30\times 1}^{c,0} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_{30} \end{pmatrix} = \mathbf{Y}_{30\times 1}^{c,\Delta},$$

$$\beta_{4\times 1}^{c,0} = \begin{pmatrix} \beta_0^{c,0} \\ \beta_1^{c,0} \\ \beta_2^{c,0} \\ \beta_3^{c,0} \end{pmatrix} = \begin{pmatrix} 3 \\ -1 \\ 1 \\ 4 \end{pmatrix} = \begin{pmatrix} \beta_0^{c,\Delta} \\ \beta_1^{c,\Delta} \\ \beta_2^{c,\Delta} \\ \beta_3^{c,\Delta} \end{pmatrix} = \beta_{4\times 1}^{c,\Delta},$$

$$\mathbb{X}_{30\times 4}^{c,\Delta} = \mathbb{X}_{30\times 4}^{0} + \mathbb{X}_{30\times 4}^{\Delta} =$$

$$= \begin{pmatrix} 1 & x_{1,1} & x_{1,1}+z_1^1 & 2x_{1,1}+z_1^2 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{30,1} & x_{30,1}+z_{30}^1 & 2x_{1,1}+z_{30}^2 \end{pmatrix} + \begin{pmatrix} 0 & \Delta x_{1,1} & \Delta x_{1,2} & \Delta x_{1,3} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \Delta x_{30,1} & \Delta x_{30,2} & \Delta x_{30,3} \end{pmatrix} =$$

$$= \begin{pmatrix} 1 & x_{1,1}+\Delta x_{1,1} & x_{1,2}+z_1^1+\Delta x_{1,2} & 2x_{30,1}+z_1^2+\Delta x_{30,1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{30,1}+\Delta x_{30,1} & x_{30,2}+z_{30}^1+\Delta x_{30,2} & 2x_{30,3}+z_{30}^2+\Delta x_{30,3} \end{pmatrix},$$

kde *c,0* označuje dáta, ktoré nezohľadňujú chybu merania a c,Δ označuje dáta, ktoré zohľadňujú chybu meriania.

Korelačná matica medzi 2.,3. a 4. regresným koeficientom matice plánu je

X_2	1		
X_3	0.9999898	1	
X_4	0.9999985	0.9999874	1
	X_2	X_3	X_4

Vidieť, že výberové korelačné koeficienty majú vysoké hodnoty. To značí prítomnosť silnej pozitívnej korelácie a tieto stĺpce zapríčiňujú multikolinearitu. Podmienenosť matice plánu je veľmi zlá, číslo podmienenosti κ je veľmi vysoké, má hodnotu 4274.8.

V tabuľke 2.7 a 2.8 sa nachádzajú odhady všetkých regresných koeficientov prislúchajúce uvažovaným regresným metódam. V prípade ignorovania chýb merania sú odhady koeficientov $\beta_i^{c,0}, i = 0,\ldots,3$, označené ako $\hat{\beta}c,0_i, i = 0,\ldots,3$. Pomocou krížovej validácie sa v hrebeňovej regresii určila hodnota penalizujúci parameter λ rovná 0.105 a parameter s pre metódu Lasso má hodnotu 0.5062.

Na druhej strane, pri zohľadnení chýb merania označujeme odhady koeficientov $\beta_i^{c,\Delta}$, $i = 0,\ldots,3$, ako $\hat{\beta}_i^{c,\Delta}$, $i = 0,\ldots,3$. Pomocou krížovej validácie sa v hrebeňovej regresii určila hodnota penalizujúci parameter $\lambda = 0.402$. Parameter s, použitý v metóde Lasso, má hodnotu 0.937. V tabuľke 2.8 nájdeme absolútne rozdiely jednotlivých odhadov. V prípade zle podmienej matice sú rozdiely v metódach najviac výrazné. Metóda Lasso je jednoznačne najviac robustná voči prítomnosti chýb merania a multikolinearity pre zle podmienenú maticu. Ďalej nasleduje hrebeňová regresia a najviac citlivou metódou je metóda najmenších štvorcov.

regresná metóda	$\hat{\beta}_0^{c,0}$	$\hat{\beta}_1^{c,0}$	$\hat{\beta}_2^{c,0}$	$\hat{\beta}_3^{c,0}$	$\mathrm{RSS}^{c,0}$
OLS	2.9850	-57.3881	10.9533	27.2026	1.0968
Ridge	2.9967	2.6273	2.6613	1.3349	1.1186
Lasso	2.9887	-27.0793	6.2840	14.3824	1.1023

Tabuľka 2.7: Odhady vektoru regresných koeficientov a príslušnej odchýlky pre dáta bez ohľadu na chyby merania vysvetľujúcich parametrov.

regresná metóda	$\hat{\beta}_0^{c,\Delta}$	$\hat{\beta}_1^{c,\Delta}$	$\hat{\beta}_2^{c,\Delta}$	$\hat{\beta}_3^{c,\Delta}$	$\mathrm{RSS}^{c,\Delta}$
OLS	2.9916	1.6707	-3.5550	4.9306	1.1118
Ridge	3.0030	2.6218	2.6436	1.3398	1.1244
Lasso	2.9911	1.7111	-3.0786	4.6723	1.1118

Tabuľka 2.8: Odhady vektoru regresných koeficientov a príslušnej odchýlky pre dáta s ohľadom na chyby merania vysvetľujúcich parametrov.

regresná metóda	$\lvert \hat{\beta}_0^{c,0} - \hat{\beta}_0^{c,\Delta} \rvert$	$\lvert \hat{\beta}_1^{c,0} - \hat{\beta}_1^{c,\Delta} \rvert$	$\lvert \hat{\beta}_2^{c,0} - \hat{\beta}_2^{c,\Delta} \rvert$	$\lvert \hat{\beta}_3^{c,0} - \hat{\beta}_3^{c,\Delta} \rvert$
OLS	0.00660	59.05878	14.50830	22.27207
Ridge	0.00637	0.00550	0.01765	0.00486
Lasso	0.00239	28.79039	9.36258	9.71004

Tabuľka 2.9: Absolútne hodnoty rozdielov odhadov regresných parametrov pre dáta s ohľadom a bez ohľadu na chyby merania vysvetľujúcich parametrov.

