UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

Robustné metódy vo faktorovej analýze

DIPLOMOVÁ PRÁCA

Bratislava 2013 Bc. Zuzana Kuižová

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY
KATEDRA APLIKOVANEJ MATEMATIKY A ŠTATISTIKY

ROBUSTNÉ METÓDY VO FAKTOROVEJ ANALÝZE

DIPLOMOVÁ PRÁCA

Bc. Zuzana Kuižová

Vedúci diplomevej práce: Mgr. Ján Somorčík, PhD.
Študijný odbor: 1114 aplikovaná matematika
Študijný program: Ekonomická a finančná matematika

BRATISLAVA 2013

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Zuzana Kuižová
Študijný program: ekonomická a finančná matematika (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: 9.1.9. aplikovaná matematika
Typ záverečnej práce: diplomová
Jazyk záverečnej práce: slovenský

Názov: Robustné metódy vo faktorovej analýze.

Cieľ: Faktorová analýza patrí k pokročilejším oblastiam štatistiky. Úlohou študenta bude zoznámiť sa s menej tradičnými postupmi, ktoré sa tam v poslednej dobe používajú, a porovnať ich správanie (najmä pomocou simuláci).

Vedúci: Mgr. Ján Somorčík, PhD.
Katedra: FMFI.KAMŠ - Katedra aplikovanej matematiky a štatistiky

Dátum zadania: 25.01.2012

Dátum schválenia: 26.01.2012 prof. RNDr. Daniel Ševčovič, CSc.
garant študijného programu

.....
študent

.....
vedúci práce

Čestné prehlásenie

Vyhlasujem, že som diplomovú prácu vypracovala samostatne s použitím teoretických vedomostí a uvedenej odbornej literatúry.

Bratislava 22.4.2013 .....

Vlastnoručný podpis

Poďakovanie

Rada by som poďakovala vedúcemu práce Mgr. Jánovi Somorčíkovi za odborné rady, ochotu a čas pri vypracovávaní tejto práce. Chcela by som sa poďakovať aj mojej rodine a priateľom za ich podporu počas doterajšieho štúdia.

Abstrakt

Táto diplomová práca sa zaoberá faktorovou analýzou, ktorá sa používa na zjednodušenie štatistických analýz a troma robustnými metódami na odhad kovariančnej matice. Ak je pôvodný počet premenných vysoký a existuje medzi nimi lineárna závislosť, tak pomocou faktorovej analýzy môžeme získať menší počet nových nezávislých premenných. Avšak ak pôvodné dáta obsahujú outlierov alebo nespĺňajú predpoklady modelu faktorovej analýzy, tak odhad faktorového modelu nemusí byť správny. Jeho najdôležitejším krokom je odhad kovariančnej matice, preto ju v tejto práci budeme odhadovať pomocou robustných metód a vzniknuté odhady faktorového modelu navzájom porovnávať. Našim cieľom je ukázať, že existujú metódy, ktoré sú odolné voči vplyvu outlierov.

Kľúčové slová: Faktorová analýza, Kovariančná matica, Robustné metódy

Abstract

This Master´s thesis deals with the Factor Analysis, which is useful for simplification of statistical analyses and it also deals with three robust methods to estimate covariance matrices. As long as the original number of variables is high and there is a linear dependance it is possible to obtain smaller number of new independent variables. However, if the data contain outliers or they do not satisfy the assumptions of factor model then the estimation of the factor model may not be correct. The most important step is to estimate the covariance matrix therefore we use robust methods to do it. Subsequently we compare these robust methods to each other. Our aim is to point out the existence of methods that can resist the effect of outliers.

Keywords: Factor analysis, Covariance matrix, Robust methods

Obsah

Úvod

Faktorová analýza je viacrozmerná metóda, ktorej cieľom je aproximovať pozorovania lineárnou kombináciou menšieho počtu nepozorovateľných premenných. Snaží sa odhaliť skryté vzťahy medzi pôvodnými premennými a jej výsledkom by mal byť menší počet nezávislých premenných, ktoré budú zachovávať podstatnú časť informácie. Výsledky faktorového modelu závisia od kvality odhadu kovariančnej, resp. korelačnej matice pôvodných premenných. Prítomnosť outlierov v dátach môže ľubovoľne zmeniť hodnoty odhadov strednej hodnoty a kovariancie, ktoré sú potrebné pri mnohých štatistických analýzach. Odhalenie outlierov vo viacrozmerných dátových súboroch je považovaný za dôležitý a ťažký problém v inžinierskych vedných odboroch, ako aj v ekonómii a financiách. Aj pri získaní viacnásobných meraní je vždy možný vznik zhlukov outlierov. Častokrát reálne dáta nespĺňajú ani predpoklady o rozdelení. Preto sa budeme v tejto práci snažiť nájsť robustné metódy na odhad kovariančnej matice, ktoré budú odolné voči outlierom i voči porušeniu predpokladov.

V prvej kapitole si opíšeme faktorový model, jeho základné princípy a predpoklady. Vysvetlíme si kľúčové pojmy v tejto oblasti štatistiky, ktoré budeme používať v celej práci. Samotné výpočty faktorovej analýzy pozostávajú z viacerých krokov. Po získaní kovariančnej matice nasleduje odhad modelu, ktorý môžeme urobiť viacerými metódami. Nakoniec môžeme ešte výsledky transformovať rotáciou pre lepšiu interpretovateľnosť.

Robustným odhadom kovariančnej matice sa budeme venovať v druhej kapitole. Popíšeme si tri metódy, z ktorých prvá je najznámejšia a nebudeme ju ani sami programovať. Jej základným princípom je vybrať podmnožinu dát, z ktorej sa bude počítať odhad. Vysvetlíme si jej základnú myšlienku a ideu algoritmu. Ďalšie dve metódy a ich algoritmy si popíšeme detailnejšie pre účely programovania. Posledná tretia metóda sa tiež snaží nájsť podmnožinu dát očistenú od outlierov. Výpočtovo je najzložitejšia, preto jej venujeme najviac priestoru.

V tretej kapitole budeme testovať kvalitu odhadu faktorového modelu, ktorého východiskom budú robustné odhady a klasický odhad kovariančnej matice. Tieto testy budeme robiť na simulovaných dátach, v ktorých budeme meniť spôsob generova-

nia outlierov a porušení predpokladov faktorového modelu. Výsledkom budú krivky znázorňujúce veľkosť strednej kvadratickej chyby odhadovaných výsledkov v porovnaní so skutočnými hodnotami. Faktorový model budeme navyše pre každý odhad kovariančnej matice odhadovať dvoma metódami, z ktorých jedna má predpoklad normality dát. Očakávame, že odhad rozkladu kovariančnej matice pomocou robustných metód sa bude viac približovať skutočnosti ako odhad faktorového modelu, ktorého vychodiskom je klasický odhad kovariančnej matice. Zaujíma nás ale aj ich vzájomné porovnanie, ako aj porovnanie metód odhadu faktoroveého modelu.

1 Faktorová analýza

V tejto kapitole je spracovaná teória faktorovej analýzy na základe knihy [1] v zozname literatúry.

1.1 Základná charakteristika

Faktorová analýza patrí do skupiny viacrozmerných štatistických metód. Začala sa rozvíjať začiatkom 20. storočia, a to v psychológii zásluhou Karla Pearsona, Charlesa Spearmana a iných. Základný význam faktorovej analýzy je opísať, ak je to možné, kovariancie medzi premennými pomocou menšieho počtu nepozorovateľných náhodných veličín, ktoré sa nazývajú faktory. Predpokladajme teda, že premenné môžeme rozdeliť do skupín na základe ich korelácii. V rámci každej skupiny sú premenné medzi sebou vysoko korelované, ale medzi skupinami sú relatívne malé korelácie. Každá skupina potom predstavuje už spomínaný faktor. Na určenie týchto vzájomných vzťahov je podstatným krokom odhad kovariančnej, resp. korelačnej matice. Jej prvoradou otázkou je, či sú dáta konzistentné s predpísanou štruktúrou. Ak to skúmaný problém vyžaduje, nové premenné sa snažíme čo najlepšie interpretovať. Faktorová analýza môže byť považovaná za rozšírenie analýzy hlavných komponentov.

1.2 Ortogonálny faktorový model

Máme daný p-rozmerný náhodný pozorovateľný vektor $\boldsymbol{X}= (X_1, X_2, ..., X_p)'$ so strednou hodnotou $\boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_p)'$ a kovariančnou maticou $\boldsymbol{\Sigma}$. Faktorový model predpokladá, že premenné vektora $\boldsymbol{X}$ sú navzájom lineárne závislé. Výsledkom faktorovej analýzy je k nepozorovateľných náhodných premenných, spoločných faktorov $F_1, F_2, ..., F_k$, a p dodatočných zdrojov variancie $\varepsilon_1, \varepsilon_2, ..., \varepsilon_p$, ktoré nazývame chyby alebo špecifické faktory. Faktorový model vyzerá takto:

$$\begin{aligned} X_1 - \mu_1 &= l_{11}F_1 + l_{12}F_2 + ... + l_{1k}F_k + \varepsilon_1 \\ X_2 - \mu_2 &= l_{21}F_1 + l_{22}F_2 + ... + l_{2k}F_k + \varepsilon_2 \\ &\vdots \\ X_p - \mu_p &= l_{p1}F_1 + l_{p2}F_2 + ... + l_{pk}F_k + \varepsilon_p \end{aligned} \tag{1.2.1}$$

Môžeme ho zapísať aj v maticovom tvare:

$$\boldsymbol{X} - \boldsymbol{\mu} = \boldsymbol{LF} + \boldsymbol{\varepsilon} \tag{1.2.2}$$

Koeficienty l_{ij} sú nenáhodné a nazývame ich *náklady* i-tej premennej na j-tom faktore. Matica $\boldsymbol{L}$ sa potom nazýva matica faktorových nákladov. Poznamenajme, že i-ty špecifický faktor je spojený len s i-tou premennou. Faktorový model sa od viacrozmerného regresného modelu líši tým, že má $p + k$ nepozorovateľných náhodných premenných $(F_1, F_2, ..., F_k,\ \varepsilon_1, \varepsilon_2, ..., \varepsilon_p)$. Keďže je ich viac ako pozorovateľných premenných, musíme prijať dodatočné predpoklady, aby sme mohli preskúmať model. Predpokladáme, že

$$\begin{aligned} E(\boldsymbol{F}) &= \boldsymbol{0}\ \ (k \times 1) \\ Cov(\boldsymbol{F}) &= E[\boldsymbol{FF'}] = \boldsymbol{I}\ \ (k \times k) \\ E(\boldsymbol{\varepsilon}) &= \boldsymbol{0}\ \ (p \times 1) \\ Cov(\boldsymbol{\varepsilon}) &= E[\boldsymbol{\varepsilon\varepsilon'}] = \boldsymbol{\Psi} = \begin{bmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \cdots & \psi_p \end{bmatrix}\ \ (p \times p) \end{aligned} \tag{1.2.3}$$

$$Cov(\boldsymbol{\varepsilon}, \boldsymbol{F}) = E(\boldsymbol{\varepsilon F'}) = \boldsymbol{0}\ \ (p \times k)\ \ \text{- nekorelovanosť } \boldsymbol{F} \text{ a } \boldsymbol{\varepsilon}$$

Dôležitým predpokladom je aj lineárny vzťah medzi pozorovanými premennými a spoločnými faktormi, aby bola zaručená formulácia modelu.

Na základe faktorového modelu (1.2.2) vieme odvodiť vzťah pre kovariančnú maticu pozorovaných premenných (matica $\mathbf{X}$):

$$\begin{aligned} (\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})' &= (\boldsymbol{LF} + \boldsymbol{\varepsilon})(\boldsymbol{LF} + \boldsymbol{\varepsilon})' \\ &= (\boldsymbol{LF} + \boldsymbol{\varepsilon})((\boldsymbol{LF})' + \boldsymbol{\varepsilon}') \\ &= \boldsymbol{LF}(\boldsymbol{LF})' + \boldsymbol{\varepsilon}(\boldsymbol{LF})' + \boldsymbol{LF\varepsilon}' + \boldsymbol{\varepsilon\varepsilon}' \end{aligned} \tag{1.2.4}$$

Takže na základe predchádzajúceho a vzťahov (1.2.3) platí:

$$\begin{aligned} \boldsymbol{\Sigma} = Cov(\boldsymbol{X}) &= E(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})' \\ &= \boldsymbol{L}E(\boldsymbol{FF'})\boldsymbol{L'} + E(\boldsymbol{\varepsilon F'})\boldsymbol{L'} + \boldsymbol{L}E(\boldsymbol{F\varepsilon'}) + E(\boldsymbol{\varepsilon\varepsilon'}) \\ &= \boldsymbol{LL'} + \boldsymbol{\Psi} \end{aligned} \tag{1.2.5}$$

kde $\boldsymbol{\Sigma}$ je kovariančná matica premenných $X_1, X_2, ..., X_p$.

Podobne si môžeme vyjadriť kovarianciu pôvodných premenných a spoločných faktorov (matice $\boldsymbol{X}$ a $\boldsymbol{F}$):

$$(\boldsymbol{X} - \boldsymbol{\mu})\boldsymbol{F}' = (\boldsymbol{L}\boldsymbol{F} + \boldsymbol{\varepsilon})\boldsymbol{F}' = \boldsymbol{L}\boldsymbol{F}\boldsymbol{F}' + \boldsymbol{\varepsilon}\boldsymbol{F}' \tag{1.2.6}$$

Potom:

$$Cov(\boldsymbol{X}, \boldsymbol{F}) = E(\boldsymbol{X} - \boldsymbol{\mu})\boldsymbol{F}' = \boldsymbol{L}E(\boldsymbol{F}\boldsymbol{F}') + E(\boldsymbol{\varepsilon}\boldsymbol{F}') = \boldsymbol{L} \tag{1.2.7}$$

Vzťah (1.2.5) si prepíšeme v podrobnejšej forme a osobitne vyjadríme kovarianciu každých dvoch pozorovaných premenných:

$$Cov(X_i, X_h) = l_{i1}l_{h1} + \cdots + l_{ik}l_{hk} \qquad i, h = 1, 2, ..., p \tag{1.2.8}$$

a varianciu každej pozorovanej premennej:

$$\begin{aligned} Var(X_i) &= l_{i1}^2 + \cdots + l_{ik}^2 + \psi_i \\ \sigma_{ii} &= \quad h_i^2 \quad + \psi_i \qquad i = 1, 2, ..., p \end{aligned} \tag{1.2.9}$$

kde $Var(X_i) = \sigma_{ii}$, $h_i^2 = l_{i1}^2 + \cdots + l_{ik}^2$ sa nazývajú komunality a ψ_i špecifické variancie alebo unicity. Teda i-ta komunalita predstavuje sumu štvorcov nákladov i-tej premennej na k spoločných faktoroch. Dôležité je, aby komunality boli vyššie ako špecifické variancie pre danú premennú. Pretože nízka komunalita môže znamenať, že model faktorovej analýzy nie je pre danú premennú správny.

Vzťah (1.2.7) pre každú dvojicu pozorovateľnej premennej a spoločného faktora vyzerá takto:

$$Cov(X_i, F_j) = l_{ij} \qquad i = 1, 2, ..., p, \;\; j = 1, 2, ..., k \tag{1.2.10}$$

1.3 Metódy odhadu

Pre každú premennú $X_1, X_2, ...X_p$ je daných n pozorovaní. Základnou otázkou faktorovej analýzy pri daných pozorovaniach je, či faktorový model s daným malým počtom spoločných faktorov reprezentuje dáta primerane. Zaujíma nás, či spoločné faktory zachovávajú informáciu, ktorá je obsiahnutá v pozorovaných premenných. Prvým krokom je odhad kovariančnej matice $\boldsymbol{\Sigma}$, ktorý budeme označovať ako maticu $\boldsymbol{S}$. Ak sú mimodiagonálne prvky tejto matice malé, tak pozorované premenné nie sú príbuzné

a faktorová analýza by nebola užitočná. Ak sa ale kovariančná matica signifikantne líši od diagonálnej matice, môžeme uvažovať o faktorovej analýze. Ďalším krokom je potom odhad faktorových nákladov, čiže matice $\boldsymbol{L}$ a špecifických variancií, matice $\boldsymbol{\Psi}$. Budeme sa zaoberať viacerými metódami odhadu, a to metódou hlavných komponentov, metódou hlavných faktorov a metódou maximálnej vierohodnosti.

