



**MATEMATICKO-FYZIKÁLNÍ
FAKULTA**
Univerzita Karlova

DIPLOMOVÁ PRÁCE

Bc. Anna Zavadilová

Hlavní komponenty

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: doc. RNDr. Zdeněk Hlávka, Ph.D.

Studijní program: Matematika

Studijní obor: Pravděpodobnost, matematická statistika
a ekonometrie

Praha 2018

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

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

V dne

Podpis autora

Na tomto místě bych ráda poděkovala panu doc. RNDr. Zdeňku Hlávkovi, Ph.D. za cenné připomínky a podněty, trpělivost, ochotu a vstřícné jednání při vedení této diplomové práce. Poděkování rovněž patří rodičům za výraznou podporu během mého dosavadního studia. Také velmi děkuji svému příteli, který mi byl velkou oporou při psaní této práce.

Hlavní komponenty

Bc. Anna Zavadilová

Katedra pravděpodobnosti a matematické statistiky

doc. RNDr. Zdeněk Hlávka, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Práce představuje hlavní komponenty jako užitečný nástroj pro snížení dimenze datového souboru. V první části jsou uvedeny teoretické vlastnosti hlavních komponent a je zde odvozena konstrukce biplotu. Dále jsou shrnuty heuristické procedury pro volbu optimálního počtu hlavních komponent. Následně jsou uvedeny asymptotické vlastnosti výběrových vlastních čísel kovarianční a bílé Wishartovy matice, rozliší se případy rovnosti některých vlastních čísel. Ve druhé části je podrobně popsáno asymptotické rozdělení největšího vlastního čísla bílé Wishartovy matice doplněné o grafické ilustrace. Na základě tohoto asymptotického rozdělení odvodíme test počtu signifikantních vlastních čísel a představíme souvislost testu s volbou vhodného počtu hlavních komponent. V závěrečné části práce shrneme pokročilé výpočetní metody pro volbu počtu hlavních komponent. Práce je doplněna grafickými ilustracemi a simulační studií v softwarech Wolfram *Mathematica* a R.

Hlavní komponenty, Tracyho-Widomovo rozdělení, výběrová vlastní čísla

Principal Components

Bc. Anna Zavadilová

Department of Probability and Mathematical Statistics

doc. RNDr. Zdeněk Hlávka, Ph.D., Department of Probability and Mathematical Statistics

This thesis presents principal components as a useful tool for data dimensionality reduction. In the first part, the basic terminology and theoretical properties of principal components are described and a biplot construction is derived there as well. Besides, heuristic methods for a choice of the optimum number of principal components are summarised there. Subsequently, asymptotical properties of sample eigenvalues of covariance and white Wishart matrices are described and cases of equality of some eigenvalues are distinguished at the same time. In the second part of the thesis, asymptotic distribution of the largest eigenvalue of white Wishart matrices is described, completed with graphic illustrations. A test of the number of significant eigenvalues is suggested on the basis of this limiting distribution, and the connection of this test to the number of suitable principal components is presented. The final part of the thesis provides an overview of advanced computational methods for the choice of an adequate number of principal components. The thesis is completed with graphical illustrations and a simulation study using Wolfram *Mathematica* and R.

Principal components, sample eigenvalues, Tracy-Widom distribution

Obsah

Úvod	3
1 Základní pojmy	4
2 Hlavní komponenty – vlastnosti a grafická interpretace	8
2.1 Výběrové hlavní komponenty	11
2.2 Grafické znázornění	12
2.3 Volba počtu hlavních komponent – přehled literatury	17
2.4 Heuristická pravidla	17
2.4.1 Součtový podíl variability	18
2.4.2 Kaiserovo-Guttmanovo kritérium	18
2.4.3 Broken stick model	19
2.4.4 Porovnání rozdělení vlastních čísel a broken stick modelu .	22
2.4.5 Scree graf	23
2.4.6 LEV diagram	24
2.5 Shrnutí heuristických pravidel	24
3 Asymptotické vlastnosti výběrových vlastních čísel	27
3.1 Přesné rozdělení výběrových vlastních čísel	27
3.2 Asymptotické vlastnosti výběrových vlastních čísel za předpokladu jejich různosti	29
3.3 Tracyho-Widomovo rozdělení	30
3.3.1 Aproximace Tracyho-Widomova rozdělení	31
3.4 Rozdělení největšího vlastního čísla bílé Wishartovy matice	35
3.5 Asymptotické rozdělení nejmenšího vlastního čísla	40
4 Volba počtu hlavních komponent založená na Tracyho-Widomově rozdělení	43
4.1 Test přítomnosti právě jednoho nejednotkového vlastního čísla . .	43
4.2 Sekvenční verze testu	45
4.3 Shrnutí testů založených na Tracyho-Widomově rozdělení	46
4.4 Souvislost s modelem broken stick	47
5 Přehled dalších metod	49
5.1 Křížové ověřování	49
5.2 Bayesovský přístup	53
Závěr	55
Seznam použité literatury	57

A Přílohy	60
A.1 Zemědělská data	60
A.2 Standardizovaná zemědělská data	62
A.3 Charakteristiky použitých dat z balíčku SMSdata	63
A.4 Rovnost pro harmonická čísla	65
A.5 Sdružené a marginální hustoty vlastních čísel bílé Wishartovy matice	66

Úvod

Tato práce si klade za cíl seznámit čtenáře s jedním z významných prostředků pro analýzu mnohorozměrných dat, kterým je metoda hlavních komponent. Díky ní můžeme docílit snížení dimenze datového souboru, aniž bychom ztratili příliš mnoho informace. V první části práce si představíme teoretický základ, ze kterého pojem hlavních komponent vychází. Popíšeme rovněž grafický nástroj, kterým je biplot, jenž dovoluje zobrazovat vícerozměrná data vzhledem k prvním dvěma hlavním komponentám.