Kapitola 3

Klasifikačná analýza

Táto kapitola uvádza princíp klasifikácie. Zavádza definície a ilustruje praktický význam. Popisuje vytváranie algoritmov a ich výpočetnú náročnosť. Ďalej predstavuje konkrétny model, ktorý sa pri problémoch klasifikácie využíva.

3.1 Vstupné dáta

Na začiatku máme n známych vektorov pozorovania a každý z nich prislúcha práve do jednej z g skupín. Vektor pozorovania je tvorený p premennými. Triviálne predpokladáme $n > g$. Za týchto podmienok definujeme pojmy a maticu vstupných údajov klasifikačného problému.

Definícia 1. *Nech pre každé $i \in \{1,\dots,n\}$, $j \in \{1,\dots,p\}$ je i-ty vektor pozorovania $\mathbf{x}_i$, ktorého hodnota v j-tej premennej je $x_{i,j}$. Potom výberová matica $\mathbf{X} \in \mathbb{R}^{p\times n}$ je matica po stĺpcoch tvorená vektormi pozorovaní, tj.*

$$\mathbf{X} = \begin{pmatrix} | & & | \\ \mathbf{x}_1 & \dots & \mathbf{x}_n \\ | & & | \end{pmatrix} \in \mathbb{R}^{p\times n}$$

Maticu kódovania skupín voláme takú maticu $G \in \mathbb{R}^{n\times g}$, pre ktorú platí $G_{i,j} = 1$, ak i-te pozorovanie patrí do j-tej skupiny, a $G_{i,j} = 0$ v opačnom prípade. Po vhodnom zoradení pozorovaní $\mathbf{x}_i$ môže mať matica $\mathbb{G}$ tvar

$$\mathbb{G} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ \vdots & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \\ & \ddots & \ddots & \\ 0 & \dots & 0 & 1 \\ 0 & \dots & 0 & 1 \\ 0 & \dots & 0 & 1 \end{pmatrix} \in \mathbb{R}^{n\times g}.$$

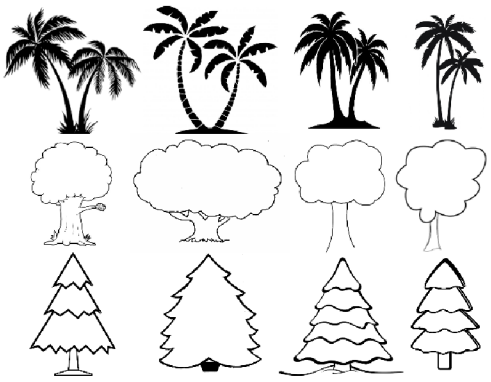
Takto definovaná matica $\mathbb{G}$ je našim vstupným údajom. Cieľom klasifikácie je prijať nové pozorovanie (ozn. z) do jednej z g skupín pomocou kasifikátoru skonštruovaného z matice X a G.

3.1.1 Aplikácia v praxi

Klasifikačné algoritmy hrajú významnú úlohu pri spracovaní signálov. Uplatňujú sa pri rozpoznávaní zvuku, hlasu a hudby. Zastúpenie majú aj v medicíne pri diagnostike na základe symptómov či génovej expresii. Podrobne sa týmto smerom zaoberá [8].

Rozlišovanie obrázkov, identifikácia fotografií, písma a symbolov tiež uplatňujú klasifikáciu. Predstavme si situáciu, kde je cieľom rozpoznať obrázok stromu a zaradiť ju k príslušnému druhu, ilustrované na obrázku 3.1. Nech už disponujeme množstvom klasifikovaných vzorkov n zaradených do g skupín. V našom konkrétnom prípade sú stromy vzorkami a skupinami sú druhy stromov (palmy, listnaté a ihličnaté stromy). Počet vzorkov je $n = 12$ a počet skupín $g = 3$. Využitím definície 1 vieme, že každý stĺpec $\mathbf{x}_i$ matice $\mathbf{X}$ reprezentuje jednu vzorku a jeho zložky sú jednotlivé pixely. Počet pixelov záleží na veľkosti obrázku. Nová vzorka (nový obrázok stromu) bude novým vektorom z, ktorú chceme priradiť do jednej z troch existujúcich skupín.

Na každú klasifikáciu sú základné algoritmy individuálne modifikované, musí sa brať v úvahu otočenie, posunutie, oddialenie, priblíženie... Problematikou v aplikáciách sa zaoberá [9].



Obr. 3.1: Vlastná sada vzorkov ilustrácií stromov.

3.2 Proces klasifikácie

Základný proces klasifikácie je možné popísať dvoma krokmi:

1. Každá skupina je nahradená svojim reprezentantom, napr. skupinovým priemerom. To znamená že pre každú skupinu $j \in \{1,\ldots,g\}$ s množinou indexov N_j a počtom indexov $n_j = |N_j|$ spočítame jej priemer

$$\tilde{\mathbf{x}}_j = \frac{1}{n_j} \sum_{i \in N_j} \mathbf{x}_i, \tag{3.1}$$