1.3.1 Metóda hlavných komponentov

Pomocou spektrálneho rozkladu si vieme maticu $\boldsymbol{\Sigma}$ zapísať v tvare:

$$\begin{aligned}\boldsymbol{\Sigma} &= \lambda_1 e_1 e_1' + \lambda_2 e_2 e_2' + \cdots + \lambda_p e_p e_p' \\ &= \left[\sqrt{\lambda_1}e_1|\sqrt{\lambda_2}e_2|\cdots|\sqrt{\lambda_p}e_p\right]\begin{bmatrix}\sqrt{\lambda_1}e_1' \\ \sqrt{\lambda_2}e_2' \\ \vdots \\ \sqrt{\lambda_p}e_p'\end{bmatrix} = \boldsymbol{L}\boldsymbol{L}'\end{aligned} \tag{1.3.1}$$

kde λ_i sú vlastné čísla kovariančnej matice a e_i jej vlastné vektory. Pričom platí, že $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Táto stanovená štruktúra faktorovej analýzy má toľko spoločných faktorov ako pozorovaných premenných ($k = p$) a špecifické variancie $\psi_i = 0$ pre každé $i = 1, 2, ..., p$. Môžeme teda napísať:

$$\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}' + \boldsymbol{0} = \boldsymbol{L}\boldsymbol{L}' \quad (p \times p) \tag{1.3.2}$$

Tento model faktorovej analýzy je exaktný ale nie je užitočný, pretože preferujeme model s menším počtom spoločných faktorov. Ak je posledných $p-k$ vlastných hodnôt malých, tak môžeme zanedbať príspevok $\lambda_{k+1}e_{k+1}e_{k+1}' + \cdots + \lambda_p e_p e_p'$ k matici $\boldsymbol{\Sigma}$ v (1.3.1). Tým získame aproximáciu:

$$\boldsymbol{\Sigma} \doteq \left[\sqrt{\lambda_1}e_1|\sqrt{\lambda_2}e_2|\cdots|\sqrt{\lambda_k}e_k\right]\begin{bmatrix}\sqrt{\lambda_1}e_1' \\ \sqrt{\lambda_2}e_2' \\ \vdots \\ \sqrt{\lambda_k}e_k'\end{bmatrix} = \boldsymbol{L}\boldsymbol{L}' \quad (p \times k)(k \times p) \tag{1.3.3}$$

Ak do modelu zahrnieme špecifické variancie, tak to budú diagonálne prvky matice $\boldsymbol{\Sigma} - \boldsymbol{L}\boldsymbol{L}'$, kde $\boldsymbol{L}\boldsymbol{L}'$ je definovaná v (1.3.3). Aproximácia matice $\boldsymbol{\Sigma}$ bude mať potom

tvar:

$$\mathbf{\Sigma} \doteq \left[\sqrt{\lambda_1}e_1|\sqrt{\lambda_2}e_2|\cdots|\sqrt{\lambda_k}e_k\right]\begin{bmatrix}\sqrt{\lambda_1}e_1'\\ \sqrt{\lambda_2}e_2'\\ \vdots\\ \sqrt{\lambda_k}e_k'\end{bmatrix}+\begin{bmatrix}\psi_1 & 0 & \cdots & 0\\ 0 & \psi_2 & \cdots & 0\\ \vdots & \vdots & \ldots & \vdots\\ 0 & 0 & \cdots & \psi_p\end{bmatrix} \tag{1.3.4}$$

kde $\psi_i = \sigma_{ii} - \sum_{j=1}^{k} l_{ij}^2$ pre $i = 1, 2, ...p$. V prípade, že premenné nie sú v rovnakých meracích jednotkách, tak je potrebné údaje štandardizovať (od každej premennej odčítame jej strednú hodnotu a vydelíme ju príslušnou odchýlkou):

$$\mathbf{z_h} = \begin{bmatrix}\frac{(x_{h1}-\overline{x}_1)}{\sqrt{s_{11}}}\\ \frac{(x_{h2}-\overline{x}_2)}{\sqrt{s_{22}}}\\ \vdots\\ \frac{(x_{hp}-\overline{x}_p)}{\sqrt{s_{pp}}}\end{bmatrix} \qquad h = 1, 2, ...n \tag{1.3.5}$$

V tomto prípade potom namiesto kovariančnej matice vlastne odhadujeme korelačnú maticu meraní $x_1, x_2, ..., x_n$. Štandardizácia premenných nám umožňuje predísť situácii, že by jedna premenná mala prehnane veľký vplyv na určenie faktorových nákladov.

Riešenie modelu hlavných faktorov spočíva v odhade matice $\boldsymbol{S}$ (kovariančnej matice premenných) a jej spektrálnom rozklade ($\widehat{\lambda}_i$ - vlastné čísla a $\widehat{e}_i$ - vlastné vektory pre $i = 1, 2, ..., p$). Pre odhadnuté vlastné čísla platí, že $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \cdots \geq \widehat{\lambda}_p$. Ak je k počet spoločných faktorov, potom odhad matice faktorových nákladov je daný:

$$\widetilde{\boldsymbol{L}} = \left[\sqrt{\widehat{\lambda}_1}\widehat{e}_1|\sqrt{\widehat{\lambda}_2}\widehat{e}_2|\cdots|\sqrt{\widehat{\lambda}_k}\widehat{e}_k\right] \tag{1.3.6}$$

Špecifické variancie sú diagonálne prvky matice $\boldsymbol{S}-\widetilde{\boldsymbol{L}}\widetilde{\boldsymbol{L}}'$:

$$\widetilde{\psi}_i = s_{ii} - \sum_{j=1}^{k}\widetilde{l}_{ij}^2 \tag{1.3.7}$$

Potom môžeme odhadnúť komunality:

$$\widetilde{h}_i^2 = \widetilde{l}_{i1}^2 + \widetilde{l}_{i2}^2 + \cdots + \widetilde{l}_{ik}^2 \tag{1.3.8}$$

Ak je potrebné premenné štandardizovať, tak vychádzame z odhadnutej korelačnej matice $\boldsymbol{R}$.

Pred odhadom faktorovej analýzy musíme zadať počet spoločných faktorov. Avšak táto metóda neposkytuje kritérium na ich presné určenie. Niekedy je počet spoločných faktorov určený dopredu na základe teórie alebo iných výskumov. Ak tento počet nepoznáme, tak faktorový model odhadneme viackrát s rôznymi počtami a porovnáme výsledky, z ktorých sa snažíme vybrať ten najvhodnejší. Určenie k môže byť založené aj na odhadnutých stredných hodnotách. Uvažujme reziduálnu maticu:

$$\boldsymbol{S} - (\widetilde{\boldsymbol{L}}\widetilde{\boldsymbol{L}}' + \widetilde{\boldsymbol{\Psi}}) \tag{1.3.9}$$

Diagonálne prvky tejto matice sú nulové a ak ostatné prvky sú veľmi malé, tak môžeme považovať k za správny počet spoločných faktorov. Potom teda platí:

$$\text{Suma štvorcov prvkov}(\boldsymbol{S} - (\widetilde{\boldsymbol{L}}\widetilde{\boldsymbol{L}}' + \widetilde{\boldsymbol{\Psi}})) \leq \widehat{\lambda}_{k+1}^2 + \cdots + \widehat{\lambda}_p^2 \tag{1.3.10}$$

Dôležité je, aby spoločné faktory vysvetľovali čo najviac celkového rozptylu. Malo by to byť $70 - 90\%$. Závisí to aj od typu skúmaných premenných. Príspevok prvého spoločného faktora k celkovej variancii $s_{11} + s_{22} + \cdots + s_{pp} = tr(\boldsymbol{S})$ vieme vyjadriť pomocou faktorových nákladov:

$$\widehat{\lambda}_1 = \left(\sqrt{\widehat{\lambda}_1}\widehat{e}_1\right)' \left(\sqrt{\widehat{\lambda}_1}\widehat{e}_1\right) = \widetilde{l}_{11}^2 + \widetilde{l}_{21}^2 + \cdots + \widetilde{l}_{p1}^2 \tag{1.3.11}$$

kde vlastný vektor $\widehat{e}_1$ má jednotkovú dĺžku. Vo všeobecnosti pre podiel celkovej variancie vzhľadom na j-ty faktor platí:

$$\begin{array}{ll} \dfrac{\widehat{\lambda}_j}{s_{11} + s_{22} + \cdots + s_{pp}} & \text{pre faktorovú analýzu matice } \boldsymbol{S} \\ \dfrac{\widehat{\lambda}_j}{p} & \text{pre faktorovú analýzu matice } \boldsymbol{R} \end{array} \tag{1.3.12}$$

Ďalším pravidlom pre určenie počtu spoločných faktorov môže byť počet kladných vlastných hodnôt v prípade faktorovej analýzy kovariančnej matice $\boldsymbol{S}$ alebo počet vlatných hodnôt väčších ako jedna v prípade faktorovej analýzy korelačnej matice $\boldsymbol{R}$.

Tieto kritéria ale nemôžu byť použité bez premyslenia a pochopenia štruktúry modelu. Cieľom je aproximácia pozorovaných premenných malým počtom spoločných faktorov pri čo najlepšej možnej interpratácii.

1.3.2 Metóda hlavných faktorov

Táto metóda odhadu je vlastne modifikovaným prístupom metódy hlavných komponentov. Odvodíme ju z korelačnej matice, hoci túto procedúru môžeme aplikovať aj na kovariančnú maticu. Takže odhadujeme faktorový model:

$$\boldsymbol{\rho} = \boldsymbol{L}\boldsymbol{L}' + \boldsymbol{\Psi} \tag{1.3.13}$$

Komunality tvoria časť diagonálnych prvkov matice $\boldsymbol{\rho}$:

$$\rho_{ii} = 1 = h_i^2 + \psi_i \tag{1.3.14}$$

Predpokladajme, že poznáme odhad ψ_i^* špecifických variancií. Potom nahradením diagonálnych prvkov matice $\boldsymbol{R}$ za $h_i^{*2} = 1 - \psi_i^*$ získame "redukovanú"korelačnú maticu:

$$\boldsymbol{R_r} = \begin{bmatrix} h_1^{*2} & r_{12} & \cdots & r_{1p} \\ r_{12} & h_2^{*2} & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1p} & r_{2p} & \cdots & h_p^{*2} \end{bmatrix} \tag{1.3.15}$$

Aproximácia tejto redukovanej korelačnej matice je určená cez k spoločných faktorov:

$$\boldsymbol{R_r} = \boldsymbol{L_r^*}\boldsymbol{L_r^{*\prime}} \tag{1.3.16}$$

kde $\boldsymbol{L_r^*}= \{l_{ij}^*\}$ sú odhadnuté faktorové náklady.

V tejto metóde hlavných faktorov musíme urobiť nasledujúce odhady:

$$\begin{aligned} \boldsymbol{L_r^*} &= \left[\sqrt{\widehat{\lambda}_1^*}\widehat{e}_1^* | \sqrt{\widehat{\lambda}_2^*}\widehat{e}_2^* | \cdots | \sqrt{\widehat{\lambda}_k^*}\widehat{e}_k^*\right] \\ \psi_i^* &= 1 - \sum_{j=1}^{k} l_{ij}^{*2} \qquad i = 1, 2, ..., p \end{aligned} \tag{1.3.17}$$

kde $\widehat{\lambda}_j^*$ sú vlastné čísla a $\widehat{e}_j^*$ vlastné vektory matice $\boldsymbol{R_r}$ pre $j = 1, 2, ..., k$. Pričom vlastné čísla sú usporiadné zostupne, kde $\widehat{\lambda}_1^*$ je najväčšia hodnota. Komunality môžeme potom znovu odhadnúť:

$$\widetilde{h}_i^{*2} = \sum_{j=1}^{k} l_{ij}^{*2} \qquad i = 1, 2, ..., p \tag{1.3.18}$$

Riešenie metódy hlavných faktorov môžeme získať iteračne, kde odhad komunalít (1.3.18) bude začiatočným odhadom v ďalšej iterácii. Podobne ako v metóde hlavných

komponentov môžeme určiť počet spoločných faktorov na základe odhadnutých vlastných hodnôt $\widehat{\lambda}_1^*, \widehat{\lambda}_2^*, ..., \widehat{\lambda}_p^*$. Druhou možnosťou je určiť k rovné hodnosti redukovanej korelačnej matice. Avšak táto hodnosť nemusí byť vždy určená správne, preto je nutné ďalšie zhodnotenie počtu spoločných faktorov.

Na počiatočný odhad špecifických variancií existuje viacero možností, no najpoužívanejší, ak vychádzame z korelačnej matice, je $\psi_i^* = 1/r^{ii}$, kde r^{ii} je i-ty diagónalny prvok matice $\boldsymbol{R^{-1}}$. Začiatočným odhadom komunalít potom bude:

$$h_i^{*2} = 1 - \psi_i^* = 1 - \frac{1}{r^{ii}} \tag{1.3.19}$$

1.3.3 Metóda maximálnej vierohodnosti

Ak sú spoločné faktory (matica $\boldsymbol{F}$) a špecifické faktory (matica $\boldsymbol{\varepsilon}$) združene normálne rozdelené, tak môžeme faktorový model odhadnúť metódou maximálnej vierohodnosti. Pozorovania $\boldsymbol{X}_j - \boldsymbol{\mu} = \boldsymbol{L}\boldsymbol{F}_j + \boldsymbol{\varepsilon}_j$ sú potom normálne rozdelené. Potom maximalizujeme vierohodnostnú funkciu, ktorá závisí na matici $\boldsymbol{L}$ a $\boldsymbol{\Psi}$ cez známy vzťah $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}' + \boldsymbol{\Psi}$:

$$\begin{aligned} L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= (2\pi)^{-\frac{np}{2}} |\boldsymbol{\Sigma}|^{-\frac{n}{2}} e^{-\left(\frac{1}{2}\right) tr\left[\boldsymbol{\Sigma}^{-1}\left(\sum_{j=1}^{n}(x_j - \overline{x})(x_j - \overline{x})' + n(\overline{x} - \boldsymbol{\mu})(\overline{x} - \boldsymbol{\mu})'\right)\right]} \\ &= (2\pi)^{-\frac{(n-1)p}{2}} |\boldsymbol{\Sigma}|^{-\frac{(n-1)}{2}} e^{-\left(\frac{1}{2}\right) tr\left[\boldsymbol{\Sigma}^{-1}\left(\sum_{j=1}^{n}(x_j - \overline{x})(x_j - \overline{x})'\right)\right]} \\ &\times (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} e^{-\left(\frac{n}{2}\right)(\overline{x} - \boldsymbol{\mu})'\Sigma^{-1}(\overline{x} - \boldsymbol{\mu})} \end{aligned} \tag{1.3.20}$$

Tento model ale ešte nie je dobre definovaný, pretože existuje viacero možností pre určenie matice $\boldsymbol{L}$. Musíme definovať doplňujúcu podmienku jednoznačnosti:

$$\boldsymbol{L'\Psi^{-1}L} = \boldsymbol{\Delta} \tag{1.3.21}$$

kde $\boldsymbol{\Delta}$ je diagonálna matica.