Následně se budeme věnovat metodám vedoucím k určení počtu hlavních komponent, které pak nahradí stávající zkoumané náhodné veličiny, jichž je obvykle příliš mnoho. Nejprve budou představeny „heuristické“ procedury hojně v praxi používané pro svou snadnou aplikovatelnost. Z těchto procedur zmíníme součtový podíl variability, Kaiserovo-Guttmanovo kritérium, broken stick model, scree graf a LEV diagram. Jelikož v této práci používáme pojem „optimální počet hlavních komponent“, popíšeme, v jakém smyslu tuto optimalitu chápeme.

Náplní třetí kapitoly je shrnutí asymptotických vlastností vlastních čísel výběrové kovarianční matice, přičemž se rozlišují dva případy – případ kdy, jsou vlastní čísla navzájem různá, a případ, kdy je tato vlastnost porušena. Dále popíšeme Tracyho-Widomovo rozdělení řádu 1, což je asymptotické rozdělení největšího vlastního čísla bílé Wishartovy matice, a toto rozdělení aproximujeme posunutým i zobecněným gama rozdělením. Odvodíme test počtu, v určitém smyslu, signifikantních vlastních čísel, který platí za specifických podmínek, a popíšeme jeho využití při volbě optimálního počtu hlavních komponent.

V závěrečné části diplomové práce se podíváme na metody sestávající z výpočetně složitějších technik, než byly ty, které vystupovaly v předchozích kapitolách. Mezi tyto výpočetně náročné procedury patří metody vycházející z křížového ověřování (cross validation, CV) a zejména v posledních letech rozvinuté metody založené na pravděpodobnostních modelech, jež používají bayesovský princip. Simulační studie a grafické ilustrace provádíme ve výpočetních softwarech Wolfram *Mathematica* (Wolfram Research, Inc., 2017) a R (R Core Team, 2017).

1. Základní pojmy

Představíme si základní matematické pojmy, se kterými se v průběhu práce budeme setkávat. Jedná se zejména o běžnou terminologii z oblasti statistiky a teorie matic. Během vytváření této kapitoly jsme zejména čerpali z knih Tebbens a kol. (2012) a Mardia a kol. (2003).

V souvislosti s hlavními komponentami se používá tzv. *singulární rozklad matice*. Využijeme jej v rámci kapitoly 2, kde najde své uplatnění při samotné konstrukci hlavních komponent, a také jej použijeme v rámci důkazů. Tento rozklad budeme ve většině případů aplikovat na datovou nebo kovarianční matici. V definici se objeví důležitá vlastnost některých matic, a tou je *ortonormalita*. Matici $\mathbb{A} \in \mathbb{R}^{n \times n}$ nazveme *ortonormální*, pokud $\mathbb{A}\mathbb{A}^\top = \mathbb{A}^\top\mathbb{A} = \mathbb{I}_{n \times n}$, kde pod symbolem \mathbb{I}_n rozumíme diagonální matici typu $n \times n$ se samými jednotkami na hlavní diagonále.

Věta 1. (*Singulární rozklad matice*) Pro každou matici $\mathbb{A} \in \mathbb{R}^{n \times m}$ hodnosti r existují ortonormální matice $\mathbb{U} \in \mathbb{R}^{n \times n}$ a $\mathbb{V} \in \mathbb{R}^{m \times m}$ a kladná čísla $l_1 \geq l_2 \geq \dots \geq l_r > 0$ tak, že platí

$$\mathbb{A} = \mathbb{U}\mathbb{L}\mathbb{V}^\top,$$

kde

$$\mathbb{L} = \begin{pmatrix} l_1 & 0 & \dots & 0 & \dots & 0 \\ 0 & l_2 & \dots & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & l_r & \dots & 0 \\ 0 & 0 & \dots & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & \dots & 0 \end{pmatrix}, \quad \mathbb{L} \in \mathbb{R}^{n \times m}.$$

Vektory $\mathbf{u}_j, j = 1, \dots, n$, tvořící sloupce matice \mathbb{U} se nazývají levými singulárními vektory, vektory $\mathbf{v}_j, j = 1, \dots, m$, tvořící sloupce matice \mathbb{V} nazýváme pravými singulárními vektory a l_1, l_2, \dots, l_r jsou singulární čísla.