   Výpočet skupinových priemerov v maticovom zápise vyzerá nasledovne

$$\tilde{X} = XGH^{-1}, \tag{3.2}$$

$$\text{kde } X = \begin{pmatrix} | & & | \\ \tilde{\mathbf{x}}_1 & \dots & \tilde{\mathbf{x}}_n \\ | & & | \end{pmatrix} \in \mathbb{R}^{p\times g},$$

$$H = diag(1_n^T G) = \begin{pmatrix} n_1 & & \\ & \ddots & \\ & & n_g \end{pmatrix} \in \mathbb{R}^{g\times g},$$

$$1_n^T = \underbrace{(1,\dots,1)}_{\text{n}}.$$

2. Nové pozorovanie z je zaradené do skupiny k, príslúchajúcej najbližšiemu reprezentantovi. Výber metriky záleží na čitateľovi, pre rôzne klasifikácie sú vhodné rôzne metriky. Pri použití Euklidovskej vzdielenosti vyzerá predpis

$$k = argmin_{j\in\{1,\dots,g\}} ||\mathbf{z} - \tilde{\mathbf{x}}_j||. \tag{3.3}$$

V maticovom zápise

$$k = argmin||A||_{col}.$$

Pričom

$$A = \mathbf{z}1_g^T - \tilde{X} = \begin{pmatrix} | & & | \\ \mathbf{z} - \tilde{\mathbf{x}}_1 & \dots & \mathbf{z} - \tilde{\mathbf{x}}_n \\ | & & | \end{pmatrix} \in \mathbb{R}^{p\times g}.$$

Značenie $argmin||.||_{col}$ znamená index stĺpca s najmenšou normou.

Tento základný algoritmus je dostatočný pre jednoduché prípady. Jeho nevýhodou je, že neberie v úvahu variabilitu jednotlivých pozorovaní, preto je pri množstve situácií nevhodný. Tento nedostatok je pri rafinovanejších metódach odstránený, ako napríklad u lineárnej diskriminačnej analýzy.

3.2.1 Test správnosti

Navrhnutie algoritmu vyžaduje experimentálne otestovanie výsledného klasifikátoru. Jedným z prístupov je rozdeliť n pozorovaní na množinu testovacích a trénovacích dát nasledovne:

$$X_{train} = \begin{pmatrix} | & & | \\ \mathbf{x}_1 & \dots & \mathbf{x}_m \\ | & & | \end{pmatrix} \in \mathbb{R}^{p\times m},$$

$$X_{test} = \begin{pmatrix} | & & | \\ \mathbf{x}_{m+1} & \dots & \mathbf{x}_n \\ | & & | \end{pmatrix} \in \mathbb{R}^{p\times(n-m)}.$$

Problém klasifikácie riešime s dátami $X = X_{train}$. Príslušnosť skupiny pre každé "nové" pozorovanie $z = x_i$, kde $i \in \{m+1,\dots,n\}$ poznáme. Preto vieme rozhodnúť, či klasifikácia prebehla správne. Úspešnosť klasifikátora súdime podľa percentuálnej správnosti zaradenia do skupiny. Tento postup je obľúbený vďaka svojej rýchlosti. Vyhodnocuje sa len $n-m$ klasifikácií.

3.3 Klasifikácia pomocou klasickej LDA

Lineárna diskriminačná analýza (LDA) hľadá vektory, ktoré majú v danom priestore najvyšší diskriminant medzi skupinami (radšej než vektory, ktoré najlepšie vytihujú dáta). Predpokladáme $p < n$, t.j. počet pozorovaní je väčší než počet hodnôt pozorovania. Hlavným nástrojom lineárnej diskriminačnej analýzy (LDA) je variančná matica, preto zavedieme značenie

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{j=1}^{n} \mathbf{x}_j \in \mathbb{R}^p,$$
$$\bar{\mathbf{x}}_{i,.} = \frac{1}{n}\sum_{j=1}^{n} x_{i,j} \in \mathbb{R}.$$

Maticový zápis je

$$\bar{\mathbf{x}} = \frac{1}{n}\mathbf{X}\mathbf{1}_n, \text{ kde } \mathbf{1}_n = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^{n\times 1}.$$

Výberovú kovarianciu i-tej a j-tej premennej výberovej matice $X \in \mathbb{R}^{p\times n}$ nazveme

$$cov(i,j) = \frac{1}{n-1}\sum_{k=1}^{k} (x_{i,k} - \bar{x}_{i,.})(x_{j,k} - \bar{x}_{j,.}).$$

Pre rozptyl platí

$$var(i) = cov(i,i).$$

Variančnú maticu S tvoria jednotlivé výberové kovariancie

$$\mathbf{S} = (s_{i,j})_{i,j=1}^{p} = (cov(i,j))_{i,j=1}^{p}.$$

Maticový zápis je

$$\mathbf{S} = \frac{1}{n-1}(\mathbf{X} - \bar{x}\mathbf{1}_n^T)(\mathbf{X}^T - \mathbf{1}_n\bar{x}^T) \in \mathbb{R}^{p\times p}.$$

Keďže S je súčinom dvoch vzájomne transponovaných matíc, je táto matica vždy symetrická a pozitívne semidefinitná. V klasickej LDA je pridanie nového pozorovania z do skupiny k vyjadrené v (3.3). Zakomponovaním vzájomného vzťahu premenných a nahradením Euklidovskej vzdialenosti *Mahalanobisovou vzdialenosťou* vznikne nový predpis

$$k = argmin_{j\in\{1,\ldots,g\}}(\mathbf{z} - \bar{\mathbf{x}}_j)^T S^{-1}(\mathbf{z} - \bar{\mathbf{x}}_j) \tag{3.4}$$