Ak sú merania $x_1, x_2, ..., x_n$ náhodné premenné z normálneho rozdelenia $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, potom metódou maximálnej vierohodnosti získame aj odhad komunalít:

$$\widehat{h}_i^2 = \widehat{l}_{i1}^2 + \widehat{l}_{i2}^2 + \cdots + \widehat{l}_{ik}^2 \qquad i = 1, 2, ..., p \tag{1.3.22}$$

Podiel celkovej variancie vysvetlený j-tym faktorom sa potom rovná:

$$\frac{\widehat{l}_{1j}^2 + \widehat{l}_{2j}^2 + \cdots + \widehat{l}_{pj}^2}{s_{11} + s_{22} + \cdots + s_{pp}} \tag{1.3.23}$$

Ak sú premenné v rôznych meracích jednotkách, tak odhadujeme kovariančnú maticu $\boldsymbol{\rho}$ štandardizovaných premenných $\boldsymbol{Z} = \boldsymbol{V}^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu})$, čo je vlastne korelačná matica centrovaných premenných:

$$\boldsymbol{\rho} = \boldsymbol{V}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{V}^{-1/2} = (\boldsymbol{V}^{-1/2}\boldsymbol{L})(\boldsymbol{V}^{-1/2}\boldsymbol{L})' + \boldsymbol{V}^{-1/2}\boldsymbol{\Psi}\boldsymbol{V}^{-1/2} \tag{1.3.24}$$

Táto korelačná matica má maticu nákladov $\boldsymbol{L_Z}$=$\boldsymbol{V}^{-1/2}\boldsymbol{L}$ a maticu špecifických faktorov $\boldsymbol{\Psi_Z}$=$\boldsymbol{V}^{-1/2}\boldsymbol{\Psi}\boldsymbol{V}^{-1/2}$. Potom z invariantnosti odhadu metódou maximálnej vierohodnosti môžeme odvodiť odhad matice $\boldsymbol{\rho}$:

$$\begin{aligned} \widehat{\boldsymbol{\rho}} &= (\widehat{\boldsymbol{V}}^{-1/2}\widehat{\boldsymbol{L}})(\widehat{\boldsymbol{V}}^{-1/2}\widehat{\boldsymbol{L}})' + \widehat{\boldsymbol{V}}^{-1/2}\widehat{\boldsymbol{\Psi}}\widehat{\boldsymbol{V}}^{-1/2} \\ &= \widehat{\boldsymbol{L}}_{\boldsymbol{Z}}\widehat{\boldsymbol{L}}_{\boldsymbol{Z}}{}' + \widehat{\boldsymbol{\Psi}}_{\boldsymbol{Z}} \end{aligned} \tag{1.3.25}$$

kde $\widehat{\boldsymbol{V}}^{-1/2}$ a $\widehat{\boldsymbol{L}}_{\boldsymbol{Z}}$ sú odhadmi matice $\boldsymbol{V}^{-1/2}$ a $\boldsymbol{L_Z}$ metódou maximálnej vierohodnosti. V tomto prípade faktorovej analýzy korelačnej matice sa podiel celkovej variancie vzhľadom na j-ty faktor rovná:

$$\frac{\widehat{l}_{1j}^2 + \widehat{l}_{2j}^2 + \cdots + \widehat{l}_{pj}^2}{p} \tag{1.3.26}$$

kde $\widehat{l}_{ij}^2$ sú prvky matice $\widehat{\boldsymbol{L}}_{\boldsymbol{Z}}$.

Metóda maximálnej vierohodnosti ponúka oproti predchádzajúcim metódam test adekvátneho počtu spoločných faktorov.

1.4 Rotácia faktorov

Rotácia faktorov je dôležitou vlastnosťou faktorovej analýzy z hľadiska interpretácie faktorov. Aby sme zachovali nekorelovanosť spoločných faktorov, musíme vykonať ortogonálnu transformáciu. Ide o tzv. prerozdelenie faktorových nákladov, pričom ostatné parametre modelu zostávajú rovnaké (komunality, variancie,...).
Ak $\widehat{\boldsymbol{L}}$ je $p \times k$ matica odhadnutých faktorových nákladov, rotáciu faktorov potom získame jej prenásobením ortogonálnou maticou $\boldsymbol{T}$ ($\boldsymbol{T}\boldsymbol{T}'$ =$\boldsymbol{T}'\boldsymbol{T} = \boldsymbol{I}$):

$$\widehat{\boldsymbol{L}}^* = \widehat{\boldsymbol{L}}\boldsymbol{T} \tag{1.4.1}$$

Odhad kovariančnej matice zostáva nezmenený:

$$\widehat{\boldsymbol{L}}\widehat{\boldsymbol{L}}' + \widehat{\boldsymbol{\Psi}} = \widehat{\boldsymbol{L}}\boldsymbol{T}\boldsymbol{T}'\widehat{\boldsymbol{L}}' + \widehat{\boldsymbol{\Psi}} = \widehat{\boldsymbol{L}}^*\widehat{\boldsymbol{L}}^{*\prime} + \widehat{\boldsymbol{\Psi}} \tag{1.4.2}$$

Keďže rovnaké zostávajú aj komunality a špecifické variancie, tak z matematického hľadiska je jedno, či odhadujeme $\widehat{\boldsymbol{L}}$ alebo $\widehat{\boldsymbol{L}}^*$. Rotáciou sa menia faktorové náklady a našim cieľom je získať čo najviac vysokých a čo najviac nízkych hodnôt. Chceme, aby každá premenná vysoko korelovala s jedným spoločným faktorom (príslušné faktorové náklady sú vysoké), pričom s ostatnými bude súvisieť menej. Ak sú pôvodné premenné lineárne závislé a chceme, aby výsledkom faktorovej analýzy bol nízky počet premenných, tak medzi faktorovými nákladmi, ktoré prislúchajú k prvému spoločnému faktoru, by malo byť viac vysokých hodnôt. Platí to pre všetkých k spoločných faktorov. Ideálne je, ak každý spoločný faktor významne vplýva na iné premenné. Môže sa stať, že nejaká premenná vysoko nekoreluje so žiadnym spoločným faktorom. V takom prípade táto premenná možno nie je správne zahrnutá do modelu. Existuje viacero možností ortogonálnej transformácie, ale v tejto práci sa rotácii faktorov nebudeme venovať.

2 Robustné odhady kovariančnej matice

Cieľom tejto diplomovej práce je získať robustné odhady vo faktorovej analýze. Jej východiskom je odhad kovariančnej matice, preto je vhodné, aby bol čo najlepší. V dátach sa často vyskytujú chyby a niektoré merania je lepšie vylúčiť. V tejto kapitole si predstavíme metódy odhadu kovariančnej matice, ktoré budú odolné voči efektu outlierov. Výsledky faktorovej analýzy (matica faktorových nákladov a matica špecifických variancií) získané z jednotlivých robustných odhadov kovariančnej matice budeme potom porovnávať navzájom a s výsledkami získanými z klasického odhadu kovariančnej matice.

2.1 Odhad MCD

Prvá metóda, ktorou sa budeme zaoberať, poskytuje vysoko odolný odhad rozptylovej matice. Princípom tejto metódy ja minimálny determinant kovariančnej matice a tento odhad nazveme *MCD* (z angl. Minimum Covariance Determinant). Snažíme sa nájsť nejakú podmnožinu meraní veľkosti h, keď bude determinant kovariančnej matice najmenší. Zvyčajne sa $h = 3n/4$, kde n je počet všetkých meraní na začiatku. Vo faktorovej analýze potom vychádzame z menšieho počtu meraní (h), ktoré by už mali byť očistené od outlierov. Takýto odhad kovariančnej matice je vysoko robustný a môže byť veľmi rýchlo vypočítaný pomocou algoritmu z článku [3], ktorého autorom je *Rousseeuw* a *Van Driessen*.

2.1.1 Idea algoritmu

Kľúčovým krokom tohto algoritmu je fakt, že štartovaním z akejkoľvek kovariančnej matice k odhadu MCD, je možné vypočítať inú kovariančnú maticu s nižším determinantom. Máme množinu p-rozmerných dát $\boldsymbol{X_n} = \{x_1, x_2, ...x_n\}$. Položme $H_1 \subset \{1, ..., n\}$, kde $|H_1| = h$ a $\boldsymbol{m_1} = (1/h) \sum_{i \in \boldsymbol{H_1}} x_i$, $\boldsymbol{S_1} = (1/h) \sum_{i \in \boldsymbol{H_1}} (x_i - \boldsymbol{m_1})(x_i - \boldsymbol{m_1})'$. Ak $det(\boldsymbol{S_1}) \neq 0$, definujeme Mahalanobisove vzdialenosti:

$$d_1(i) = \sqrt{(x_i - \boldsymbol{m_1})' \boldsymbol{S_1}^{-1} (x_i - \boldsymbol{m_1})} \quad \text{pre } i = 1, ..., n \tag{2.1.1}$$

Teraz vezmeme H_2 takú, že $\{d_1(i); i \in H_2\} := \{(d_1)_{1:n}, ..., (d_1)_{h:n}\}$, kde $(d_1)_{1:n} \leq (d_1)_{2:n} \leq \cdots \leq (d_1)_{n:n}$ sú usporiadané vzdialenosti $d_1(1), ..., d_1(n)$ a vypočítame $\boldsymbol{m_2}$ a $\boldsymbol{S_2}$ na základe množiny H_2. Potom bolo v článku [3] dokázané, že

$$det(\boldsymbol{S_2}) \leq det(\boldsymbol{S_1}) \tag{2.1.2}$$

s rovnosťou vtedy a len vtedy, ak $\boldsymbol{m_2} = \boldsymbol{m_1}$ a $\boldsymbol{S_2} = \boldsymbol{S_1}$.

Predchádzajúci výpočet nazveme C-krok, ktorého opakovanie vedie k iteračnému procesu. Ak $det(\boldsymbol{S_2}) = 0$ alebo $det(\boldsymbol{S_2}) = det(\boldsymbol{S_1})$, zastavíme. Inak počítame ďalší C-krok, čo vedie k $det(\boldsymbol{S_3})$. Potom postupnosť $det(\boldsymbol{S_1}) \geq det(\boldsymbol{S_2}) \geq det(\boldsymbol{S_3}) \geq ...$ je nezáporná a musí konvergovať. Keďže počet podmnožín veľkosti h je konečný, musí existovať index q, kedy $det(\boldsymbol{S_q}) = 0$ alebo $det(\boldsymbol{S_q}) = det(\boldsymbol{S_{q-1}})$, čo zabezpečuje konvergenciu. Základnou ideou algoritmu je vziať veľa počiatočných množín H_1 a aplikovať C-krok kým dosiahneme konvergenciu a uložiť riešenie s najnižším determinantom. Vzniká tu ale niekoľko dôležitých otázok: Ako budeme generovať množiny H_1, s ktorými budeme začínať? Koľko týchto množín H_1 potrebujeme? Ako sa vyhneme opakovaniu, pretože niektoré množiny H_1 môžu viesť k rovnakým výsledkom?

Existujú dve možnosti počiatočného odhadu množiny H_1:

1. Náhodný výber podmožiny veľkosti h.

2. Náhodný výber podmnožiny J veľkosti $(p+1)$ a počítame $\boldsymbol{m_0} := ave(J)$ a $\boldsymbol{S_0} := cov(J)$, kde ave značí priemer a cov kovarianciu. Ak $det(\boldsymbol{S_0}) = 0$, potom rozšírime množinu J pridaním ďalšieho náhodného pozorovania kým $det(\boldsymbol{S_0}) > 0$. Potom vypočítame vzdialenosti $d_0^2(i) := (x_i - \boldsymbol{m_0})' \boldsymbol{S_0}^{-1} (x_i - \boldsymbol{m_0})$ pre $i = 1, ..., n$. Usporiadame ich do postupnosti s permutáciou π: $d_0(\pi(1)) \leq \cdots \leq d_0(\pi(n))$ a vytvoríme množinu $H_1 := \{\pi(1), ..., \pi(h)\}$

Možnosť 1 je síce jednoduchšia, ale ak pri niektorých dátach začneme zo zlej množiny H_1, iterácie nebudú konvergovať k správnemu výsledku. Preto autori algoritmu používajú vždy druhú možnosť. Veľkosť h-podmnožiny si užívateľ môže zvoliť, ale prednastavená je hodnota $[(n + p + 1)/2]$.

Ďalšie výpočty algoritmu závisia od rozmerov pozorovaní:

1. Ak $h = n$, potom odhad MCD priemeru a kovariančnej matice je počítaný ako klasický odhad zo všetkých dát.

2. ak $p = 1$ (jednorozmerné dáta), tak MCD odhad sa počíta iným algoritmom, ktorého autormi sú *Rousseeuw* a *Leroy* (1987) a v tejto práci sa mu nebudeme venovať.

3. Ak $h < n$, $p \geq 2$ a $n \leq 600$, potom:

   - opakujeme (povedzme) 500-krát:
     - ◇ vytvoríme počiatočnú h-podmnožinu H_1 použitím druhej metódy (začíname s podmnožinou veľkosti $(p + 1)$);
     - ◇ pokračujeme dvoma C-krokmi;
   - pre 10 výsledkov z 500 s najnižším determinantom matice S_3 potom pokračujeme C-krokom kým dosiahneme konvergenciu;
   - výstupom je riešenie $(\boldsymbol{m}, \boldsymbol{S})$ s najnižším $det(\boldsymbol{S})$.

4. Ak $n > 600$, potom:

   - vytvoríme päť disjunktných náhodných podmnožín veľkosti n_{sub} podľa článku [3], pričom veľkosti týchto podmnožín by mali byť približne rovnaké;
   - v rámci každej podmnožiny opakujeme 100-krát:
     - ◇ vytvoríme počiatočnú podmnožinu H_1 veľkosti $h_{sub} = [n_{sub}(h/n)]$;
     - ◇ pokračujeme dvoma C-krokmi s n_{sub} a h_{sub};
     - ◇ ponecháme si 10 najlepších výsledkov $(\boldsymbol{m_{sub}}, \boldsymbol{S_{sub}})$;
   - všetky podmnožiny spojíme do množiny veľkosti n_{merged};
   - v spojenej množine opakujeme pre každé z 50 riešení $(\boldsymbol{m_{sub}}, \boldsymbol{S_{sub}})$:
     - ◇ pokračujeme dvoma C-krokmi s n_{merged} a $h_{merged} = [n_{merged}(h/n)]$;
     - ◇ ponecháme si 10 najlepších riešení $(\boldsymbol{m_{merged}}, \boldsymbol{S_{merged}})$

- v rámci celej množiny dát opakujeme pre w_{full} najlepších výsledkov:
  - ◇ urobíme ešte niekoľko C-krokov s n a h;
  - ◇ ponecháme si najlepšie výsledné riešenie $(\boldsymbol{m_{full}}, \boldsymbol{S_{full}})$.

Počet w_{full} a počet C-krokov na konci závisí od veľkosti množiny dát, kde dôležitou je konvergencia. Tento algoritmus je navyše afinne invariantný. Počas iterácii môže nastať prípad, že $det(\boldsymbol{S}) = 0$. Táto možnosť je samozrejme v algoritme *Rousseeuwa* a *Driessena* z článku [3] ošetrená potrebnými krokmi.

2.2 Priestorový medián

Priestorový medián $\widehat{\boldsymbol{\mu}}$ je jedna z viacerých možností odhadu parametra polohy pre p-rozmerné dáta $x_1, ..., x_n$. Je považovaný za zovšeobecnenie jednorozmerného mediánu. Na základe článku [4] vieme, že priestorový medián je jediný, ak je dimenzia dát väčšia ako jedna. Ale jeho nevýhoda je neekvivariantnosť pri afinnej trasformácii dát.

Odhad kovariančnej matice priestorového mediánu získame pomocou aproximácie rozdelenia $\widehat{\boldsymbol{\mu}}$ normálnym rozdelením. Pre tento druhý robustný odhad kovariancie budeme používať skratku *SM* (z angl. Spatial Median).
Priestorový medián $\widehat{\boldsymbol{\mu}}$ minimalizuje sumu:

$$\sum_{i=1}^{n} |x_i - \boldsymbol{\mu}| \tag{2.2.1}$$

kde $|\cdot|$ je Euklidova norma. Táto minimalizácia je známa aj ako *Fermat-Weberov* problém. Ak $\widehat{\boldsymbol{\mu}}$ rieši rovnicu $\sum_{i=1}^{n} \{\mathbf{U}(x_i - \widehat{\boldsymbol{\mu}})\} = 0$, kde $\mathbf{U}(x) = |x|^{-1}x$, potom $\widehat{\boldsymbol{\mu}}$ je priestorový medián.

Popíšme si formálnejšie nejaké teoretické aspekty priestorového mediánu. Majme p-rozmerný náhodný vektor $\boldsymbol{X}$ s kumulatívnou distribučnou funkciou F a $p > 1$. Priestorový medián $\boldsymbol{\mu}$ funkcie F minimalizuje účelovú funkciu:

$$D(\boldsymbol{\mu}) = E\{|x - \boldsymbol{\mu}| - |x|\} \tag{2.2.2}$$

Poznamenajme, že na základe [4] nie sú potrebné žiadne momentové predpoklady, ale pre asymptotickú teóriu predpokladáme, že:

(P1) p-rozmerná funkcia hustoty vektora $\boldsymbol{X}$ je spojitá a ohraničená,

(P2) priestorový medián rozdelenia $\boldsymbol{X}$ sa rovná nule a je jediný.