Singulární čísla jsou rovněž odmocněná vlastní čísla matice $\mathbb{A}^\top\mathbb{A}$. Pokud je matice \mathbb{A} symetrická, pak $\mathbb{U} = \mathbb{V}$ a hovoříme o tzv. *spektrálním rozkladu*. Výše uvedený tvar singulárního rozkladu je vhodný spíše pro teoretické účely. V praktických situacích se používá tzv. *ekonomický tvar* singulárního rozkladu, jenž vznikne z výše uvedených matic odstraněním nulových bloků. Označíme-li $\mathbb{U}_r = (\mathbf{u}_1, \dots, \mathbf{u}_r) \in \mathbb{R}^{n \times r}$ a $\mathbb{V}_r = (\mathbf{v}_1, \dots, \mathbf{v}_r) \in \mathbb{R}^{m \times r}$ matice tvořené prvními r sloupci matic \mathbb{U} a \mathbb{V} a $\mathbb{L}_r = \text{diag}(l_1, l_2, \dots, l_r) \in \mathbb{R}^{r \times r}$, pak je možné singulární rozklad zapsat jako

$$\mathbb{A} = \mathbb{U}_r\mathbb{L}_r\mathbb{V}_r^\top.$$

Na základě ekonomického tvaru singulárního rozkladu pak můžeme prvek a_{ij} matice \mathbb{A} zapsat následovně:

$$a_{ij} = \sum_{k=1}^r u_{ik}l_kv_{jk}, \quad (1.1)$$

kde u_{ik}, v_{jk} jsou prvky na pozicích $(i, k), (j, k)$ matic \mathbb{U}, \mathbb{V} .

Nyní věnujme pozornost zápisu rozdělení náhodných veličin a vektorů. Zápisem $X \sim \mathcal{L}$, resp. $X \stackrel{as}{\sim} \mathcal{L}$ značíme, že X má přesné rozdělení \mathcal{L} , resp. že konverguje v distribuci k náhodné veličině mající rozdělení \mathcal{L} .

Uvažujme náhodný výběr $(\mathbf{X}_1, \dots, \mathbf{X}_n)$ vzniklý realizacemi p -rozměrného generického vektoru \mathbf{X} s existujícím vektorem středních hodnot $\boldsymbol{\mu} \in \mathbb{R}^p$ a kovarianční maticí $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$. Sestavme matici $\mathbb{X} \in \mathbb{R}^{n \times p}$, $\mathbb{X} = (\mathbf{X}_1, \dots, \mathbf{X}_n)^\top$. Pak kovarianční matici $\boldsymbol{\Sigma}$ generického vektoru odhadneme pomocí výběrové kovarianční matice

$$\mathbb{S} = \frac{1}{n} \mathbb{X}^\top \mathbb{X} - \bar{\mathbf{X}} \bar{\mathbf{X}}^\top,$$

kde $\bar{\mathbf{X}} = (\bar{X}_1, \dots, \bar{X}_p)^\top$ představuje vektor tvořený výběrovými průměry sloupců matice \mathbb{X} . Pokud navíc $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (viz definice 3), pak je \mathbb{S} maximálně věrohodným odhadem $\boldsymbol{\Sigma}$.

Užitečným nástrojem pro porovnávání vzdáleností vektorů pozorování \mathbf{x}_i a \mathbf{x}_j v prostoru je kromě klasické Euklidovské vzdálenosti definované jako

$$\|\mathbf{x}_i - \mathbf{x}_j\| = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^\top (\mathbf{x}_i - \mathbf{x}_j)}$$

také Mahalanobisova vzdálenost, což je obecnější měřítko dovolující porovnávat pozorování v systému souřadnic, jehož osy na sebe nejsou kolmé. Uvedme si její nejčastěji používanou výběrovou podobu.

Definice 1. (*Mahalanobisova vzdálenost – výběrová verze*) Uvažujme pozorování \mathbf{x}_i a \mathbf{x}_j příslušná datové matici \mathbb{X} s plnou hodnotí. Nechť \mathbb{S} značí příslušnou výběrovou kovarianční matici. Pak Mahalanobisovu vzdálenost mezi pozorováními \mathbf{x}_i a \mathbf{x}_j definujeme jako

$$\|\mathbf{x}_i - \mathbf{x}_j\|_M = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^\top \mathbb{S}^{-1} (\mathbf{x}_i - \mathbf{x}_j)}.$$

Poznamenejme, že inverze výběrové kovarianční matice vždy existuje, neboť se jedná o symetrickou a pozitivně definitní matici. Dalším užitečným maticovým nástrojem je Frobeniova norma.

Definice 2. (*Frobeniova norma*) Nechť $\mathbb{A} \in \mathbb{R}^{n \times m}$. Frobeniovu normu matice \mathbb{A} definujeme vztahem

$$\|\mathbb{A}\|_F = \sqrt{\sum_{i=1}^n \sum_{j=1}^m a_{ij}^2}.$$

Frobeniovu normu lze zapsat také jinými způsoby jako např. (viz Tebbens a kol., 2012, kap. 1, str. 22)

$$\|\mathbb{A}\|_F = \sqrt{\sum_{j=1}^m \|\mathbf{a}_j\|^2} = \sqrt{\text{tr}(\mathbb{A}^\top \mathbb{A})},$$

kde \mathbf{a}_j je j -tý sloupec matice \mathbb{A} a $\text{tr}(\mathbb{A}^\top \mathbb{A})$ označuje stopu matice $\mathbb{A}^\top \mathbb{A}$, tj. součet prvků na hlavní diagonále. Později použijeme druhou mocninu Frobeniovy normy, jež aplikujeme na rozdíly matic, a která říká, jak moc se matice od sebe odlišují.

Rovněž připomeňme definici několika významných rozdělení objevujících se v mnohorozměrné analýze. Konkrétně se bude jednat o mnohorozměrné normální rozdělení náhodného vektoru a Wishartovo rozdělení náhodné matice.