Poznámka 9 (Mahalanobisova vzdialenosť)**.** *Na rozdiel od iných metrík zahŕňa do vzťahu pre vzdialenosť aj väzby medzi bodmi vyjadrené variančnou maticou S. Ide vlastne o vzdialenosť bodov v priestore, ktorého osi nemusia byť ortogonálne. Metrika je definovanú nasledovne*

$$d(x,y) = \sqrt{(x-y)^T S^{-1}(x-y)}.$$

Hľadanie inverznej matice je často analyticky komplikované a numericky nestabilné. Do takej situácie sa môžme dostať aj pri výpočte S^{-1}. Nemusíme použiť explicitný výpočet, pretože variančná matica je symetrická a pozitívne (semi)definitná. Môžme tak pracovať so spektrálnym rozkladom

$$S = QDQ^T,$$

kde D je diagonálna matica s vlastnými číslami na diagonále a Q je ortogonálna matica s príslušnými vlastnými vektormi vo stĺpcoch. Aj vďaka predpokladu o hodnoti S, $\text{rank}(S) = p$, má zmysel výraz $D^{-1/2}$. Vďaka tomu počítame

$$\begin{aligned}(\mathbf{z}-\bar{\mathbf{x}}_j)^T S^{-1}(\mathbf{z}-\bar{\mathbf{x}}_j) &= (\mathbf{z}-\bar{x}_j)^T QD^{-1/2}D^{-1/2}Q^T(\mathbf{z}-\bar{\mathbf{x}}_j) \\ &= ||D^{-1/2}Q^T(\mathbf{z}-\bar{\mathbf{x}}_j)||\end{aligned} \tag{3.5}$$

a ekvivalentne nahradíme klasifikačný výpočet (3.4)

$$k = argmin_{j\in\{1,\dots,g\}}||D^{-1/2}Q^T(\mathbf{z}-\bar{\mathbf{x}}_j)||. \tag{3.6}$$

V niektorých prípadoch je výhodnejšie nahradiť spektrálny rozklad Choleského rozkladom. Výpočetná náročnosť spektrálneho rozkladu našej matice S rádovo činia p^3 operácií a Choleského $p^3/3$, podľa [14].

Poznámka 10 (Choleského rozklad)**.** *Veta o Choleskom rozklade hovorí, že ak je matica $A \in \mathbb{R}^{n\times n}$ symetriká a pozitívne definitná, potom existuje dolná trojuholníková matica $G \in \mathbb{R}^{n\times n}$, jednoznačne určená. Pričom má G kladné prvky na diagonále a platí $A = GG^T$. Dôkaz vety je v [14], sekcia* 4.2.

Pre stabilitu choleského rozladu je nutným predpokladom pozitívna definitnosť variančnej matice, čo nemusí byť vždy splnené. Preto v základnej implementácii LDA, načrtnutej v **Algoritme 1**, je použitý spektrálny rozklad.

Algoritmus 1: Klasická LDA

Vstup : výberová matica X
nové pozorovanie z
matica kódovania skupín G

Výstup: priradená skupina k

1: $H = \text{diag}(\mathbb{1}_n^T G)$
2: $\tilde{X} = XGH^{-1}$
3: $A = z\mathbb{1}_g^T - \tilde{X}$
4: $\bar{x} = X\mathbb{1}_n/n$
5: $S = (X - \bar{x}\mathbb{1}_n)(X^T - \mathbb{1}_n^T\bar{x}^T)$
6: $S = QDQ^T$
7: $k = argmin||D^{-1/2}Q^TA||_{col}$

3.4 Vysokodimenzionálna klasifikácia

Doteraz sme pracovali s predpokladom, že máme viac pozorovaní, než premenných, tj. $p < n$. Ďalším redpokladom bola plná hodnosť variančnej matice $S \in \mathbb{R}^{p\times p}$, $\text{rank}(S) = p$. V množstve aplikácií, napríklad v medocínskom prostredí, narazíme však na problém získania nového pozorovania, pričom je prítomných mnoho premenných. Potom prichádzame do situácie $p > n$, alebo dokonca $p >> n$ a preto $p > \text{rank}(S)$. Klasifikácia nového pozorovania podľa klasickej LDA (3.4) pracuje s inverziou variančnej matice, čo teraz nedáva zmysel. Ukážeme si preto metodu, ktorá pracuje s pseudoinverziou.

3.4.1 Pseudoinverzné matice

Sú známe rôzne spôsoby ako invertovať singulárnu maticu. Nasledujúci text je čerpaný z [1]. Máme maticu $A_{n\times m}$. Ak je $A^{-}_{n\times m}$ taká matica, že platí

$$AA^{-}A = A, \tag{3.7}$$

potom nazývame A^{-}*pseudoinverznou maticou* k matici A. Pseudoinverzná matica existuje vždy, ale nie je určená jednoznačne. V aplikácii sa počíta s výrazmi, kde na konkrétnom vyjadrení pseudoinverzie nezávisí. Už z výrazu (3.7) vidieť, že sa súčin $AA^{-}A$ vždy bude rovnať A bez ohľadu na konkrétnej voľbe A^{-}.

Moore a Penrose zaviedli pseudoinverziu A^{+} k matici A, pričom pre A^{+} platí

1. $AA^{+}A = 1$,
2. $A^{+}AA^{+} = A^{+}$,
3. $(AA^{+})^{T} = AA^{+}$,
4. $(A^{+}A)^{T} = A^{+}A$.

Táto inverzia existuje vždy a je jediná. V numerickej matematike má bohaté využitie, mnoho matematických softwarom s ňou tiež pracuje.