Ďalej definujme vektorové a maticové hodnotové funkcie:

$$\begin{aligned} \mathbf{U}(x) &= \frac{x}{|x|} \\ \mathbf{A}(x) &= \frac{1}{|x|}\left[I_p - \frac{xx'}{|x|^2}\right] \\ \mathbf{B}(x) &= \frac{xx'}{|x|^2} \end{aligned} \tag{2.2.3}$$

pre $x \neq 0$ a $\mathbf{U}(0) = 0$ a $\mathbf{A}(0) = \mathbf{B}(0) = 0$. Taktiež zavedieme značenie $\mathbf{A} = E\{\mathbf{A}(x)\}$ a $\mathbf{B} = E\{\mathbf{B}(x)\}$. Asymptotický odhad kovariančnej matice sa potom rovná $\mathbf{A^{-1}BA^{-1}}$. Predpokladáme, že skutočná hodnota $\boldsymbol{\mu} = \mathbf{0}$ a odhadneme $\mathbf{A}$ a $\mathbf{B}$:

$$\begin{aligned} \widehat{A} &= ave\{\mathbf{A}(x_i - \widehat{\boldsymbol{\mu}})\} \\ \widehat{B} &= ave\{\mathbf{B}(x_i - \widehat{\boldsymbol{\mu}})\} \end{aligned} \tag{2.2.4}$$

kde *ave* značí priemer.

Rozdelenie $\widehat{\boldsymbol{\mu}}$ sa dá podľa [4] aproximovať normálnym rozdelením $N_p(\boldsymbol{\mu}, \frac{1}{n}\widehat{A}^{-1}\widehat{B}\widehat{A}^{-1})$. Priestorový medián vypočítame iteračne v nasledujúcich 2 krokoch:

Krok 1: $e_i \leftarrow x_i - \boldsymbol{\mu}$, $i = 1, ..., n$

Krok 2: $\boldsymbol{\mu} \leftarrow \boldsymbol{\mu} + (\sum_{i=1}^{n} |e_i|^{-1})^{-1} \sum_{i=1}^{n} \mathbf{U}(e_i)$

Ako počiatočný odhad $\widehat{\boldsymbol{\mu}}$ môžeme určiť priemer. Odhad kovariančnej matice sa potom rovná $\frac{1}{n}\widehat{A}^{-1}\widehat{B}\widehat{A}^{-1}$, ktorý vieme na základe vzťahov (2.2.3) a (2.2.4) ľahko dopočítať. Iteračný cyklus ukončíme, keď sa posledné 2 iterácie líšia iba o ϵ.

2.3 Identifikácia outlierov na základe koeficientu špicatosti

V tejto podkapitole si predstavíme ďalšiu alternatívnu procedúru pre robustný odhad kovariančnej matice, ktorý budeme ďalej označovať skratkou *OD*. V množine pozorovaní sa táto metóda snaží odhaliť outlierov, ktoré vylúči a odhad kovariančnej matice sa potom počíta z očistenej podmnožiny dát. Je založená na analýze projekcií meraní do určitých $2p$ smerov, kde p je rozmer priestoru meraní. Tieto smery sú získané maximalizáciou a minimalízáciu koeficientu špicatosti. Idea použitia projekcií je základom pre niekoľko podobných algoritmov identifikácie outlierov. Tieto

procedúry sa spoliehajú na fakt, že vo viacrozmerných dátových súboroch je každý outlier extrémnym bodom. Avšak ich nevýhodov je, že vyžadujú projekciu dát do náhodne generovaných smerov, ktorých počet musí byť vysoký, aby boli tieto metódy úspešné. My sa budeme snažiť vybrať konečný počet smerov na základe hodnoty koeficientu špicatosti pozorovaní. Prítomnosť outlierov v dátach môže na základe článku [5] zvýšiť, ale aj znížiť koeficient špicatosti pozorovaných dát. Preto sa v tejto metóde budeme snažiť identifikovať outlierov projekciou dát do smerov, ktoré maximalizujú alebo minimalizujú špicatosť meraní, čo by malo byť postačujúce. Nízky počet outlierov spôsobuje ťažké chvosty a zvyšuje špicatosť. Ale zvýšenie počtu outlierov môže znížiť koeficient špicatosti.

2.3.1 Popis algoritmu

Predpokladáme, že máme dané pozorovania $x_1, x_2, ..., x_n$ p-rozmerného náhodného vektora X. Algoritmus je založený na projektovaní každého pozorovania do $2p$ smerov a následnom analyzovaní jednorozmerných projekcií na tieto smery, ktoré sú získané ako riešenie $2p$ jednoduchých optimalizačných problémov. Postupujeme podľa nasledujúcich krokov:

1. Originálne dáta preškálujeme:

$$y_i = S^{-1/2}(x_i - \overline{x}), \quad i = 1, ...n$$

   kde $\overline{x}$ je priemer a S kovariančná matica pôvodných premenných.

2. Vypočítame p ortogonálnych smerov a projekcií, ktoré maximalizujú koeficient špicatosti:

   a. Nastavíme $y_i^{(1)} = y_i$ a iteračný index $j = 1$

   b. Smer, ktorý maximalizuje koeficient špicatosti je riešením problému:

$$d_j = \arg\max_d \frac{1}{n} \sum_{i=1}^{n} (d' y_i^{(j)})^4$$

$$d'd = 1$$

c. Merania projektujeme do menej rozmerného podpriestoru, ortogonálneho na smer d_j. Definujeme:

$$v_j = d_j - e_1, \quad Q_j = I - \frac{v_j v_j'}{v_j' d_j} \text{ ak } v_j' d_j \neq 0$$
$$\text{inak} \quad Q_j = I$$

kde e_1 je prvý jednotkový vektor. Výsledná matica Q_j je ortogonálna a môžeme vypočítať nové hodnoty:

$$u_i^j \equiv \begin{pmatrix} z_i^j \\ y_i^{j+1} \end{pmatrix} = Q_j y_i^j, \quad i = 1, ..., n$$

kde $z_i^{(j)}$ je prvý komponent $u_i^{(j)}$, ktorý spĺňa $z_i^{(j)} = d_j' y_i^{(j)}$ (jednorozmerná projekčná hodnota) a $y_i^{(j+1)}$ zodpovedá zvyšným $(p-j)$ komponentom $u_i^{(j)}$. Nastavíme $j = j + 1$ a ak $j < p$, vrátime sa ku kroku 2b. Inak položíme $z_i^{(p)} = y_i^{(p)}$

3. Vypočítame ďalších p ortogonálnych smerov a projekcií, ktoré minimalizujú koeficient špicatosti:

   a. Nastavíme $y_i^{(p+1)} = y_i$ a $j = p + 1$

   b. Opakujeme predchádzajúce kroky 2b a 2c, ale riešime minimalizačný problém:

$$d_j = \arg\min_d \frac{1}{n} \sum_{i=1}^{n} (d' y_i^{(j)})^4$$
$$d'd = 1$$

4. Aby sme mohli určiť, či $z_i^{(j)}$ je outlierom v niektorom z $2p$ smerov, vypočítame jednorozmernú hodnotu pre každé meranie:

$$r_i = \max_{1 \leq j \leq 2p} \frac{|z_i^{(j)} - median(z^{(j)})|}{MAD(z^{(j)})},$$

kde $MAD(z^{(j)}) = median(|z^{(j)} - median(z^{(j)})|)$.

5. Tieto hodnoty r_i sú použité na testovanie, či dané meranie môže byť označené ako outlier. Ak je $r_i > \beta_p$, potom meranie i je považované za outliera. Hraničné hodnoty β_p, ktoré závisia od rozmeru premenných, sú uvedené v tabuľke 1. Boli vypočítané simuláciami v článku [5] a ich ďalšie hodnoty ľahko dopočítame lineárnou interpoláciou $\log \beta_p$ na $\log p$ ako to navrhli v článku [5].

6. Ak bola podmienka v kroku 5 splnená pre niektoré i, tak vytvoríme novú vzorku všetkých pozorovaní, pre ktoré platí $r_i \leq \beta_p$. Procedúru, teda kroky 1.-5., potom opakujeme s redukovanou vzorkou pozorovaní kým pre žiadne pozorovanie neplatí $r_i > \beta_p$ alebo počet zostávajúcich pozorovaní by bol menší ako $\lfloor (n+p+1)/2 \rfloor$.

7. Nakoniec vypočítame Mahalanobisovu vzdialenosť pre všetky pozorovania označené ako outlier v predchádzajúcich krokoch použitím dát (stredná hodnota a kovariančná matica) z množiny ostatných pozorovaní, ktorú si označíme U. Počítame:

$$\widetilde{m} = \frac{1}{|U|} \sum_{i \in U} x_i$$

$$\widetilde{S} = \frac{1}{|U|-1} \sum_{i \in U} (x_i - \widetilde{m})(x_i - \widetilde{m})'$$

$$v_i = (x_i - \widetilde{m})' \widetilde{S}^{-1} (x_i - \widetilde{m}) \qquad \forall i \notin U$$

Tie pozorovania, pre ktoré platí, že $i \notin U$ a $v_i < \chi^2_{p,0.99}$, pridáme do množiny U.

Rozmer premenných p	5	10	20
Hraničná hodnota β_p	4,1	6,9	10,8

Tabuľka 1: Hraničné hodnoty jednorozmerných projekcií

2.3.2 Výpočet projekčných smerov

Hlavný výpočtový problém v aplikácii predchádzajúceho algoritmu je spojený s určením optimálneho smeru d_j. Tento výpočet môžeme uskutočniť dvoma spôsobmi:

1. Aplikovaním modifikovanej verzie Newtonovej metódy.

2. Získaním riešenia priamo z prvých podmienok optimality, ktoré sú rovnaké pre maximalizačnú, aj minimalizačnú úlohu:

$$\begin{aligned} 4 \sum_{i=1}^{n} \left(d' y_i^{(j)} \right)^3 y_i^{(j)} - 2\lambda d = 0 \\ d'd = 1 \end{aligned} \tag{2.3.1}$$

Vynásobením prvej rovnice vektorom d a dosadením obmedzenia, dostaneme hodnotu λ. Výsledná podmienka potom vyzerá takto:

$$\left(\sum_{i=1}^{n}\left(d'y_i^{(j)}\right)^2 y_i^{(j)}y_i^{(j)\prime}\right)d = \sum_{i=1}^{n}\left(d'y_i^{(j)}\right)^4 d \tag{2.3.2}$$

Z predchádzajúcej rovnice potom vyplýva, že optimálne d bude jednotkovým vlastným vektorom matice:

$$M(d) = \sum_{i=1}^{n}\left(d'y_i^{(j)}\right)^2 y_i^{(j)}y_i^{(j)\prime} \tag{2.3.3}$$

Iteračnú procedúru na výpočet smeru d môžeme zapísať v nasledovných krokoch:

1. Výber počiatočného smeru, ktorý spĺňa, že $\|\bar{d}_0\| = 1$
2. V iterácii $l+1$ počítame $\bar{d}_{l+1}$ ako jednotkový vektor prislúchajúci k najväčšej (resp. najmenšej) vlastnej hodnote matice $M(\bar{d}_l)$
3. Iterácie ukončíme, ak $\|\bar{d}_{l+1} - \bar{d}_l\| < \epsilon$ a nastavíme $d_j = \bar{d}_{l+1}$

V druhom kroku počítame vlastný vektor prisluchájúci najväčšej vlastnej hodnote, ak v algoritme v podkapitole 2.3.1 riešime v kroku 2b maximalizačný problém. Ak riešime minimalizačný problém, čo zodpovedá kroku 3b, tak smer d počítame ako vlastný vektor matice $M(d)$ prislúchajúci najmenšej vlastnej hodnote. Počiatočný smer určíme na základe článku [5] ako hlavný komponent rozkladu kovariančnej matice normalizovaných meraní $y_i^{(j)}/\|y_i^{(j)}\|$. V prípade maximalizácie to bude najväčší hlavný komponent a v prípade minimalizácie najmenší. Keďže pozorovania sú štandardizované, tak tieto smery sú afinne invariantné.

Pri programovaní algoritmu sme zistili, že optimalizácia v kroku 3b nekonverguje použitím tejto druhej možnosti získania optimálneho smeru d_j. Preto pri minimalizácii použijeme Newtonovu metódu:

$$\bar{d}_{l+1} = \bar{d}_l - [Hf(\bar{d}_l)]^{-1}\nabla f(\bar{d}_l), \tag{2.3.4}$$

kde $Hf(\bar{d}_l)$ je Hesián a $\nabla f(\bar{d}_l)$ gradient účelovej funkcie $f(d) = \frac{1}{n}\sum_{i=1}^{n}(d'y_i)^4$. Keďže optimalizačná úloha má ohraničenie $d'd = 1$, použijeme penaltovú funkciu na základe knihy [6] a budeme minimalizovať fukciu:

$$F(d, r) = f(d) + r(d'd - 1)^2 \tag{2.3.5}$$

Pridanie penaltovej funkcie nám zabezpečí, že dostaneme riešenie, ktoré bude spĺňať ohraničenie, čiže bude ležať na jednotkovej kružnici. Penaltová funkcia má potom na kružnici nulovú hodnotu a mimo nej rýchlo rastie. Optimálne riešenie minimalizačného problému v kroku 3b získame iteračným procesom:

1. Vstup: d_0, r_0, $k = 0$, tolerancia ϵ

2. Opakujeme nasledujúce výpočty kým $(d'd - 1) > -\epsilon$:
   $k = k + 1$
   $r_k = 10 * r_{k-1}$
   Nájdeme optimálne d_k Newtonovou metódou:
   $F(d, r_k) = \frac{1}{n} \sum_{i=1}^{n} (d'y_i)^4 + r_k(d'd - 1)^2$,
   kde $\bar{d_0} = d_{k-1}$

3. Výstup $d_j = d_k$

Výstup d_j predchádzajúceho iteračného algoritmu je optimálnym riešením minimalizačného problému algoritmu v podkapitole 2.3.1 v kroku 3b.

Rozmer premenných p	5	10	20
Hraničná hodnota k_d	0,98	0,95	0,92

Tabuľka 2: Hraničné hodnoty jednorozmerných projekcií

2.3.3 Robustný odhad kovariančnej matice

Po aplikovaní predchádzajúceho algoritmu v podkapitole 2.3.1 získame vzorku pozorovaní očistených od outlierov (množina U), z ktorej môžeme vypočítať robustný odhad strednej hodnoty a kovariančnej matice:

$$\tilde{m} = \frac{1}{|U|} \sum_{i \in U} x_i \tag{2.3.6}$$

$$\widetilde{S} = \frac{1}{(|U| - 1)k_d} \sum_{i \in U} (x_i - m)(x_i - m)' \tag{2.3.7}$$

kde U je množina všetkých pozorovaní, ktoré nie sú považované za outlierov. $|U|$ značí počet pozorovaní v tejto množine a k_d je konštanta, ktorá má zabezpečiť, že stopa odhadnutej matice bude nevychýlená. Hodnoty k_d, ktoré sú uvedené v tabuľke 2 závisia od rozmeru premenných a boli vypočítané simuláciami autormi článku [5]. Ďalšie hodnoty vieme dopočítať lineárnou interpoláciou $\log k_p$ na $\log p$.