Definice 3. (*Mnohorozměrné normální rozdělení*) Řekneme, že náhodný vektor \mathbf{X} má p -rozměrné normální rozdělení s vektorem středních hodnot $\boldsymbol{\mu} \in \mathbb{R}^p$ a pozitivně definitní kovarianční maticí $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$, pokud je jeho hustota tvaru

$$f(\mathbf{x}) = |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp \left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \right\}, \quad \mathbf{x} \in \mathbb{R}^p.$$

Zapisujeme jako $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Poznamenejme, že předpoklad pozitivní definitnosti se vyžaduje kvůli existenci inverze matice $\boldsymbol{\Sigma}$. S Wishartovým rozdělením matice se setkáme v kapitole 3, kde se budeme věnovat asymptotickému rozdělení jejich vlastních čísel.

Definice 4. (*Wishartovo rozdělení*) Necht je matice $\mathbf{X} \in \mathbb{R}^{n \times p}$ sestavena jako náhodný výběr z rozdělení $\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$ uspořádaný do řádků a necht $\mathbf{M} = \mathbf{X}^\top \mathbf{X}$. Pak říkáme, že matice \mathbf{M} má Wishartovo rozdělení se škálovou maticí $\boldsymbol{\Sigma}$ a n stupni volnosti. Zapisujeme jako $\mathbf{M} \sim \mathcal{W}_p(\boldsymbol{\Sigma}, n)$ a rovněž \mathbf{M} nazýváme Wishartovou maticí. V případě, že $\boldsymbol{\Sigma} = \mathbb{I}_p$, hovoříme o tzv. bílé Wishartově matici.

Nyní popíšeme jednorozměrná rozdělení, která využijeme v kapitole 3 k aproximačním účelům. Konkrétně se bude jednat o klasické gama rozdělení, u něhož budeme navíc uvažovat parametr posunutí, a pak definujeme zobecněné gama rozdělení.

Definice 5. (*Posunuté gama rozdělení*) Řekneme, že náhodná veličina X má posunuté gama rozdělení s parametry tvaru $k > 0$, měřítka $\theta > 0$ a polohy $\mu > 0$ (zapisujeme $X \sim \mathbf{SG}\Gamma(k, \theta, \mu)$), pokud je její hustota dána předpisem

$$f(x) = \begin{cases} \frac{1}{\Gamma(k)\theta^k} (x - \mu)^{k-1} e^{-\frac{x-\mu}{\theta}}, & x > \mu, \\ 0, & x \leq \mu, \end{cases}$$

kde $\Gamma(k)$ značí hodnotu gama funkce v bodě k .

Distribuční funkce posunutého gama rozdělení je tvaru

$$F(x) = \begin{cases} \frac{1}{\Gamma(k)} \gamma\left(k, \frac{x-\mu}{\theta}\right), & x > \mu, \\ 0, & x \leq \mu, \end{cases}$$

kde uvedené symboly $\Gamma(k)$, resp. $\gamma\left(k, \frac{1}{\theta}\right)$ představují gama funkci, resp. spodní neúplnou gama funkci, které jsou určeny následujícími předpisy:

$$\begin{aligned} \Gamma(s) &= \int_0^\infty t^{s-1} e^{-t} dt, \quad s > 0, \\ \gamma(s, x) &= \int_0^x t^{s-1} e^{-t} dt, \quad s > 0, \quad x > 0. \end{aligned}$$

Střední hodnota a rozptyl pro $X \sim \mathbf{SG}\Gamma(k, \theta, \mu)$ mají vyjádření:

$$\begin{aligned} \mathbb{E} X &= k\theta + \mu, \\ \text{var} X &= k\theta^2. \end{aligned}$$

Případná volba $\mu = 0$ by vedla na klasické gama rozdělení. Definici 5 můžeme rozšířit na tzv. *zobecněné gama rozdělení*. Oproti běžné definici navíc zavedeme koeficient polohy, což rovněž odpovídá implementaci v softwaru *Mathematica*. Takto definované rozdělení pak použijeme v kapitole 3 k aproximačním účelům.

Definice 6. (*Zobecněné gama rozdělení*) Řekneme, že náhodná veličina X má zobecněné gama rozdělení s parametry tvaru $k > 0$, $\beta > 0$, měřítka $\theta > 0$ a polohy $\mu > 0$ (zapisujeme $X \sim \mathbf{SGT}(k, \theta, \beta, \mu)$), pokud má její hustota tvar

$$f(x) = \begin{cases} \frac{\beta}{\Gamma(k)\theta} \left(\frac{x-\mu}{\theta}\right)^{k\beta-1} e^{-\left(\frac{x-\mu}{\theta}\right)^\beta}, & x > \mu, \\ 0, & x \leq \mu. \end{cases}$$

Distribuční funkce zobecněného gama rozdělení je

$$F(x) = \begin{cases} \frac{1}{\Gamma(k)} \gamma\left(k, \left(\frac{x-\mu}{\theta}\right)^\beta\right), & x > \mu, \\ 0, & x \leq \mu. \end{cases}$$