Mooreova-Penrosovu pseudonverziu môžeme spočítať s použitím singulárneho rozkladu nasledovne

$$A = PDQ^{T}, \quad D = \begin{bmatrix} \Delta & (0) \\ (0) & (0) \end{bmatrix},$$

kde Δ je diagonálna matica s kladnými singulárminy číslami na diagonále. O tomto vyjadrení sa dočítame aj v [16] v kapitole 5. Potom je matica D^{+} v tvare

$$D^{+} = \begin{bmatrix} \Delta^{-1} & (0) \\ (0) & (0) \end{bmatrix} \tag{3.8}$$

a A^{+} má tvar

$$A^{+} = QD^{+}P^{T}. \tag{3.9}$$

Pre symetrickú maticu A môžeme využiť spektrálny rozklad

$$A = QDQ^{T}, \quad D = \begin{bmatrix} D_r & (0) \\ (0) & (0) \end{bmatrix}, \tag{3.10}$$

kde D_r je diagonálna matica tvorená kladnými vlastnými číslami matice A. Potom pseudoinverziu počítame nasledovne

$$A^+ = Q \begin{bmatrix} D_r & (0) \\ (0) & (0) \end{bmatrix} Q^T. \tag{3.11}$$

Taktiež môžme použiť redukovaný tvar spektrálneho rozkladu

$$A = Q_r D_r Q_r^T, \tag{3.12}$$

kde Q_r je tvorená prvými r stĺpcami matice Q. Pri práci s redukovaným tvarom má pseudoinverzia tvar

$$A^+ = Q_r D_r^{-1} Q_r^T. \tag{3.13}$$

3.4.2 Klasifikácia pomocou LDA s pseudoinverziou

V lineárnej diskriminačnej analýze pracujeme s variančnou maticou S, ktorá je reálna a symetrická. Vo vysokodynemzionálnych prípadoch nastáva problém s inverziou. Preto môžeme odvodiť Moore-Penroseova pseudoinverziu matice S pomocou spektrálneho rozkladu.

Nech $S \in \mathbb{R}^{p\times p}$ a $\text{rank}(S) = r$. Potom analogicky ku (3.12) nazývame výraz $S = Q_r D_r Q_r^T$ redukovaným spektrálnym rozkladom. M-P pseudoinverzia matice S je S^+, definovaná podľa (3.13) ako

$$S^+ = Q_r D_r^{-1} Q_r^T \in \mathbb{R}^{r\times r}.$$

Vďaka vyjadreniu S^+ aktualizujeme klasifikačný výpočet (3.3) v LDA algoritme novým predpisom

$$k = argmin_{j\in\{1,\dots,g\}} ||D_r^{-1/2} Q_r^T (z - \bar{x}_j)||. \tag{3.14}$$

Ak používame štandardnú metódu výpočtu spektrálneho rozkladu (zabudovanú v matematických softwaroch ako QR metóda), činia výpočetné náklady rádovo p^3 operácií, rovnako ako to bolo pre klasickú LDA v **Algoritme 1**. V prípade $p > n$ alebo dokonca $p >> n$, potom sú náklady dosť nepriaznivé. Pre efektívnejšie počítanie Q_r a D_r zformulujeme tvrdenie.

Veta 4. *Nech $Y \in \mathbb{R}^{p\times n}$, kde $p > n$. Ďalej nech $\tilde{Q}_r D_r \tilde{Q}_r^T$ je redukovaným spektrálnym rozkladom matice $Y^T Y \in \mathbb{R}^{n\times n}$. Ak označíme $Q_r = Y\tilde{Q}_r D_r^{-1/2}$, potom platí, že $Q_r D_r Q_r^T$ je redukovaným rozkladom matice $YY^T \in \mathbb{R}^{p\times p}$.*

Dôkaz. Z vety plynie, že

$$Q_r^T Q_r = D_r^{-1/2} \tilde{Q}_r^T Y^T Y \tilde{Q}_r D_r^{-1/2} = D_r^{-1/2} \tilde{Q}_r^T \left(\tilde{Q}_r D_r \tilde{Q}_r^T\right) \tilde{Q}_r D_r^{-1/2} = I.$$

Zároveň z predpokladu vety je zjavné a vďaka tomu počítame nasledovne

$$\begin{aligned} Y^T Y \tilde{Q}_r &= \tilde{Q}_r D_r \\ YY^T Y \tilde{Q}_r &= Y \tilde{Q}_r D_r \\ YY^T Y Q_r &= Q_r D_r. \end{aligned}$$

Záverom je, že Q_r je ortogonálna a $Q_r D_r Q_r^T$ je redukovaným rozkladom YY^T. □

Algoritmus 2: LDA s pseudoinverziou

Vstup : výberová matica X
nové pozorovanie z
matica kódovania skupín G

Výstup: priradená skupina k

1: $H = \text{diag}(\mathbb{1}_n^T G)$
2: $\tilde{X} = XGH^{-1}$
3: $A = z\mathbb{1}_g^T - \tilde{X}$
4: $\bar{x} = X\mathbb{1}_n/n$
5: $Y = (X - \bar{x}\mathbb{1}_n)$
6: $Y^TY = Q_r D_r Q_r^T$
7: $k = argmin||D_r^{-1}Q_r^T A||_{col}$

Výpočetné náklady sa znížia, spektrálneho rozkladu YY^T vyžaduje iba n^3 operácií oproti pôvodným p^3. Pomocou vety (4) dostaneme **Algoritmus 2**.

Poznámka 11. *Použijeme značenie z* 5. *bodu v algoritme,* $Y = (X - \bar{x}\mathbb{1}_n)$. *Treba si uvedomiť, že variančná matica* $S = Y^TY$ *nemá nikdy plnú hodnosť. Platí*

$$Y^TY\mathbb{1}_n = Y^T(X\mathbb{1}_n - n\bar{x}) = Y^T\mathbf{0}_p = \mathbf{0}_n.$$

Preto je dôležité, aby bol spektrálny rozklad redukovaný.

Zdá sa, že aplikácia pseudoinverzie je postačujúca, no v prípade $p >> n$ je otázne, či nami spočítané vzdialenosti odpovedajú tým Mahalanobisovým. Vo vysokodimenzionálnej štatistike sa preto dáva prednosť metódam, ktoré nahradzujú S jej blízkou regulárnou maticou S^*. Tieto metódy sa súhrnne nazývajú regularizačné a sú bližšie popísané v [17] v sekcii 5.5.