3 Simulácie Monte Carlo

V tejto kapitole budeme skúmať vplyv outlierov na odhady členov rozkladu (1.2.2) pomocou simulácií, ktorých scenár pochádza z článku [2]. Najprv si vygenerujeme skutočnú $p \times k$ maticu faktorových nákladov $\boldsymbol{L} \sim N(0, \frac{1}{9})$ a $p \times p$ diagonálnu maticu špecifických variancií $\boldsymbol{\Psi}$, ktorej prvky sú z rovnomerného rozdelenia na intervale $[0, 1]$. Potom budeme simulovať množinu dát $\boldsymbol{X}^{(s)}$, kde s označuje simulované dáta, podľa faktorového modelu, ku ktorému ešte pridáme maticu Out, ktorá vytvorí z niektorých meraní outlierov:

$$\boldsymbol{X}^{(s)} = \boldsymbol{L}\boldsymbol{F}^{(s)} + \boldsymbol{\varepsilon}^{(s)} + Out^{(s)}. \tag{3.0.8}$$

Pre každé s generujeme $k \times n$ maticu faktorových skóre $\boldsymbol{F}^{(s)}$ z $N(0, 1)$ a $p \times n$ maticu špecifických faktorov, ktorej prvky $\varepsilon_{ij}^{(s)}$ pochádzajú z $N(0, \psi_j)$. Prvky matice Out sú väčšinou nulové až na n_{out} hodnôt. V každej simulácii náhodne vyberieme n_{out} meraní a pre každé toto meranie náhodne vyberieme jeden prvok, pre ktorý vygenerujeme hodnotu z normálneho rozdelenia $N(10, (0, 05)^2)$. Následne odhadneme faktorový model, ktorého vstupom bude kovariančná matica simulovaného vektora pozorovaní $\boldsymbol{X}^{(s)}$ pre $s = 1, ..., m$. Počet spoločných faktorov k považujeme pri odhadovaní za známe. Výstupom odhadu bude matica faktorových nákladov $\widehat{\boldsymbol{L}}^{(s)}$ a matica špecifických variancií $\boldsymbol{P}^{(s)}$ pre $m = 1000$ simulácií. Tieto výsledky budeme porovnávať so skutočnými dátami. Keďže matica faktorových nákladov nie je vzhľadom na ortogonálne transformácie určená jednoznačne, tak namiesto nej vezmeme $p \times p$ maticu $\boldsymbol{A}^{(s)} = \widehat{\boldsymbol{L}}^{(s)}(\widehat{\boldsymbol{L}}^{(s)})'$ a porovnáme ju so skutočnou $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}'$. Pre tento účel si vypočítame strednú kvadratickú chybu:

$$MSE(a_{ij}) = \frac{1}{m} \sum_{s=1}^{m} \left(a_{ij}^{(s)} - a_{ij} \right)^2$$

kde $a_{ij}^{(s)}$ sú prvky matice $\boldsymbol{A}^{(s)}$ a a_{ij} prvky matice $\boldsymbol{A}$ pre $i, j = 1, ..., p$. Potom definujeme priemer strednej kvadratickej chyby ako $MSE(\boldsymbol{A}) = \frac{1}{p^2} \sum_{i=1}^{p} \sum_{j=1}^{p} MSE(a_{ij})$.
Podobne porovnáme druhú odmocninu špecifických variancií z odhadu simulovaných dát so skutočnými hodnotami. Takže počítame:

$$MSE(P_j) = \frac{1}{m} \sum_{s=1}^{m} \left(\sqrt{P_j^{(s)}} - \sqrt{\psi_j} \right)^2$$

pre $j = 1, ..., p$ a definujeme priemer MSE ako $MSE(\boldsymbol{P}) = \frac{1}{p} \sum_{j=1}^{p} MSE(P_j)$.

Výsledkom simulácií budú krivky znázorňujúce priemernú strednú kvadratickú chybu v závislosti od podielu outlierov. Faktorový model budeme odhadovať metódou maximálnej vierohodnosti (MLE) a metódou hlavných faktorov (PFA) na základe podkapitoly 1.3.2, ktorého východiskom bude klasický odhad kovariančnej matice a tri robustné odhady (MCD, SM, OD) popísané v predchádzajúcej kapitole.

Na výpočty sme použili štatistický program R, ktorý ponúka množstvo naprogramovaných metód. My sme využili funkciu na odhad faktorovej analýzy metódou maximálnej vierohodnosti (`factanal(...)`) a funkciu na *MCD*-odhad kovariančnej matice (`CovMcd(...)`) z balíka `rrcov`. Vstupom pre funkciu `factanal` je odhad kovariančnej matice, ale naprogramované je to tak, že výsledkom je rozklad korelačnej matice. Keďže naše skutočné matice $\boldsymbol{A}$ a $\boldsymbol{\Psi}$ sú rozkladom kovariančnej matice, musíme si výsledky zo simulácií transformovať. Každý prvok $a_{ij}^{(s)}$ matice $\boldsymbol{A}^{(s)}$ musíme prenásobiť výberovými smerodajnými odchýlkami z dát $S_i S_j$ a každý prvok $P_j^{(s)}$ vektora špecifických variancií musíme prenásobiť druhou mocninou výberových smerodajných odchýlok z dát S_j^2 pre $i, j = 1, 2, ..., p$. Metódu hlavných faktorov sme si museli naprogramovať a podľa teórie sme odhad robili tiež z korelačnej matice. To znamená, že klasický aj všetky robustné odhady rozkladu kovariančnej matice sme previedli na korelačnú maticu, odhadli sme faktorový model a výslednú maticu $\boldsymbol{A}^{(s)}$ a vektor špecifických variancií $\boldsymbol{P}^{(s)}$ sme transformovali rovnako ako pri metóde maximálnej vierohodnosti, aby sa rozklad rovnal rozkladu kovariančnej matici. Je to dôležité pre vzájomné porovnanie so skutočným rozkladom. Pri odhade faktorového modelu môže nastať, že všetky špecifické variancie ψ_j, $j = 1, ..., p$ nebudú kladné. Nazýva sa to *Heywoodov* prípad a riešením je nahradenie záporných hodnôt nulou. Jedným z krokov metódy hlavných faktorov je odhad vlastných hodnôt kovariančnej matice a následne výpočet ich prvých k (počet spoločných faktorov) odmocnín pre odhad matice faktorových nákladov. Avšak počas našich výpočtov nastal prípad, keď týchto $k-1$ vlastných hodnôt bolo záporných. Prvá vlastná hodnota je vždy kladná, pretože kovariančná matica je kladne semidefinitná. V každej sade 1000 simulácií ich ale nebolo viac ako 10 a tieto prípady sme z výpočtov vylúčili, aby bol počet spoločných faktorov vždy rovnaký. Funkcie na výpočet *OD* a *SM*-odhadu kovariančnej matice sme si taktiež naprogramovali podľa teórie predchádzajúcej kapitoly. Dôležitým krokom algoritmu na *OD*-odhad kovariančnej

matice je určenie optimálneho projekčného smeru, ktorý je opísaný v podkapitole 2.3.2. V iteráciách s maximalizáciou sme optimálny smer vypočítali druhým spôsobom ako najväčší vlastný vektor matice $M(d)$, ktorá je definovaná vzťahom (2.3.3). Pri minimalizácií sme pre zabezpečenie konvergencie použili Newtonovu metódu s penaltovou funkciou.

3.1 Výsledky

V simuláciách budeme generovať maticu $X^{(s)}$ s rozmermi $p = 5$ a $n = 100$ a počet spoločných faktorov $k = 2$. Robustnosť odhadov rozkladu kovariančnej matice budeme testovať na viacerých spôsoboch generovania outlierov. Na základe článku [2] sme simulovali dáta a odhadovali faktorový model (3.0.8). Pre každé percento outlierov s diferenciou 2% od 0% po 20% sa počíta $m = 1000$ simulácií a výsledkom je priemerná hodnota strednej kvadratickej chyby. Vo funkcii na výpočet *MCD*-odhadu musíme zadať veľkosť podmnožiny, z ktorej sa bude počítať odhad kovariančnej matice. Pri maximálnom podiele 20% outlierov sme túto veľkosť na základe článku [2] určili na 3/4, aby sme mali zaručené, že dáta budú dostatočne očistené od všetkých chýb. R-ko poskytuje viacero úvodných nastavení generátora náhodných čísel. Pri dostatočnom počte simulácií by ale výsledky pri rôznych nastaveniach generátora mali byť skoro rovnaké. Keď sa naše prvé výsledky rádovo nepodobali s výsledkami z článku [2], zistili sme, že 1000 simulácií nie je dostatočný počet na to, aby sa výsledky pri rôznych typoch generovania náhodných čísel nelíšili. Avšak výpočet celého programu je časovo náročný, preto sme sa rozhodli, že počet simulácií nebudeme zvyšovať. Dôležité bude porovnávať vzájomný vývoj a trend kriviek strednej kvadratickej chyby pre jednotlivé odhady kovariančnej matice. Pre ilustráciu sme faktorový model (3.0.8) odhadli pre dva typy generovania náhodných čísel s 1000 simuláciami pre každé percento outlierov a raz s 2000 simuláciami.

3.1.1 Výsledky 1

Na obrázku 1 a v tabuľke 3 môžeme vidieť prvé získané výsledky pre 1000 simulácií. Štruktúra ďalších obrázkov a tabuliek bude rovnaká, preto si teraz na začiatku vysvetlí-

me, čo znázorňujú. Na každom obrázku sú dva grafy, pričom graf naľavo znázorňuje priemerné stredné kvadratické chyby (*MSE.A*) v závislosti od percenta outlierov pre maticu $A = LL'$, kde L je matica faktorových nákladov a napravo pre špecifické variancie (*MSE.P*). Výpočet týchto stredných kvadratických chýb sme si vysvetlili na začiatku tejto kapitoly. V oboch grafoch sú krivky pre odhad faktorového modelu metódou maximálnej vierohodnosti (*A.cl, A.mcd, A.sm, A.od* a *P.cl, P.mcd, P.sm, P.od*) a metódou hlavných faktorov (*A.clP, A.mcdP, A.smP, A.odP* a *P.clP, P.mcdP, P.smP, P.odP*) pre štyri rôzne odhady kovariančnej matice. V ich názvoch sú už spomenuté skratky odhadov, pričom pre klasický odhad kovariančnej matice sme si zvolili skratku *cl*. V hornej časti tabuľky sú konkrétne hodnoty pre faktorové náklady a v spodnej časti pre špecifické variancie.

Vráťme sa teraz k prvým výsledkom. Vidíme, že veľkosť MSE pre špecifické variancie je rádovo vyššia oproti MSE pre faktorové náklady. Výsledky pre metódu hlavných faktorov sú pre všetky odhady kovariančnej matice lepšie, iba pre odhad pomocou priestorového mediánu (*SM*) sú pre faktorové náklady rovnaké. *MCD*-odhad kovariančnej matice je známy a je považovaný za veľmi dobrý robustný odhad, ale v grafe pre faktorové náklady vidíme, že modrá krivka, ktorá predstavuje metódu hlavných faktorov z klasického odhadu kovariančnej matice, je pod červenou krivkou, ktorá zodpovedá metóde maximálnej vierohodnosti z *MCD*-odhadu kovariančnej matice. Dôvodom bude zrejme zadaná veľkosť podmnožiny, z ktorej sa má počítať *MCD*-odhad. Stanovili sme ho pevne pre všetky generované percentá outlierov na 3/4. Pozitívnym výsledkom je, že nami programované funkcie na výpočet *SM* a *OD*-odhadu kovariančnej matice dávajú pre faktorové náklady v tomto prvom prípade generovania outlierov lepšie výsledky. Avšak pre špecifické variancie je výška strednej kvadratickej chyby odhadu kovariančnej matice priestorového mediánu (*SM*) dosť vysoká.

3.1.2 Výsledky 2

Obrázok 2 a tabuľka 4 zodpovedajú rovnakým výpočtom ale s iným úvodným nastavením generátora náhodných čísel. Keďže rozsah osí je vždy pre jednotlivé grafy rovnaký, hneď vidíme, že v tomto druhom prípade sú kvadratické chyby nižšie. Až na jednu zmenu sa zachoval trend jednotlivých kriviek. V predchádzajúcom obrázku bola

červená krivka pre faktorové náklady nad modrou, ale v tomto prípade sa približne v strede pretínajú a červená sa dostáva pod modrú. Znamená to, že ak je v dátach viac ako 10% outlierov, tak odhad faktorového modelu metódou maximálnej vierohodnosti z *MCD*-odhadu kovariančnej matice je lepší ako odhad metódou hlavných faktorov z klasického odhadu kovariančnej matice. Tento výsledok viac podporuje predpoklad, že faktorová analýza vychádzajúca z *MCD*-odhadu je vždy lepšia ako faktorová analýza vychádzajúca z klasického odhadu kovariančnej matice. Dôležité je správne určiť percento outlierov v dátach a čím viac sa priblížime ku skutočnosti, tak výsledky by mali byť robustnejšie. Na obrázku 3 a v tabuľke 5 sú výsledky pre 2000 simulácií pre každé percento outlierov a skôr sa podobajú našim druhým výsledkom, preto v ďalších výpočtoch s iným vytváraním outlierov budeme používať rovnaký typ generátora náhodných čísel.

3.1.3 Výsledky 3

Pri ďalšom testovaní sme sa rozhodli simulovať dáta podľa rovnakého modelu (3.0.8), ale zmeníme maticu Out. Znova v každej simulácii náhodne vyberieme n_{out} meraní, ale pre každé meranie vygenerujeme namiesto jedného prvku všetkých p prvkov z rozdelenia $N(10, (0,05)^2)$. Výsledky výpočtov sú na obrázku 4 a v tabuľke 6. Pri porovnaní s druhými výsledkami z obrázku 2 vidíme, že krivky pre robustné odhady kovariančnej matice sa v podstate nezmenili. Trochu sa iba zhoršili hodnoty *MSE* faktorových nákladov *OD*-odhadu pri 20% outlierov. Tento druhý spôsob generovania outlierov negatívnejšie ovplyvnil výsledky oboch metód odhadu vychádzajúcich z klasického odhadu kovariančnej matice. Hoci je modrá a čierna krivka pre špecifické variancie dosť nízko, pre faktorové náklady stúpajú ich hodnoty veľmi vysoko. Môžeme povedať, že aj v tomto prípade sú použité robustné odhady kovariančnej matice naozaj lepšie.

3.1.4 Výsledky 4

Dôležitým predpokladom faktorovej analýzy je normalita spoločných faktorov a špecifických faktorov, preto v nasledujúcich troch výpočtoch tento predpoklad porušíme. Dáta budeme generovať zo Studentovho rozdelenia, pretože má ťažšie chvosty ako normálne rozdelenie a tým je systematicky daný vznik outlierov. Čím je vyšší počet stupňov

voľnosti Studentovho rozdelenia, tým viac sa podobá normálnemu rozdeleniu. Podľa predpokladov ale zachováme výšku strednej hodnoty a disperzie. Matica *Out* už nebude vstupovať do modelu, ktorý bude pri ďalších troch prípadoch generovania pozorovaní vyzerať takto:

$$\boldsymbol{X}^{(s)} = \boldsymbol{L}\boldsymbol{F}^{(s)} + \boldsymbol{\varepsilon}^{(s)}. \tag{3.1.1}$$

Maticu $\boldsymbol{F}^{(s)}$ a $\boldsymbol{\varepsilon}^{(s)}$ budeme postupne vytvárať pre 33 až 3 stupne voľnosti s diferenciou 3 a pre všetky stupne voľnosti počítame 1000 simulácií. Dáta už ale nebudeme generovať zo Studentovho rozdelenia s počtom stupňov voľnosti 1 a 2, pretože pre dané parametre nie je definovaná variancia rozdelenia. V prvom prípade budeme v každej simulácii generovať iba maticu $\boldsymbol{\varepsilon}^{(s)}$ zo Studentovho rozdelenia so strednou hodnotou 0 a rozptylom ψ_j, $j = 1, 2, ..., p$. Pre každé meranie generujeme p-rozmerný vektor zo Studentovho rozdelenia podľa definície:

$$Student_{df} \sim \frac{N((0, \cdots, 0)', I_p)}{\sqrt{\frac{\chi^2_{df}}{df}}},$$

kde df označuje počet stupňov voľnosti a I_p p-rozmernú jednotkovú maticu. Takéto dáta majú potom varianciu rovnú $df/(df-2)$. Aby sme dosiahli, že špecifické faktory budú mať disperziu ψ_j pre $j = 1, 2, ..., p$, musíme ich predeliť $df/(df-2)$ a vynásobiť druhou odmocninou ψ_j pre $j = 1, 2, ..., p$. Matica $\boldsymbol{F}^{(s)}$ bude z normálneho rozdelenia podľa predpokladov modelu. Tieto výsledky sú na obrázku 5 a v tabuľke 7. Na prvý pohľad vidíme, že krivky pre klasický odhad kovariančnej matice nestúpajú tak, ako v predchádzajúcich grafoch. Hodnoty strednej kvadratickej chyby pre metódu hlavných faktorov sú opäť nižšie ako pre metódu maximálnej vierohodnosti. Pri nízkom počte stupňov voľnosti už krivky pre špecifické variancie začínajú stúpať, ako aj modrá a čierna krivka pre faktorové náklady. Takto vytvárané pozorovania teda pri nízkom počte stupňov voľnosti negatívne ovplyvnili všetky odhady faktorového modelu. Ale pozitívny výsledok vidíme v tom, že hodnoty strednej kvadratickej chyby pre faktorové náklady stúpajú iba pri klasickom odhade kovariančnej matice. V porovnaní s grafmi na obrázku 2 vidíme, že porušenie predpokladu normality špecifických faktorov nemá taký významný vplyv na kvalitu odhadu ako pri generovaní dát podľa modelu (3.0.8).