Pozorujeme, že volbou $\beta = 1$, $\mu = 0$ opět získáme klasické gama rozdělení. Střední hodnota a rozptyl náhodné veličiny $X \sim \mathbf{SGT}(k, \theta, \beta, \mu)$ mají následující tvary:

$$\begin{aligned} \mathbb{E} X &= \theta \frac{\Gamma((k\beta + 1)/\beta)}{\Gamma(k)} + \mu, \\ \text{var } X &= \theta^2 \left(\frac{\Gamma((k\beta + 2)/\beta)}{\Gamma(k)} - \left(\frac{\Gamma((k\beta + 1)/\beta)}{\Gamma(k)} \right)^2 \right). \end{aligned}$$

2. Hlavní komponenty – vlastnosti a grafická interpretace

V této kapitole představíme hlavní komponenty, shrneme jejich základní vlastnosti a zadefinujeme pojmy s nimi spojené, které nás budou v rámci celé práce doprovázet. Rovněž budeme použití hlavních komponent ilustrovat na příkladu s reálnými daty. Cílem metody hlavních komponent je snížení dimenze datového souboru, aniž by došlo k velké ztrátě informace. Na data pohlížíme skrze hlavní komponenty představující lineární kombinace původních veličin, které jsou navíc normované a navzájem nekorelované.

Uvažujme p -rozměrný náhodný vektor $\mathbf{X} = (X_1, \dots, X_p)^\top$ s existující kovarianční maticí Σ a střední hodnotou $\boldsymbol{\mu} = (\mu_1, \dots, \mu_p)^\top$. Hlavní komponenta příslušná tomuto náhodnému vektoru se pak vytvoří pomocí normovaných lineárních kombinací

$$\mathbf{l}_i^\top \mathbf{X} = \sum_{j=1}^p l_{ij} X_j, \quad i = 1, \dots, p,$$

kteří zachycují maximální možnou míru *variability* obsaženou ve vektoru, tj. $\sum_{i=1}^p \text{var } X_i$, a jsou navzájem nekorelované. Formálnější popis nabízí následující definice.

Definice 7. *Nechť je \mathbf{X} p -rozměrný náhodný vektor. Lineární kombinaci $Y_1 = \mathbf{l}_1^\top \mathbf{X}$ nazveme první hlavní komponentou, pokud tato lineární kombinace maximalizuje rozptyl $\mathbf{l}_1^\top \mathbf{X}$ za podmínky $\mathbf{l}_1^\top \mathbf{l}_1 = 1$. Lineární kombinaci $Y_i = \mathbf{l}_i^\top \mathbf{X}$, $1 < i \leq p$, nazveme i -tou hlavní komponentou, pokud tato lineární kombinace maximalizuje rozptyl $\mathbf{l}_i^\top \mathbf{X}$ za podmínek $\mathbf{l}_i^\top \mathbf{l}_i = 1$ a $\text{cov}(\mathbf{l}_i^\top \mathbf{X}, \mathbf{l}_k^\top \mathbf{X}) = 0$ pro každé $k < i$.*

Nyní se podívejme na to, jakým způsobem kýžené lineární kombinace nalezneme. Podle následujícího tvrzení jsou hlavní komponenty nekorelované lineární kombinace původního náhodného vektoru s koeficienty příslušnými vlastním vektorům a jejich rozptyly odpovídají vlastním číslům kovarianční matice Σ tohoto vektoru.

Tvrzení 2. *Nechť Σ značí kovarianční matici náhodného vektoru \mathbf{X} , která má dvojice vlastní číslo – vlastní vektor $(\lambda_1, \mathbf{e}_1), \dots, (\lambda_p, \mathbf{e}_p)$, kde $\lambda_1 \geq \dots \geq \lambda_p > 0$. Pak i -tá hlavní komponenta je dána vztahem*

$$Y_i = \mathbf{e}_i^\top \mathbf{X} = \sum_{j=1}^p e_{ij} X_j, \quad i = 1, \dots, p.$$

Při této volbě platí

$$\begin{aligned} \text{var } Y_i &= \mathbf{e}_i^\top \Sigma \mathbf{e}_i = \lambda_i, \quad i = 1, \dots, p, \\ \text{cov}(Y_i, Y_j) &= \mathbf{e}_i^\top \Sigma \mathbf{e}_j = 0, \quad i \neq j. \end{aligned}$$

V případě, že si jsou některé λ_i rovny, pak volby odpovídajících vektorů \mathbf{e}_i , a tudíž ani Y_i , nejsou jednoznačné.

Důkaz. Viz Johnson a Wichern (1992, str. 358, Result 8.1.) □

Označíme-li vektor hlavních komponent $\mathbf{Y} = (Y_1, \dots, Y_p)^\top$ a Γ matici vlastních vektorů kovarianční matice Σ , můžeme výše uvedený předpis pro hlavní komponenty zapsat maticově jako $\mathbf{Y} = \Gamma^\top \mathbf{X}$. Rovněž se setkáváme s předpisem $\mathbf{Y} = \Gamma^\top (\mathbf{X} - \boldsymbol{\mu})$, na základě něhož pak mají hlavní komponenty nulovou střední hodnotu. Často dochází k situaci, kdy mají sledované statistické znaky, z nichž se náhodný vektor skládá, různé jednotky. Tato skutečnost pak působí potíže při tvorbě hlavních komponent, jež se při změně měřítka mohou podstatně odlišovat. Tento problém je možné vyřešit transformací prvků vektoru \mathbf{X} , kdy místo nich uvažujeme veličiny

$$Z_i = \frac{X_i - \mu_i}{\sqrt{\text{var } X_i}}, \quad i = 1, \dots, p. \quad (2.1)$$

Vstupem pro náhodné komponenty pak bude místo vektoru \mathbf{X} vektor $\mathbf{Z} = (Z_1, \dots, Z_p)^\top$ s nulovým vektorem středních hodnot a kovarianční maticí s jednotkami na hlavní diagonále. Postup, kdy od náhodné veličiny odčítáme výběrový průměr a dělíme výběrovou směrodatnou odchylkou jako ve vzorci (2.1), nazýváme *standardizace*.