3.5 LDA verzus PCA

Obe metódy, lineárna diskriminačná analýza (LDA) a analýza hlavných komponent (PCA), sú lineárnymi transformačnými technikami, ktoré sa bežne využívajú pri redukcii dimenzie. PCA môže byť označené ako "neriadený" algoritmus. Ignoruje rozdelenie dát do skupín, keďže jeho cieľom je nájsť smery súradnicových os, známe ako hlavné komponenty, ktoré maximaluzujú rozptyl dát. Na druhej strane, LDA je "riadený" algoritmus. Hľadá vektory, označené ako lineárne diskriminanty, tak aby reprezentovali súradnicovú os maximalizujúcu rozptyl medzi jednotlivými skupinami a súčasne minimalizujúcu rozptyl v jednotlivých skupinách.

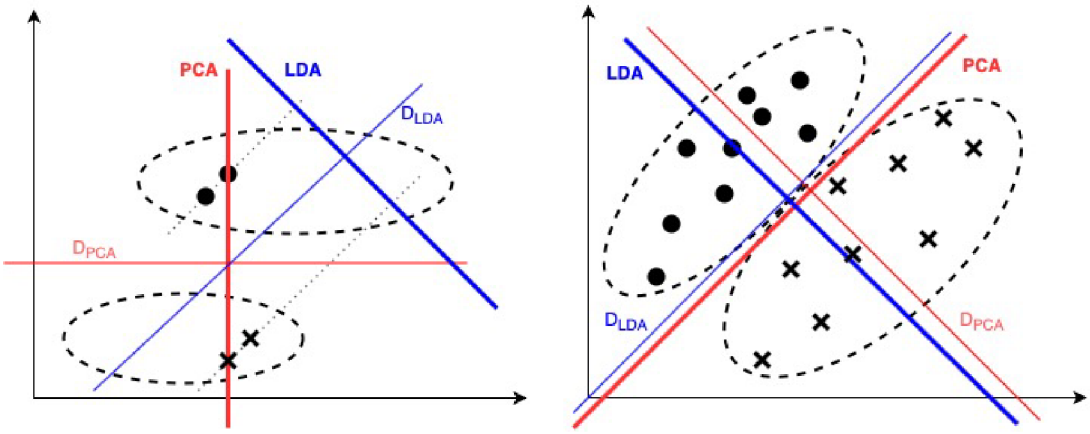
Môže sa zdať, že je LDA nadradené nad PCA v mnohoskupinovej klasifikácii, kde sú jednotlivé zaradenia do skupín známe. No nemusí to tak vždy byť. Podľa článku [18], ktorý skúmal presnosť klasifikácie pri rozpoznávaní obrazu, mala PCA lepšie výsledky oproti LDA v prípade, že bol počet pozorovaní v skupinách nízky.

Ilustratívne znázornenie metód v situácii s malým i veľkým dátovým súborom je na obrázku 3.2. Vidíme dve skupiny dát, reprezentované guličkou a krížikom, a pre každú

predpokladáme normálne rozdelenie, znázornené prerušovanou krivkou. Pri aplikácii metód sa berú v úvahu všetky dáta dokopy. PCA hľadá vektory, ktoré maximalizujú rozptyl všetkých dát. Vektor prvej hlavnej komponenty je zobrazený červenou vertikálnou krivkou označenou **PCA**. Na druhej strane LDA hľadá vektory, ktoré maximaluzujú rozptyl medzi skupinami. Vektor prvého lineárneho diskriminantu je označený modrou diagonálnou krivkou s popisom **LDA**. Hranica delenia skupín je určená metódou najbližšieho suseda a je značená **D** s indexom prislúchajúcej metódy.

V prípade s malým počtou dát (viz obrázok 3.2 vľavo) vidíme, že hranica delenia pre lineárnu diskrimináciu prechádza cez elipsy korešpondujúce rozdeleniu skupín, čo môže spôsobiť chybnú klasifikáciu nových pozorovaných dát. Preto v tomto prípade bude PCA vykazovať lepšie klasifikačné výsledky. Ak je k dispozícii veľký súbor dát, klasifikačné vlastnosti metód sú vymenené a lepšou voľbou je LDA (viz obrázok 3.2 vpravo). Dôkaz tejto ilustrativnej simulácie je realizovaný na reálnych dátach v článku [18].

V praxi nie je neobvyklé, aby sa používala kombinácia metód: napríklad PCA pre redukciu dimenzie nasledovaná LDA pre rozdelenie do skupín.



Obr. 3.2: Klasifikácia pomocou PCA a LDA v závislosti na veľkosti dátovej sady. Vľavo: malá dátová sada. Vpravo: rozsiahla dátová sada. V oboch prípadoch sú dve skupiny s vlastným normálnym rozdelením. Výsledok klasifikácie podľa PCA (vektor prvej hlavnej komponenty je označený krivkou modrej farby) je lepšia pre príprad s malým dátovým súborom. Klasifikácia podľa LDA (vektor prvého lineárneho diskriminantu je označený krivkou červenej farby) je lepšia pri väčšom dátovom súbore. D_{PCA} a D_{LDA} reprezentujú klasifikačnú hranicu delenia zostrojenú podľa metódy najbližšieho suseda. Zdroj [18].

3.5.1 Aplikácia na reálnych dátach

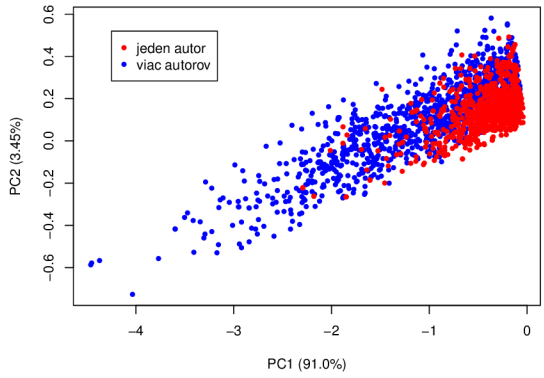
V tejto sekcii je úlohou klasifikácia dát pomocou oboch metód, PCA a LDA. V oboch príkladoch sa spracovávajú reálne dáta, ktoré majú mnoho pozorovaní v každej kategórii. Rozdiel je v množstve kategorizovaných skupín. Analýza dát je realizovaná v programe R s bežne používanými funkciami *prcomp* a *lda*.