3.1.5 Výsledky 5

Na obrázku 6 a v tabuľke 8 sú výsledky pre dáta podľa modelu (3.1.1), kde matica $\boldsymbol{\varepsilon}^{(s)}$ je v každej simulácii generovaná z normálneho rozdelenia s rovnakými parametrami ako doteraz a matica $\boldsymbol{F}^{(s)}$ je zo Studentovho rozdelenia so strednou hodnotou 0 a disperziou 1. Takéto porušenie predpokladu faktorového modelu výrazne neovplyvní jeho odhad. Všetky krivky sú skoro vodorovne. Dokonca najnižsie chyby má metóda hlavných faktorov z klasického odhadu kovariančnej matice. Z predchádzajúcich dvoch výsledkov by sme mohli povedať, že voľba odhadu kovariančnej matice nemá zásadný vplyv na faktorovú analýzu, keď sú spoločné faktory alebo špecifické faktory zo Studentovho rozdelenia.

3.1.6 Výsledky 6

Nakoniec budeme generovať zo Studentovho rozdelenia maticu $\boldsymbol{F}^{(s)}$ aj maticu $\boldsymbol{\varepsilon}^{(s)}$. Parametre rozdelenia zostávajú rovnaké. Hodnoty *MSE* pre tieto pozorovania sú na obrázku 7 a v tabuľke 9. Vidíme tu podobnosť s grafmi na obrázku 5, ktorý prislúcha dátam simulovaným podľa modelu 3.1.1, kde špecifické faktory sú zo Studentovho rozdelenia. Pri nízkom počte stupňov voľnosti sa znova zhoršuje odhad špecifických variancií pre všetky odhady kovariančnej matice.

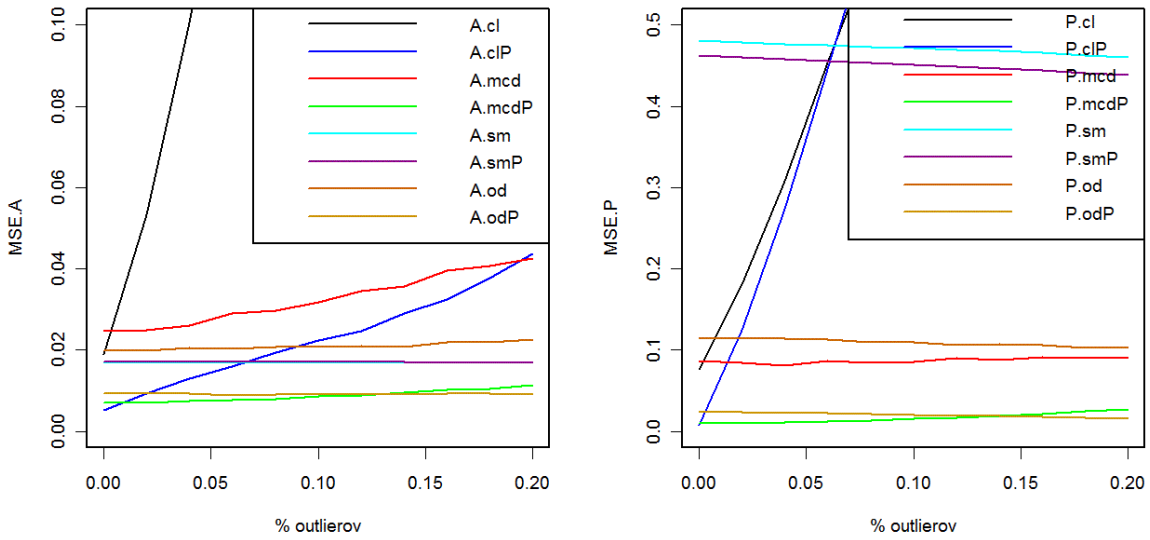
3.1.7 Výsledky 7

V predposlednom prípade budeme simulovať dáta opäť podľa modelu (3.1.1), ale zameriame sa na maticu $\boldsymbol{L}$ a vypočítame si ešte novú maticu $\bar{\boldsymbol{L}}=\boldsymbol{L}+0,5$. Niekoľko náhodných meraní matice $\boldsymbol{X}$ bude potom generovaných pomocou tejto novej matice $\bar{\boldsymbol{L}}$. Predpokladáme, že týmto spôsobom vytvoríme v dátach outlierov. Počet simulácií zostáva rovnaký, kde náhodne vybraných odlišne počítaných stĺpcov matice $\boldsymbol{X}$ bude od 0 po 20 s diferenciou 2. Tieto výsledky môžeme vidieť na obrázku 8 a v tabuľke 10. Odhady všetkých robustných metód nevykazujú vysoké chyby ako v ostatných testoch. Klasický odhad kovariančnej matice síce nedopadol až tak zle, ako v druhých výsledkoch, ale príslušné krivky rastú rýchlejšie ako ostatné. Platí to ale iba pre stredné kvadratické chyby faktorových nákladov. Zaujímave je, že takéto generovanie chýb v dátach

neovplyvnilo výrazne odhad špecifických variancií ani pre klasický odhad kovariančnej matice v porovananí s ostatnými prípadmi.

3.1.8 Výsledky 8

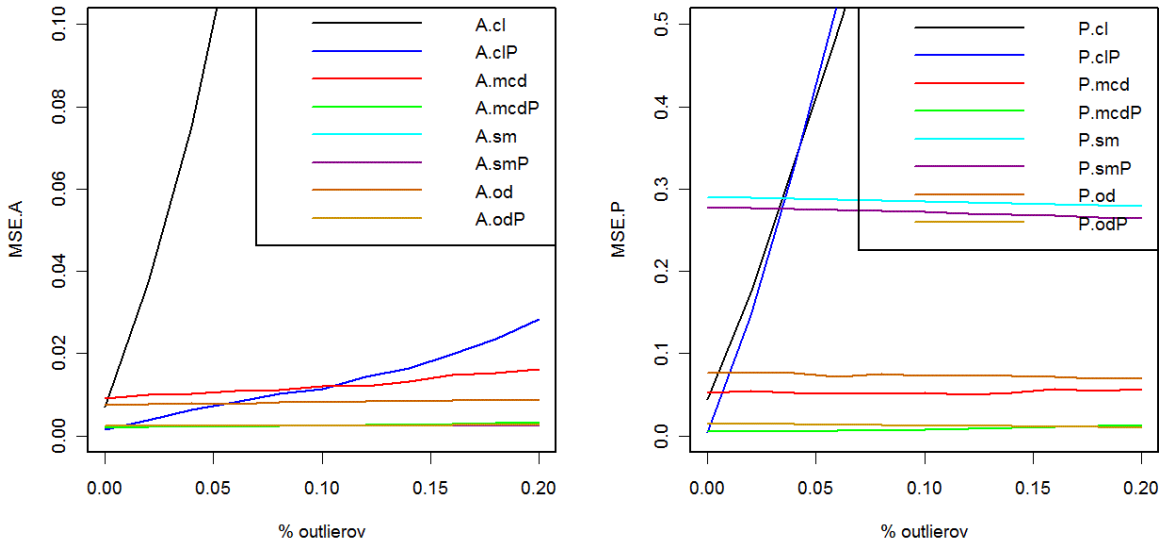
Vráťme sa k modelu (3.0.8) s maticou *Out*, kde jeden náhodne vybraný prvok pre každé z n_{out} náhodne zvolených meraní bude z rozdelenia $N(10, (0,05)^2)$. Pre *MCD*-odhad kovariančnej matice je dôležité správne určiť veľkosť podmnožiny dát, ktorá by už nemala mať outlierov. Doteraz sme odhadovali s hodnotou 3/4 a v simuláciach sme mali maximálne 20% outlierov. Grafy na obrázku 9, ku ktorým prislúcha tabuľka 11, zobrazujú krivky pre dáta až s 30% outlierov, pričom parametre pre výpočet *MCD*-odhadu zostali nezmenené, teda uvažujeme, že v dátach je najviac 25% zlých pozorovaní. Všetky krivky okrem červenej a zelenej v oboch grafoch v porovnaní s obrázkom 4 zostali približne rovnaké. Po prekročení hranice 25% sa výsledky faktorového modelu z *MCD*-odhadu kovariančnej matice významne zhoršili, ale na zvyšné dva robustné odhady to nemalo negatívny vplyv. Takže ak nepoznáme hornú hranicu podielu outlierov v dátach a odhadneme ho na menej ako je skutočnosť, tak výsledky sú oveľa horšie ako pri *OD* a *SM*-odhade.



Obr. 1: Graf pre výsledky z podkapitoly 3.1.1 na strane 26

% outlierov	A.cl	A.clP	A.mcd	A.mcdP	A.sm	A.smP	A.od	A.odP
0%	0,019	0,005	0,025	0,007	0,017	0,017	0,020	0,009
2%	0,053	0,009	0,025	0,007	0,017	0,017	0,020	0,009
4%	0,100	0,013	0,026	0,007	0,017	0,017	0,020	0,009
6%	0,161	0,016	0,029	0,008	0,017	0,017	0,020	0,009
8%	0,231	0,019	0,030	0,008	0,017	0,017	0,021	0,009
10%	0,316	0,022	0,032	0,009	0,017	0,017	0,021	0,009
12%	0,438	0,025	0,035	0,009	0,017	0,017	0,021	0,009
14%	0,587	0,029	0,036	0,010	0,017	0,017	0,021	0,009
16%	0,672	0,033	0,039	0,010	0,017	0,017	0,022	0,009
18%	0,966	0,038	0,041	0,010	0,017	0,017	0,022	0,009
20%	1,215	0,044	0,043	0,011	0,017	0,017	0,023	0,009
% outlierov	P.cl	P.clP	P.mcd	P.mcdP	P.sm	P.smP	P.od	P.odP
0%	0,077	0,008	0,087	0,010	0,480	0,462	0,116	0,025
2%	0,180	0,125	0,084	0,010	0,479	0,460	0,114	0,023
4%	0,308	0,274	0,081	0,011	0,477	0,458	0,114	0,023
6%	0,453	0,446	0,087	0,012	0,476	0,456	0,113	0,022
8%	0,592	0,627	0,085	0,013	0,473	0,454	0,109	0,021
10%	0,749	0,824	0,086	0,015	0,472	0,452	0,110	0,020
12%	0,908	1,039	0,090	0,017	0,470	0,449	0,106	0,019
14%	1,056	1,249	0,087	0,019	0,469	0,447	0,107	0,019
16%	1,249	1,470	0,091	0,021	0,466	0,444	0,106	0,018
18%	1,393	1,699	0,091	0,024	0,463	0,441	0,103	0,017
20%	1,528	1,917	0,090	0,027	0,460	0,438	0,102	0,016

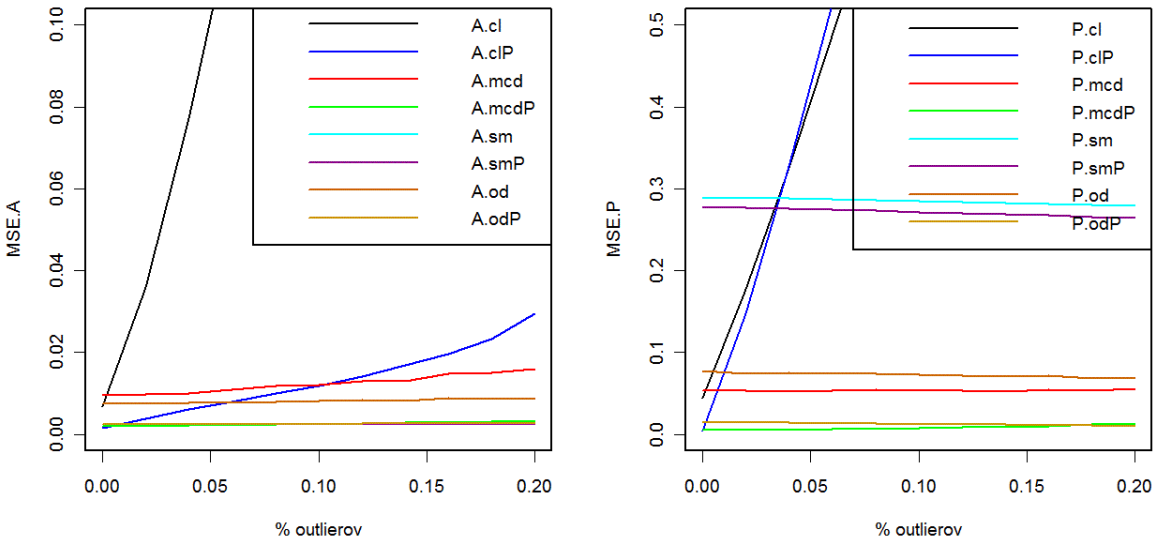
Tabuľka 3: Výsledky pre priemer MSE z podkapitoly 3.1.1



Obr. 2: Graf pre výsledky z podkapitoly 3.1.2 na strane 27 pre 1000 simulácií

% outlierov	A.cl	A.clP	A.mcd	A.mcdP	A.sm	A.smP	A.od	A.odP
0%	0,007	0,001	0,009	0,002	0,002	0,002	0,007	0,002
2%	0,037	0,004	0,010	0,002	0,002	0,002	0,008	0,002
4%	0,075	0,006	0,010	0,002	0,002	0,002	0,008	0,003
6%	0,125	0,008	0,011	0,002	0,002	0,002	0,008	0,002
8%	0,187	0,01	0,011	0,002	0,002	0,002	0,008	0,003
10%	0,291	0,011	0,012	0,002	0,002	0,002	0,008	0,002
12%	0,410	0,014	0,012	0,003	0,002	0,002	0,008	0,002
14%	0,574	0,016	0,013	0,003	0,002	0,002	0,008	0,003
16%	0,762	0,020	0,015	0,003	0,002	0,002	0,009	0,003
18%	1,017	0,024	0,015	0,003	0,002	0,002	0,009	0,003
20%	1,412	0,028	0,016	0,003	0,002	0,002	0,009	0,003
% outlierov	P.cl	P.clP	P.mcd	P.mcdP	P.sm	P.smP	P.od	P.odP
0%	0,045	0,004	0,052	0,005	0,289	0,278	0,077	0,015
2%	0,173	0,147	0,055	0,005	0,289	0,277	0,076	0,014
4%	0,331	0,326	0,052	0,005	0,288	0,276	0,077	0,015
6%	0,490	0,526	0,051	0,006	0,287	0,274	0,072	0,013
8%	0,658	0,741	0,051	0,006	0,287	0,273	0,075	0,013
10%	0,812	0,965	0,052	0,007	0,285	0,272	0,073	0,012
12%	0,987	1,205	0,049	0,008	0,283	0,270	0,072	0,012
14%	1,141	1,449	0,052	0,009	0,283	0,269	0,073	0,012
16%	1,294	1,690	0,056	0,011	0,282	0,267	0,071	0,011
18%	1,442	1,942	0,055	0,012	0,28	0,266	0,07	0,010
20%	1,531	2,193	0,057	0,014	0,279	0,264	0,071	0,010