Obecně předpokládáme, že vysoký podíl variability obsažené v náhodném vektoru \mathbf{X} bude zastoupen takovým počtem příslušných hlavních komponent m , který je výrazně nižší než dimenze náhodného vektoru p . Problémem vhodné volby počtu m se budeme v práci nadále zabývat. Dále poznamenejme, že celková variabilita, která je obsažená v hlavních komponentách, zůstává stejná jako variabilita náhodného vektoru \mathbf{X} . Blíže o tomto faktu pojednává následující tvrzení.

Tvrzení 3. *Nechť Σ značí kovarianční matici náhodného vektoru \mathbf{X} , která má dvojice vlastní číslo – vlastní vektor $(\lambda_1, \mathbf{e}_1), \dots, (\lambda_p, \mathbf{e}_p)$, kde $\lambda_1 \geq \dots \geq \lambda_p > 0$. Nechť $Y_1 = \mathbf{e}_1^\top \mathbf{X}, \dots, Y_p = \mathbf{e}_p^\top \mathbf{X}$ jsou hlavní komponenty. Pak*

$$\sum_{i=1}^p \text{var } X_i = \sum_{i=1}^p \text{var } Y_i = \sum_{i=1}^p \lambda_i.$$

Důkaz. Nejprve na kovarianční matici Σ aplikujeme spektrální rozklad popsaný v kapitole 1, tj. $\Sigma = \Gamma \Lambda \Gamma^\top$, kde Λ představuje diagonální matici s vlastními čísly $\lambda_1, \dots, \lambda_p$ na hlavní diagonále a Γ je ortonormální matice, jejíž sloupce jsou vlastní vektory $\mathbf{e}_1, \dots, \mathbf{e}_p$. Pak

$$\sum_{i=1}^p \text{var } X_i = \text{tr}(\Sigma) = \text{tr}(\Gamma \Lambda \Gamma^\top) = \text{tr}(\Lambda \Gamma^\top \Gamma) = \text{tr}(\Lambda) = \sum_{i=1}^p \lambda_i = \sum_{i=1}^p \text{var } Y_i. \quad \square$$

Jinými slovy předchozí tvrzení říká, že prvních $m = 1, \dots, p$ hlavních komponent vysvětluje

$$\frac{\sum_{i=1}^m \lambda_i}{\sum_{i=1}^p \lambda_i}$$

variability náhodného vektoru \mathbf{X} . Pokud je tento podíl např. 0,8 nebo 0,9, nahrazením původních p složek proměnných prvními m komponentami neztrácíme příliš mnoho informace obsažené ve zbývajících hlavních komponentách.

Jedním z významných prostředků pro zjišťování vztahu mezi veličinami je jejich korelace. Následující tvrzení pojednává o předpisu a zajímavé vlastnosti korelace mezi veličinami a hlavními komponentami.

Tvrzení 4. *Nechť \mathbf{X} je náhodný vektor s p -rozměrnou kovarianční maticí Σ a dvojicemi vlastní číslo – vlastní vektor $(\lambda_1, \mathbf{e}_1), \dots, (\lambda_p, \mathbf{e}_p)$, kde $\lambda_1 \geq \dots \geq \lambda_p > 0$. Příslušné hlavní komponenty mají tvar $(Y_1, \dots, Y_p)^\top = (\mathbf{e}_1^\top \mathbf{X}, \dots, \mathbf{e}_p^\top \mathbf{X})^\top$. Pak lze korelaci $\rho_{X_k Y_i}$ mezi k -tou veličinou X_k a i -tou hlavní komponentou Y_i vyjádřit vztahem*

$$\rho_{X_k Y_i} = e_{ki} \sqrt{\frac{\lambda_i}{\sigma_{kk}}},$$

kde e_{ki} představuje i -tý prvek vektoru \mathbf{e}_k a σ_{kk} rozptyl veličiny X_k . Dále také platí

$$\sum_{i=1}^p \rho_{X_k Y_i}^2 = 1.$$

Důkaz. Nejprve dokážeme první část tvrzení. Označme \mathbf{l}_k p -rozměrný vektor tvořený samými nulami a jedinou jednotkou na pozici k , což lze zapsat jako $\mathbf{l}_k = (0, \dots, 0, 1, 0, \dots, 0)^\top$. Pak můžeme k -tou veličinu vyjádřit jako $X_k = \mathbf{l}_k^\top \mathbf{X}$. Pro kovarianci mezi k -tou veličinou a i -tou hlavní komponentou platí vztah

$$\text{cov}(X_k, Y_i) = \text{cov}(\mathbf{l}_k^\top \mathbf{X}, \mathbf{e}_i^\top \mathbf{X}) = \mathbf{l}_k^\top \Sigma \mathbf{e}_i = \lambda_i e_{ki},$$

příčemž poslední rovnost vyplývá ze základního vztahu mezi vlastními čísly a vlastními vektory. Nyní již stačí dosadit do definice korelace pro dvojici X_k a Y_i :

$$\rho_{X_k Y_i} = \frac{\text{cov}(X_k, Y_i)}{\sqrt{\text{var } X_k \text{ var } Y_i}} = \frac{\lambda_i e_{ki}}{\sqrt{\sigma_{kk} \lambda_i}} = e_{ki} \sqrt{\frac{\lambda_i}{\sigma_{kk}}}.$$