Príklad 3 (2 kategórie)**.**

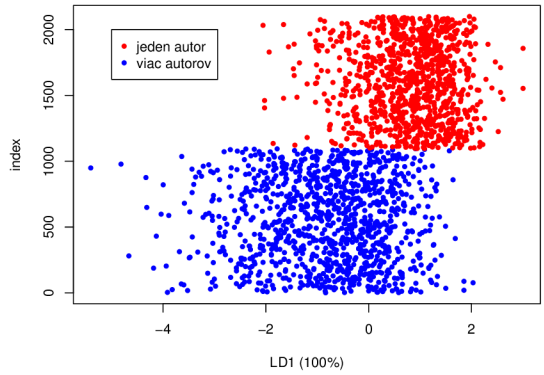
Analyzované dáta pochádzajú zo stránky *kaggle.com* a spracovávajú ručne písaný text. Každé pozorovanie je charakterizované 51 premennými a kategorizované ako písmo jedného autora, alebo viacerých autorov. Takže počet kategórií je 2 a počet všetkých pozorovaní je niečo vyše 2000.

Na obrázku 3.3 sú graficky zobrazené výsledky oboch metód. Vidieť, že pre súbor s takýmito charakteritikami je lepšie použiť LDA. Ak sú prítomné iba 2 skupiny, nehľadiac na dimenziu dátového súboru, existuje iba jeden linárny diskriminant. Podľa tohto vektoru sú skupiny rozdelé jednoznačne a je v ňom obsiahnutá celá variabilita súboru.

Prvé dve hlavné komponenty obsahujú síce vyše 94% variability pôvodného súboru, ale klasifikácia metódou PCA je v tomto prípade nevhodná. Skupiny sú vzhľadom na prvé dve hlavné komponenty premiešané a graficky nejednoznačne rozdelené.



(a)



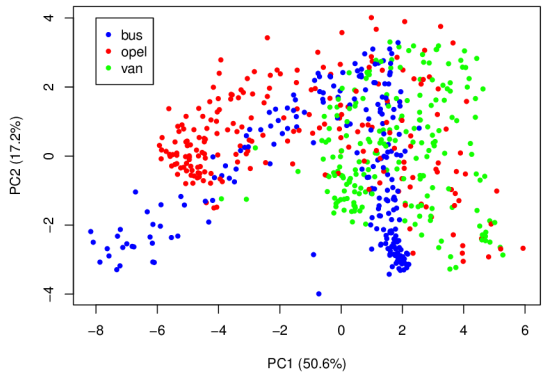
(b)

Obr. 3.3: PCA vs LDA na reálnych dátach spolu s množstvom obsiahnutej variability pôvodného súboru. (a) metóda PCA, (b) metóda LDA. Zdroj [23].

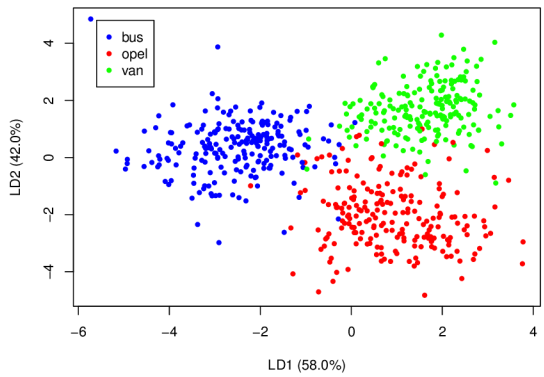
Príklad 4 (3 kategórie)**.**

Analyzované dáta v tomto príklade pochádzajú zo stránky *uci.edu*. Dáta reprezentujú tri typy motorových vozidiel na základe popisu ich siluety. Vozidlo je skúmané z rôznych uhlov a výsledky sú zaznamenané ako jednotlivé atribúty. Počet atribútov je 18 a počet kategórii je 3. Počet všetkých pozorovaní je vyše 600.

Na obrázku 3.3 sú graficky zobrazené výsledky aplikovaných metód. Vidieť, že aj v tomto prípade lepšie klasifikuje skupiny metóda LDA. Prvý a druhý lineárny diskriminant pokrývajú dokopy celú variabilitu súboru. Prvé dve hlavné komponenty obsahujú vyše 67% variability pôvodného súboru. Rovnako ako v predchádzajúcom príklade, sú skupiny vzhľadom na prvé dve hlavné komponenty premiešané a graficky nejednoznačne rozdelené. Preto nie je vhodná voľba klasifikácie metódou PCA.



(a)



(b)

Obr. 3.4: PCA vs LDA na reálnych dátach spolu s množstvom obsiahnutej variability pôvodného súboru. (a) metóda PCA, (b) metóda LDA. Zdroj [25].

Uvedené príklady demonštrujú, že pre veľké množstvo dát je lepšia klasifikácia pomocou lineárnej diskriminačnej analýzy. Situácia s málo pozorovaniami je detailne vypracovaná v [18].

Záver

Cieľom práce bolo priblíženie používania matíc v štatistike. Ukázali sa hlavne široké možnosti využitia variančných matíc a singulárneho rozkladu. Okrem algebraickej časti o rozklade matíc, sa práca venuje dvom odvetviam štatistiky – regresnej a klasifikačnej analýze.

Každá kapitola obsahuje teoretický úvod, na pochopenie preblematiky, ktorý je nasledovaný praktickou demonštráciou popísaných metód na reálnych alebo simulovaných dátach. Pri práci s reálnymi dátami sa vyskytujú viaceré problémy, napríklad chyby merania. V každej aplikácii sme brali ohľad na používaný dátový súbor a jeho vlastnosti. Preto sme museli napríklad zaviesť aj pojem pseudoinverzie.