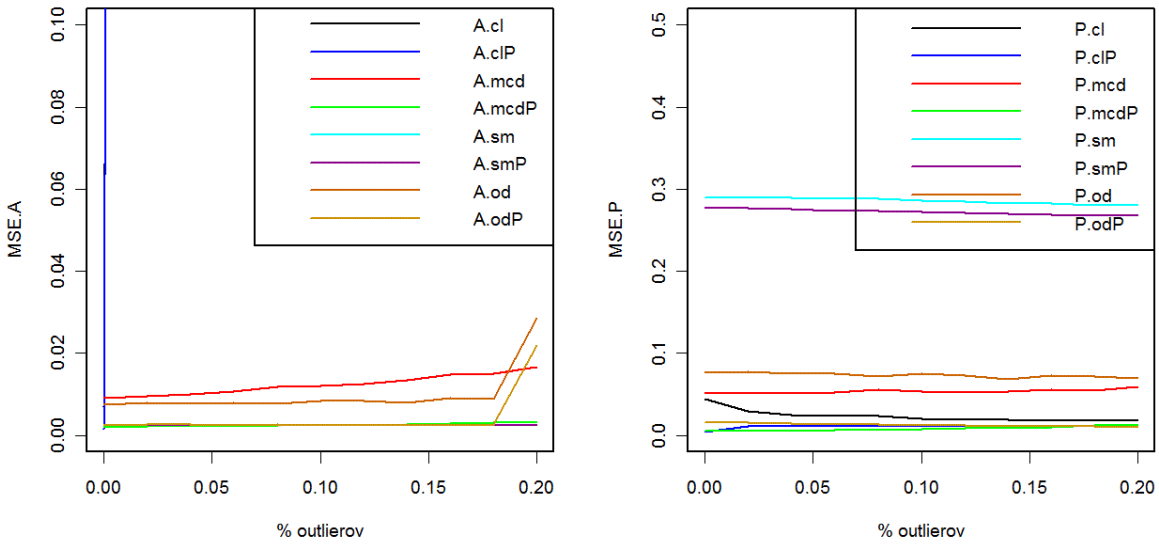
Tabuľka 4: Výsledky pre priemer MSE z podkapitoly 3.1.2 pre 1000 simulácií



Obr. 3: Graf pre výsledky z podkapitoly 3.1.2 na strane 27 pre 2000 simulácií

% outlierov	A.cl	A.clP	A.mcd	A.mcdP	A.sm	A.smP	A.od	A.odP
0%	0,007	0,001	0,009	0,002	0,002	0,002	0,007	0,002
2%	0,036	0,004	0,010	0,002	0,002	0,002	0,007	0,002
4%	0,077	0,006	0,010	0,002	0,002	0,002	0,008	0,003
6%	0,126	0,008	0,011	0,002	0,002	0,002	0,008	0,003
8%	0,202	0,010	0,012	0,002	0,002	0,002	0,008	0,003
10%	0,282	0,012	0,012	0,002	0,002	0,002	0,008	0,003
12%	0,416	0,014	0,013	0,003	0,002	0,002	0,008	0,003
14%	0,579	0,017	0,013	0,003	0,002	0,002	0,008	0,003
16%	0,768	0,020	0,015	0,003	0,002	0,002	0,009	0,003
18%	1,006	0,023	0,015	0,003	0,002	0,002	0,009	0,003
20%	1,411	0,029	0,016	0,003	0,002	0,002	0,009	0,003
% outlierov	P.cl	P.clP	P.mcd	P.mcdP	P.sm	P.smP	P.od	P.odP
0%	0,044	0,004	0,053	0,005	0,29	0,278	0,077	0,015
2%	0,177	0,147	0,053	0,005	0,289	0,277	0,074	0,015
4%	0,325	0,327	0,052	0,005	0,288	0,276	0,075	0,014
6%	0,487	0,524	0,054	0,006	0,287	0,274	0,074	0,014
8%	0,647	0,741	0,054	0,007	0,286	0,273	0,074	0,013
10%	0,82	0,968	0,053	0,007	0,285	0,272	0,073	0,012
12%	0,982	1,207	0,053	0,008	0,284	0,270	0,072	0,012
14%	1,134	1,443	0,052	0,009	0,283	0,269	0,070	0,012
16%	1,293	1,695	0,055	0,01	0,282	0,267	0,072	0,011
18%	1,435	1,938	0,054	0,012	0,281	0,266	0,069	0,010
20%	1,528	2,187	0,055	0,013	0,279	0,264	0,068	0,010

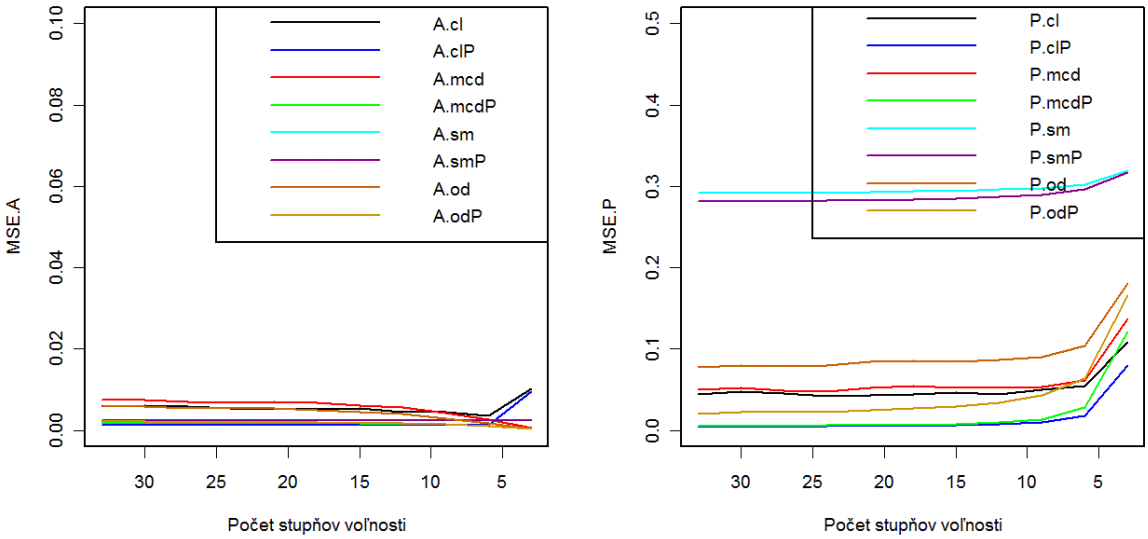
Tabuľka 5: Výsledky pre priemer MSE z podkapitoly 3.1.2 pre 2000 simulácií



Obr. 4: Graf pre výsledky z podkapitoly 3.1.3 na strane 28

% outlierov	A.cl	A.clP	A.mcd	A.mcdP	A.sm	A.smP	A.od	A.odP
0%	0,007	0,001	0,009	0,002	0,002	0,002	0,007	0,002
2%	3,954	3,849	0,01	0,002	0,002	0,002	0,008	0,003
4%	15,088	14,877	0,01	0,002	0,002	0,002	0,008	0,003
6%	32,559	32,267	0,011	0,002	0,002	0,002	0,008	0,003
8%	55,454	55,062	0,012	0,002	0,002	0,002	0,008	0,003
10%	82,768	82,313	0,012	0,002	0,002	0,002	0,008	0,003
12%	113,868	113,316	0,012	0,003	0,002	0,002	0,008	0,003
14%	148,074	147,46	0,013	0,003	0,002	0,002	0,008	0,002
16%	184,326	183,663	0,015	0,003	0,002	0,002	0,009	0,003
18%	222,154	221,401	0,015	0,003	0,002	0,002	0,009	0,003
20%	261,402	260,627	0,017	0,003	0,002	0,002	0,029	0,022
% outlierov	P.cl	P.clP	P.mcd	P.mcdP	P.sm	P.smP	P.od	P.odP
0%	0,045	0,004	0,052	0,005	0,289	0,278	0,077	0,015
2%	0,030	0,011	0,051	0,005	0,29	0,277	0,077	0,015
4%	0,025	0,011	0,051	0,005	0,29	0,275	0,075	0,014
6%	0,024	0,011	0,052	0,006	0,289	0,274	0,075	0,014
8%	0,023	0,011	0,055	0,006	0,288	0,273	0,072	0,013
10%	0,021	0,011	0,054	0,007	0,286	0,272	0,075	0,013
12%	0,020	0,011	0,052	0,008	0,285	0,271	0,073	0,012
14%	0,019	0,011	0,054	0,009	0,283	0,27	0,068	0,011
16%	0,018	0,011	0,055	0,010	0,282	0,269	0,073	0,011
18%	0,018	0,011	0,055	0,012	0,281	0,268	0,071	0,011
20%	0,017	0,011	0,059	0,013	0,280	0,267	0,069	0,010

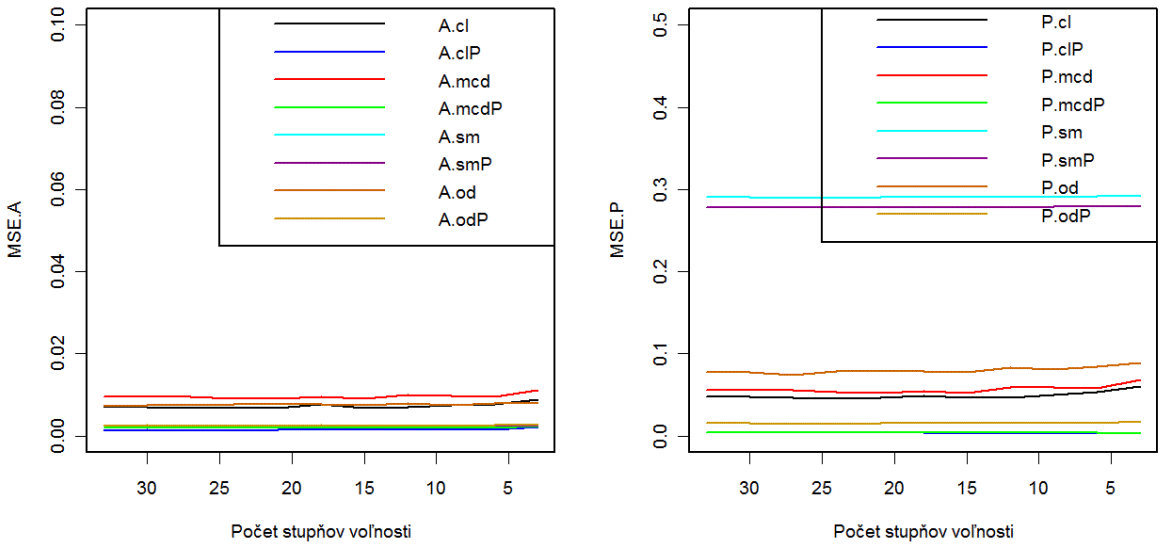
Tabuľka 6: Výsledky pre priemer MSE z podkapitoly 3.1.3



Obr. 5: Graf pre výsledky z podkapitoly 3.1.4 na strane 28

Stupne voľnosti	A.cl	A.clP	A.mcd	A.mcdP	A.sm	A.smP	A.od	A.odP
33	0,006	0,001	0,007	0,002	0,002	0,002	0,006	0,002
30	0,006	0,001	0,008	0,002	0,002	0,002	0,006	0,002
27	0,006	0,001	0,007	0,002	0,002	0,002	0,005	0,002
24	0,005	0,001	0,007	0,002	0,002	0,002	0,005	0,002
21	0,005	0,001	0,007	0,002	0,002	0,002	0,005	0,002
18	0,005	0,001	0,007	0,002	0,002	0,002	0,005	0,002
15	0,005	0,001	0,006	0,002	0,002	0,002	0,004	0,002
12	0,005	0,001	0,006	0,002	0,002	0,002	0,004	0,002
9	0,004	0,001	0,004	0,001	0,002	0,002	0,003	0,001
6	0,004	0,001	0,003	0,001	0,002	0,002	0,002	0,001
3	0,010	0,009	0,001	0,000	0,002	0,002	0,000	0,000
Stupne voľnosti	P.cl	P.clP	P.mcd	P.mcdP	P.sm	P.smP	P.od	P.odP
33	0,044	0,004	0,050	0,005	0,292	0,281	0,077	0,021
30	0,047	0,005	0,053	0,006	0,292	0,282	0,08	0,022
27	0,045	0,004	0,049	0,006	0,292	0,282	0,078	0,022
24	0,042	0,005	0,047	0,006	0,292	0,282	0,079	0,023
21	0,043	0,005	0,052	0,006	0,293	0,283	0,084	0,025
18	0,044	0,005	0,055	0,007	0,294	0,284	0,085	0,027
15	0,046	0,006	0,052	0,008	0,295	0,285	0,084	0,030
12	0,044	0,007	0,053	0,010	0,296	0,287	0,086	0,034
9	0,050	0,010	0,053	0,014	0,298	0,29	0,091	0,043
6	0,055	0,018	0,061	0,028	0,303	0,296	0,104	0,064
3	0,108	0,079	0,137	0,121	0,319	0,317	0,180	0,165

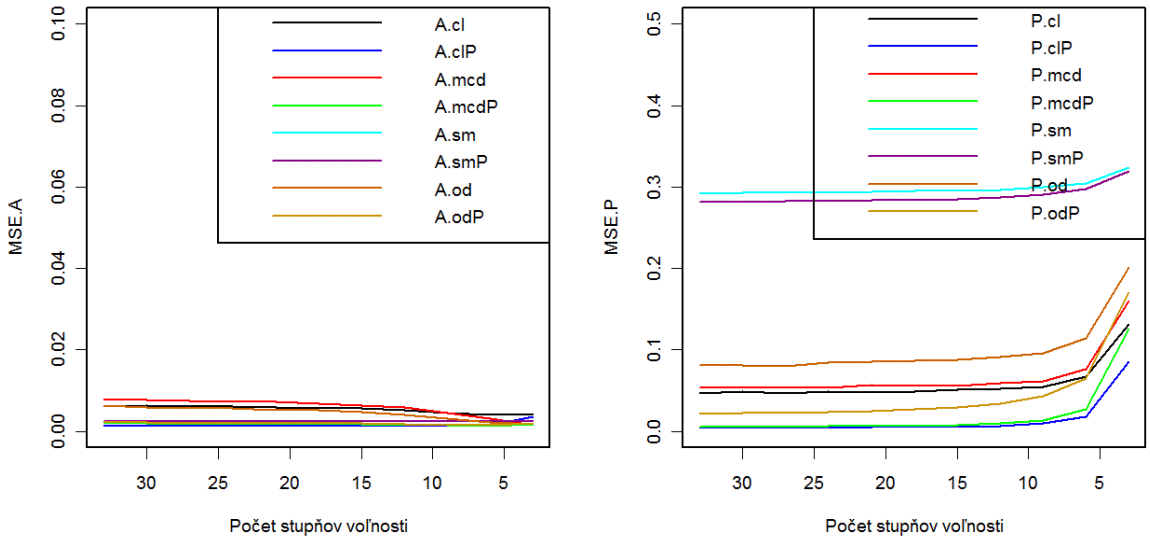
Tabuľka 7: Výsledky pre priemer MSE z podkapitoly 3.1.4



Obr. 6: Graf pre výsledky z podkapitoly 3.1.5 na strane 30

Stupne voľnosti	A.cl	A.clP	A.mcd	A.mcdP	A.sm	A.smP	A.od	A.odP
33	0,007	0,001	0,010	0,002	0,002	0,002	0,007	0,002
30	0,007	0,001	0,010	0,002	0,002	0,002	0,007	0,003
27	0,007	0,001	0,010	0,002	0,002	0,002	0,007	0,002
24	0,007	0,001	0,009	0,002	0,002	0,002	0,008	0,002
21	0,007	0,001	0,009	0,002	0,002	0,002	0,008	0,002
18	0,008	0,001	0,009	0,002	0,002	0,002	0,008	0,003
15	0,007	0,001	0,009	0,002	0,002	0,002	0,007	0,003
12	0,007	0,001	0,010	0,002	0,002	0,002	0,008	0,003
9	0,007	0,001	0,010	0,002	0,002	0,002	0,007	0,002
6	0,008	0,002	0,01	0,002	0,002	0,002	0,008	0,003
3	0,009	0,002	0,011	0,003	0,002	0,002	0,008	0,003
Stupne voľnosti	P.cl	P.clP	P.mcd	P.mcdP	P.sm	P.smP	P.od	P.odP
33	0,047	0,004	0,056	0,005	0,290	0,278	0,077	0,015
30	0,048	0,004	0,056	0,004	0,290	0,278	0,077	0,015
27	0,046	0,004	0,056	0,004	0,290	0,278	0,074	0,014
24	0,045	0,004	0,054	0,004	0,290	0,278	0,078	0,015
21	0,046	0,004	0,053	0,004	0,291	0,278	0,078	0,015
18	0,049	0,004	0,055	0,004	0,291	0,278	0,079	0,015
15	0,046	0,003	0,052	0,004	0,290	0,278	0,077	0,015
12	0,047	0,003	0,059	0,004	0,290	0,278	0,083	0,015
9	0,050	0,003	0,059	0,004	0,291	0,279	0,08	0,016
6	0,054	0,003	0,058	0,004	0,292	0,279	0,085	0,016
3	0,060	0,003	0,069	0,003	0,293	0,280	0,089	0,017