Nyní přistupme k důkazu zbytku tvrzení. Použitím již dokázané první části dostáváme:

$$\sum_{i=1}^p \rho_{X_k Y_i}^2 = \frac{1}{\sigma_{kk}} \sum_{i=1}^p \lambda_i e_{ki}^2 = 1.$$

Poslední rovnost plyne z toho, že výraz $\sum_{i=1}^p \lambda_i e_{ki}^2$ odpovídá prvku na pozici (k, k) matice $\Gamma \Lambda \Gamma^\top$, kde $\Gamma = (\mathbf{e}_1, \dots, \mathbf{e}_p)$, a Λ označuje diagonální matici s prvky $\lambda_1, \dots, \lambda_p$ na hlavní diagonále. Uvedený součin tří matic je však zároveň spektrálním rozkladem kovarianční matice Σ , jejíž prvek na místě (k, k) je právě σ_{kk} . \square

Podobně jako výše můžeme říci, že podíl variability veličiny X_k vysvětlené komponentou Y_i je $\rho_{X_k Y_i}^2$ a podíl variability vysvětlené prvními m komponentami

$$\sum_{i=1}^m \rho_{X_k Y_i}^2 = \frac{1}{\sigma_{kk}} \sum_{i=1}^m \lambda_i e_{ki}^2.$$

Jmenovatel zlomku vyjadřuje variabilitu X_k , která má být vysvětlena, a čítec pak variabilitu X_k vysvětlenou prvními m komponentami. Variabilitu obsaženou v prvních m komponentách je rovněž možné vyjádřit jako součet podílů variability přes všechny veličiny přenásobených variabilitou každé z veličin, tj.

$$\sum_{i=1}^m \lambda_i = \sum_{k=1}^p \sigma_{kk} \sum_{i=1}^m \rho_{X_k Y_i}^2.$$

2.1 Výběrové hlavní komponenty

Nyní se budeme zabývat odvozením výběrových verzí hlavních komponent v případě, že máme k dispozici náhodný výběr $(\mathbf{X}_1, \dots, \mathbf{X}_n)$ tvořený p -rozměrnými náhodnými vektory, které umístíme do řádků datové matice \mathbb{X} typu $n \times p$. Příslušnou kovarianční matici Σ odhadneme pomocí výběrové kovarianční matice zmíněné již v rámci kapitoly 1, tj.

$$\mathbb{S} = \frac{1}{n} \mathbb{X}^\top \mathbb{X} - \bar{\mathbf{X}} \bar{\mathbf{X}}^\top, \quad (2.2)$$

s vektorem výběrových průměrů sloupců datové matice značeným jako $\bar{\mathbf{X}} = (\bar{X}_1, \dots, \bar{X}_p)^\top$. V případě, že jsou složky náhodných vektorů ve výběru uvedeny v různých jednotkách, je vhodnější pracovat s výběrovou korelační maticí \mathbb{R} . Ta je vytvořena jako výběrová kovarianční matice pro standardizované složky (2.1).

Odpovídající dvojice vlastní číslo – vlastní vektor matice \mathbb{S} (lze také použít matici \mathbb{R}) označme $(\hat{\lambda}_1, \hat{\mathbf{e}}_1), \dots, (\hat{\lambda}_p, \hat{\mathbf{e}}_p)$, kde $\hat{\lambda}_1 \geq \dots \geq \hat{\lambda}_p > 0$. Pak analogicky jako v nevýběrové části, j -tá výběrová hlavní komponenta příslušná náhodnému vektoru $\mathbf{X}_i, i = 1, \dots, n$, je dána vztahem

$$\hat{Y}_{ij} = \hat{\mathbf{e}}_{j_i}^\top \mathbf{X}_i, \quad j = 1, \dots, p.$$

Opět se jedná o takovou lineární kombinaci $\mathbf{l}_i^\top \mathbf{X}_i$, která vzhledem k podmínce $\mathbf{l}_i^\top \mathbf{l}_i = 1$ maximalizuje výběrový rozptyl $\mathbf{l}_i^\top \mathbb{S} \mathbf{l}_i$ a splňuje požadavek nekorelovanosti s každou z dosud vytvořených kombinací.

Označíme-li matici vlastních čísel výběrové kovarianční matice jako $\hat{\Gamma}$, můžeme výše uvedený předpis pro hlavní komponenty zapsat do maticové podoby jako

$$\hat{\mathbf{Y}} = \mathbb{X} \hat{\Gamma}, \text{ resp. jako } \hat{\mathbf{Y}} = (\mathbb{X} - \mathbf{1}_n \bar{\mathbf{X}}^\top) \hat{\Gamma},$$

přičemž symbol $\mathbf{1}_n$ představuje sloupcový vektor tvořený n jedničkami. Podobně jako v teoretickém případě platí, že

$$\text{var } \hat{Y}_i = \hat{\lambda}_i, \quad \text{cov}(\hat{Y}_i, \hat{Y}_j) = 0 \text{ pro } i, j = 1, \dots, p, \quad i \neq j.$$

Vysvětlená variabilita p výběrovými hlavními komponentami je pak rovna $\sum_{i=1}^p \hat{\lambda}_i$, což odpovídá součtu prvků na diagonále výběrové kovarianční matice, tj.

$$\sum_{i=1}^p s_{ii} = \sum_{i=1}^p \hat{\lambda}_i, \quad (2.3)$$

kde s_{ii} značí prvek (i, i) matice \mathbb{S} . Odhad korelace mezi veličinou X_k a hlavní komponentou Y_i je dán vztahem

$$r_{X_k Y_i} = \hat{e}_{ki} \sqrt{\frac{\hat{\lambda}_i}{s_{kk}}}$$

(viz Johnson a Wichern, 1992, str. 368).