Pomocou singulárneho rozkladu matíc a hľadaniu vlastných čísel sme vysvetlili metódu analýzy hlavných komponent. Využitie tejto metódy sme demonštrovali pri zhlukovej analýze a redukcii dimenzie. Prvá hlavná komponenta prislúcha najväčšiemu vlastnému číslu, preto sa poskytuje priestor ďalším špeciálnym metódam na hľadanie vlasných čísel.

Ďalej sme sa zaoberali lineárnou regresnou analýzou a rozdielmi v hľadaní odhadov pomocou metód vychádzajúcich z klasickej metódy najmenších štvorcov. Teoreticky sme priblížili hrebeňovú regresiu a metódu Lasso. Praktická časť porovnávame robustnosť jednotlivých odhadov a rozhodujeme o najlepšej voľbe metódy pri daných charakteristikách súboru. Ako rozšírenie témy môže byť hľadanie riešenia vysoko dimenzionálnej sústave rovníc pomocou iteračných metód.

Na záver práce sme sa venovali aj klasickej klasifikačnej analýze a jej metódam, ako je lineárna disktiminačná analýza. Rozdiel medzi analýzou hlavných komponent a lineárnou diskriminačnú analýzu sme demonštrovali na reálnych dátach. Videli sme odlišné riešenia týchto metód, z čoho sa dali vyvodiť závery o preferovanej voľbe v závislosti na veľkosti dátového súboru.

Práca môže poslúžiť ako inšpirácia pre riešenie iných problémov vo vysoko dienzionálnej štatistike. Uvedené poznatky môžu byť použité pre ďalšie modely mnohorozmernej štatistiky. Je nutné podotknúť, že ku každému problému treba pristupovať zvlášť a treba zvážiť či je daný prístup riešenia konkrétneho problému vhodný. V práci predpokladáme iba výskyt hustých matíc. Pri práci s riedkymi maticami by sa využívali vlastnosti tak, ako to je napríklad v [3].

Seznam použité literatury

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

[2] J. Kalina and J. Duintjer Tebbens. *Metody pro redukci dimenze v mnoho- rozměrné statistice a jejich výpočet.* Informačný bulletin Českej štatistickšj spoločnosti, 25:13–29, 2014.

[3] J. Duintjer Tebbens and P. Schlesinger. *Improving implementation of linear discriminant analysis for the high dimension/small sample size problem.* Comput. Statist. Data Anal., 52(1):423–437, 2007.

[4] A. C. Rencher. *Methods of Multivariate Analysis*. Second Edition. John Wiley & Sons, Inc., publication, Brigham Young University, 2001.

[5] J. Kogan. *Introduction to Clustering Large and High-Dimensional Data*. Cambridge University Press, Cambridge 2007. With a foreword by Michael W. Berry.

[6] J. Duintjer Tebbens and P. Schlesinger. Improving implementation of linear discriminant analysis for the high dimension/small sample size problem *Computation Statistical Data Analysis*, 52(1):423-437, 2007.

[7] M. Pourahmadi. *High-Dimensional Covariance Estimation*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, 2013.

[8] R. Tibshirani, T. Hastie, B. Narasimhan, and G.Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Nat. Acad. Sci. USA*, 99:6567-6572, 2002.

[9] Eldén, L (2007). *Matrix Methods in Data Mining and Pattern Recognition, Fundamentals of Algorithms 4*. SIAM.

[10] Efron, B., Hastie, T., Johnstone, I. a Tibshirani, R. (2004). Least angle regression (with discussion). *Annals of Statistics*, 32, 407–499.

[11] Galton, F. (1886a). *Family likeness in stature. Proceedings of the Royal Society of London*, **40**, 42-63.

[12] Galton, F. (1886b). *Regression towards mediocrity in hereditary stature. Journal of the Anthropological Institute*, **15**, 246–263.

[13] T. Hastie, R. Tibshirani, J. Friedman. *The Elements of Statistical Learning - - Data Mining, Inference, and Prediction*. Springer Series in Statistics, 2017.

[14] G. H. Golub and Ch. F. Van Loan. *Matrix Computations.* Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Bal- timore, MD, fourth edition, 2013.

[15] Golub, G. H., Heath, M. a Wahba, G. (1979). *Generalized Cross–Validation as a Method for Choosing a Good Ridge Parameter.* Technometrics, 21(2), 215–223.

[16] J.R.Schott, *Matrix analysis for statistics*, ISBN 0-471-15409-1.

[17] Geoffrey J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*, ISBN 9780471691150.

[18] A. M. Martinez, A. C. Kak, *PCA versus LDA*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 33., No.2., pp 228-233, 2001.

[19] R. Tibshirani. *Regression Shrinkage and Selection via the Lasso.* Journal of the Royal Statistical Society, Series B, 58, 267–288.

[20] Draper, N. R. a Smith, H. (1998). *Applied Regression Analysis.* Wiley-Interscience, New York. ISBN 0-471-17082-8.

[21] Hoerl, A. E. a Kennard, R. W. (1970). *Ridge Regression: Biased Estimation for Nonorthogonal Problems and Applications to Nonorthogonal Problems.* Technometrics, 12.

[22] L. Fahrmeir, T. Kneib, S. Lang S, B. Marx, *Regression- Models, Methods and Applications.* Springer. ISBN 978-3-642-34332-2.

[23] Kaggle, verejné dátové súbory. *https://www.kaggle.com/anupamwadhwa/handwriting-verification*.

[24] Obrázok zebry. *https://goo.gl/images/GN68Bq*.

[25] UCI, Machine Learning Repository. Verejné dátové súbory. *http://archive.ics.uci.edu/ml/machine-learning-databases/statlog/vehicle/*