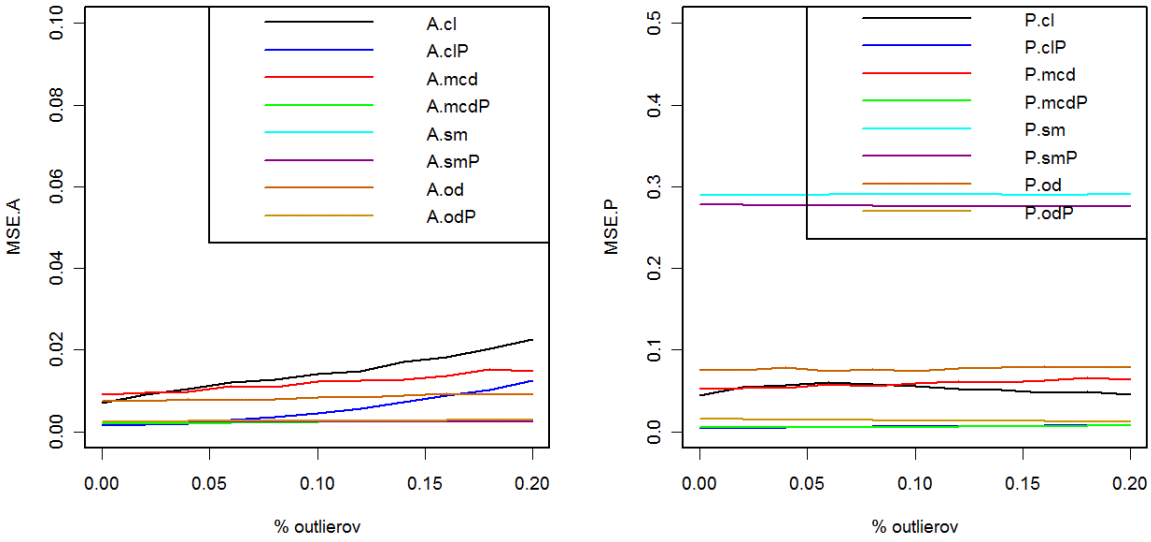
Tabuľka 8: Výsledky pre priemer MSE z podkapitoly 3.1.5



Obr. 7: Graf pre výsledky z podkapitoly 3.1.6 na strane 30

Stupne voľnosti	A.cl	A.clP	A.mcd	A.mcdP	A.sm	A.smP	A.od	A.odP
33	0,006	0,001	0,008	0,002	0,002	0,002	0,006	0,002
30	0,006	0,001	0,008	0,002	0,002	0,002	0,006	0,002
27	0,006	0,001	0,007	0,002	0,002	0,002	0,006	0,002
24	0,006	0,001	0,007	0,002	0,002	0,002	0,006	0,002
21	0,006	0,001	0,007	0,002	0,002	0,002	0,005	0,002
18	0,006	0,001	0,007	0,002	0,002	0,002	0,005	0,002
15	0,006	0,001	0,006	0,002	0,002	0,002	0,005	0,002
12	0,005	0,001	0,006	0,002	0,002	0,002	0,004	0,002
9	0,004	0,001	0,005	0,001	0,002	0,002	0,003	0,002
6	0,004	0,001	0,003	0,001	0,002	0,002	0,002	0,001
3	0,004	0,003	0,002	0,002	0,002	0,002	0,002	0,002
Stupne voľnosti	P.cl	P.clP	P.mcd	P.mcdP	P.sm	P.smP	P.od	P.odP
33	0,046	0,004	0,053	0,005	0,292	0,281	0,081	0,021
30	0,048	0,004	0,053	0,005	0,293	0,282	0,08	0,022
27	0,047	0,004	0,053	0,005	0,293	0,282	0,079	0,022
24	0,049	0,004	0,054	0,006	0,293	0,283	0,084	0,024
21	0,047	0,005	0,056	0,006	0,294	0,283	0,085	0,025
18	0,049	0,005	0,056	0,007	0,295	0,284	0,087	0,027
15	0,051	0,006	0,055	0,007	0,296	0,285	0,087	0,029
12	0,052	0,007	0,059	0,009	0,297	0,287	0,092	0,034
9	0,055	0,010	0,062	0,014	0,300	0,290	0,096	0,043
6	0,067	0,018	0,076	0,027	0,305	0,297	0,115	0,065
3	0,131	0,086	0,160	0,125	0,324	0,320	0,201	0,170

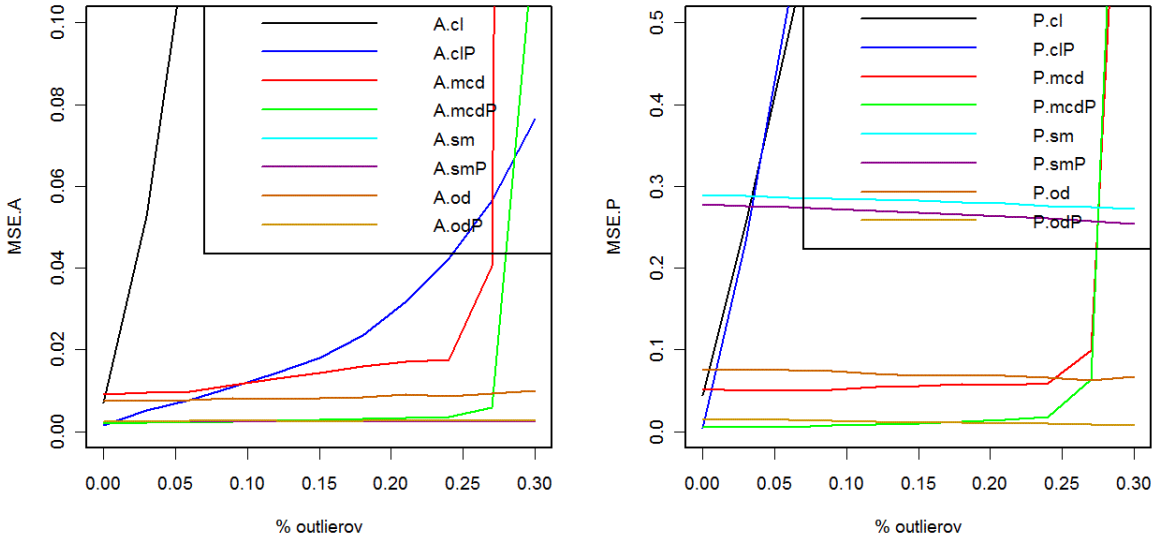
Tabuľka 9: Výsledky pre priemer MSE z podkapitoly 3.1.6



Obr. 8: Graf pre výsledky z podkapitoly 3.1.7 na strane 30

% outlierov	A.cl	A.clP	A.mcd	A.mcdP	A.sm	A.smP	A.od	A.odP
0	0,007	0,001	0,009	0,002	0,002	0,002	0,007	0,002
2	0,009	0,002	0,010	0,002	0,002	0,002	0,007	0,003
4	0,01	0,002	0,010	0,002	0,002	0,002	0,008	0,003
6	0,012	0,003	0,011	0,002	0,002	0,002	0,008	0,003
8	0,013	0,004	0,011	0,002	0,002	0,002	0,008	0,003
10	0,014	0,005	0,012	0,002	0,002	0,002	0,008	0,003
12	0,015	0,006	0,013	0,002	0,002	0,002	0,008	0,003
14	0,017	0,007	0,013	0,003	0,002	0,002	0,009	0,003
16	0,018	0,009	0,014	0,003	0,002	0,002	0,009	0,003
18	0,020	0,010	0,015	0,003	0,002	0,002	0,009	0,003
20	0,023	0,012	0,015	0,003	0,002	0,002	0,009	0,003
% outlierov	P.cl	P.clP	P.mcd	P.mcdP	P.sm	P.smP	P.od	P.odP
0	0,045	0,004	0,052	0,005	0,289	0,278	0,077	0,015
2	0,054	0,004	0,053	0,005	0,29	0,278	0,075	0,015
4	0,057	0,005	0,053	0,005	0,29	0,277	0,078	0,015
6	0,06	0,006	0,058	0,005	0,29	0,277	0,074	0,014
8	0,058	0,006	0,055	0,006	0,291	0,277	0,076	0,014
10	0,056	0,006	0,059	0,006	0,29	0,276	0,074	0,014
12	0,052	0,007	0,062	0,006	0,291	0,276	0,077	0,014
14	0,051	0,007	0,061	0,006	0,291	0,276	0,078	0,013
16	0,048	0,007	0,062	0,007	0,29	0,276	0,08	0,013
18	0,049	0,007	0,066	0,007	0,29	0,275	0,078	0,013
20	0,045	0,007	0,064	0,008	0,291	0,275	0,078	0,012

Tabuľka 10: Výsledky pre priemer MSE z podkapitoly 3.1.7



Obr. 9: Graf pre výsledky z podkapitoly 3.1.8 na strane 31

% outlierov	A.cl	A.clP	A.mcd	A.mcdP	A.sm	A.smP	A.od	A.odP
0	0,007	0,001	0,009	0,002	0,002	0,002	0,007	0,002
3	0,053	0,005	0,01	0,002	0,002	0,002	0,008	0,002
6	0,126	0,008	0,01	0,002	0,002	0,002	0,008	0,003
9	0,252	0,011	0,011	0,002	0,002	0,002	0,008	0,003
12	0,411	0,014	0,013	0,003	0,002	0,002	0,008	0,003
15	0,645	0,018	0,014	0,003	0,002	0,002	0,008	0,002
18	1,024	0,023	0,016	0,003	0,002	0,002	0,008	0,003
21	1,548	0,032	0,017	0,003	0,002	0,002	0,009	0,003
24	2,153	0,042	0,018	0,004	0,002	0,002	0,009	0,003
27	2,991	0,056	0,04	0,006	0,002	0,002	0,009	0,003
30	3,811	0,077	1,105	0,123	0,002	0,002	0,01	0,003
% outlierov	P.cl	P.clP	P.mcd	P.mcdP	P.sm	P.smP	P.od	P.odP
0	0,045	0,004	0,052	0,005	0,289	0,278	0,077	0,015
3	0,254	0,233	0,05	0,005	0,288	0,276	0,075	0,015
6	0,486	0,525	0,05	0,006	0,287	0,275	0,075	0,014
9	0,734	0,855	0,051	0,007	0,286	0,273	0,074	0,013
12	0,985	1,204	0,054	0,008	0,284	0,27	0,071	0,012
15	1,229	1,573	0,056	0,01	0,282	0,268	0,068	0,011
18	1,425	1,941	0,057	0,012	0,281	0,266	0,068	0,011
21	1,605	2,317	0,057	0,014	0,279	0,263	0,068	0,01
24	1,754	2,684	0,059	0,018	0,276	0,261	0,066	0,009
27	1,892	3,053	0,099	0,063	0,274	0,257	0,063	0,008
30	2,051	3,415	1,109	1,256	0,272	0,254	0,067	0,008

Tabuľka 11: Výsledky pre priemer MSE z podkapitoly 3.1.8

Záver

Pri vytváraní a odhadovaní rôznych matematických modelov alebo pri testovaní hypotéz v štatistike je dôležité, aby dáta, s ktorými pracujeme spĺňali predpoklady a neobsahovali outlierov. Ak to tak nie je, mali by sme používať také metódy, ktoré dokážu minimalizovať vplyv outlierov na výsledky. Keďže je vo faktorovej analýze veľmi dôležitý odhad kovariančnej matice, zamerali sme sa na robustné metódy, pomocou ktorých získame kvalitný odhad rozkladu kovariančnej matice aj z dát, ktoré obsahujú outlierov. Prvá najznámejšia metóda (MCD), ktorá je založená na minimálnom determinante kovariančnej matice, potrebuje ako vstup veľkosť podmnožiny, z ktorej sa bude počítať odhad. Metóda využívajúca normálne rozdelenie priestorového mediánu sa iteračne snaží získať čo najlepší odhad z celej množiny dát. Posledná metóda odhadu (OD) sa takisto snaží nájsť podmnožinu dát, ktoré budú očistené od outlierov. Od MCD-odhadu sa líši tým, že vstupom pre túto metódu je iba množina pôvodných dát. Počet outlierov si iteračne tento algoritmus určí sám a následne vytvorí podmnožinu dát, z ktorej počítame odhad kovariančnej matice.

Výsledky testovania metód uvádzame v tretej kapitole. Zvolili sme si viacero spôsobov generovania outlierov, ako aj porušenie predpokladu normality. Druhé a štvrté výsledky (grafy na obrázku 2 a 4) znázorňujú podobný typ vytvárania outlierov. Najlepšie, čiže najnižšie hodnoty stredných kvadratických chýb (MSE), ktoré krivky znázorňujú, sú pre metódu hlavných faktorov z MCD a OD-odhadu kovariančnej matice. Keď sme spoločné a špecifické faktory simulovali zo studentovho rozdelenia, ktoré má ťažké chvosty, všetky spôsoby odhadu faktorového modelu dávali približne rovnaké výsledky. Pre jednotlivé odhady kovariančnej matice je vždy lepší odhad metódou hlavných faktorov, čo sme aj očakávali, keďže predpokladom metódy maximálnej vierohodnosti je normalita dát. V jednom prípade sme pre niektoré pozorovania posunuli maticu faktorových nákladov o $0,5$, čím sme sa iným spôsobom snažili vytvoriť outlierov. Tieto výsledky ukazujú, že na odhad špecifických variancií to nemalo výrazný vplyv. Odhad faktorových nákladov je najmenej presný pre klasický odhad kovariančnej matice. V posledných výsledkoch (obrázok 9) sme testovali aký vplyv má správne zadanie veľkosti h-podmnožiny, ktorá by už nemala obsahovať outlierov. V

predchádzajúcich výpočtoch sa veľkosť tejto podmnožiny rovná 3/4 a v dátach bolo 20% outlierov, ale v tomto prípade generujeme až 30% outlierov. Z výsledkov vidíme, že hodnoty *MSE* pre *MCD*-odhad kovariančnej matice po prekročení 25% outlierov prudko stúpajú, pričom ostatné krivky zachovávajú svoj trend.

Cieľom práce bolo použiť nové postupy a zlepšiť odhad faktorovej analýzy. Ukázali sme, že nami použité robustné metódy na odhad faktorovej analýzy dávajú v niektorých prípadoch oveľa lepšie výsledky. Dôležité je, aby sme poznali dáta, s ktorými pracujeme a vedeli správne určiť, aké metódy je najlepšie použiť.

Literatúra

[1] Richard A. Johnson a Dean W. Wichern. 6.vyd. Applied Multivariate Statistical Analysis. New Jersey: Pearson Prentice Hall, ©2007. ISBN 0-13-187715-1.

[2] Greet Pison, et al. (2001). Robust Factor Analysis.

[3] Peter J. Rousseeuw a Katrien Van Driessen. A Fast Algorithm for the Minimum Covariance Determinant Estimator (1999). *Technometrics* **41** s. 212-223.

[4] Jyrki Mottonen, Klaus Nordhausen a Hannu Oja. Asymptotic theory of the spatial median (2010). *Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis: A Festschrift in honor of Professor Jana Jurečková* **7** s. 182-193.

[5] Daniel Pena a Francisco J. Prieto. (2001). Multivariate Outlier Detection and Robust Covariance Matrix Estimation. *Technometrix* **43** s. 286-310.

[6] James L. Buchanan a Peter R. Turnen. Numerical Methods and Analysis. McGraw-Hill, ©1992. ISBN 0-07-112922-7.

[7] Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.