2.2 Grafické znázornění

V literatuře je uvedena celá řada metod, jakými lze data pomocí hlavních komponent zobrazit. Stěžejní roli v grafických interpretacích hrají první dvě hlavní komponenty nesoucí největší míru informace. Naměřená data nebo jejich transformaci tedy můžeme promítnout do grafů, jejichž osy reprezentují první dvě hlavní komponenty, případně můžeme provést tzv. analýzu hlavních souřadnic. Pro grafickou ilustraci jsme zvolili tzv. biplot a se zbylými metodami se můžeme blíže seznámit v knize Jolliffe (2002, kap. 5).

Biplot nebo též „dvojný graf“ je grafický prostředek hojně používaný v mnohorozměrné analýze dat a existuje v několika verzích. My se budeme zabývat tou, která vychází z metody hlavních komponent. Jedná se o graf, na kterém jsou znázorněna pozorování a vztahy mezi statistickými znaky vzhledem k prvním dvěma hlavním komponentám. Data jsou tedy promítnuta do dvoudimenzionálního prostoru. Nejprve vyjdeme ze singulárního rozkladu datové matice, na jehož základě provedeme její aproximaci maticí hodnoty dvě, která je optimální z hlediska minimalizace Frobeniovy normy rozdílu původní a aproximující matice (viz definice 2 v kapitole 1).

Na základě teoretických poznatků popíšeme konkrétní způsob, jakým se biplot vytváří. Uvažujme datovou matici $\mathbb{X} \in \mathbb{R}^{n \times p}$ s hodnotami r . Ekonomický tvar singulárního rozkladu této matice značme

$$\mathbb{X} = \mathbf{U}\mathbf{L}\mathbf{V}^\top,$$

kde $\mathbf{U} \in \mathbb{R}^{n \times r}$, resp. $\mathbf{V} \in \mathbb{R}^{p \times r}$ značí ortonormální matice a $\mathbf{L} \in \mathbb{R}^{r \times r}$ je diagonální matice s prvky $l_1 \geq l_2 \geq \dots \geq l_r > 0$ na hlavní diagonále. Dále pro $\alpha \in [0, 1]$ definujme diagonální matici $\mathbf{L}^\alpha = \text{diag}(l_1^\alpha, l_2^\alpha, \dots, l_r^\alpha)$. Analogicky je zadefinována matice $\mathbf{L}^{1-\alpha}$. Označíme-li nyní matice

$$\mathbf{G} = \sqrt{n}\mathbf{U}\mathbf{L}^{1-\alpha} \quad \text{a} \quad \mathbf{H} = \frac{1}{\sqrt{n}}\mathbf{V}\mathbf{L}^\alpha,$$

dostáváme vyjádření matice pozorování jiným způsobem:

$$\begin{aligned} \mathbb{X} &= \mathbf{U}\mathbf{L}\mathbf{V}^\top = \mathbf{U}\mathbf{L}^{1-\alpha}\mathbf{L}^\alpha\mathbf{V}^\top = \\ &= \left(\sqrt{n}\mathbf{U}\mathbf{L}^{1-\alpha}\right) \left(\frac{1}{\sqrt{n}}\mathbf{L}^\alpha\mathbf{V}^\top\right) = \mathbf{G}\mathbf{H}^\top. \end{aligned} \quad (2.4)$$

Odtud také vidíme, že prvek na pozici (i, j) matice \mathbb{X} lze vyjádřit jako

$$x_{ij} = \sum_{k=1}^r u_{ik}l_kv_{jk} \quad (2.5)$$

$$= \mathbf{g}_i^\top \mathbf{h}_j, \quad (2.6)$$

kde \mathbf{g}_i a \mathbf{h}_j označuje i -tý, resp. j -tý sloupec matice \mathbf{G} a \mathbf{H} . Za předpokladu hodnoty matice \mathbb{X} alespoň dvě můžeme aproximovat její prvky tím způsobem, že ve výše uvedeném vztahu nasčítáme pouze první dva členy rozvoje. Dostaneme pak následující odhad analogický vztahům (2.5) a (2.6):

$$\tilde{x}_{ij} = \sum_{k=1}^2 u_{ik}l_kv_{jk} = \sum_{k=1}^2 g_{ik}h_{jk} = \tilde{\mathbf{g}}_i^\top \tilde{\mathbf{h}}_j,$$

kde vektory $\tilde{\mathbf{g}}_i$ a $\tilde{\mathbf{h}}_j$ obsahují první dva prvky vektorů \mathbf{g}_i a \mathbf{h}_j . Využitím rozkladu ze vzorce (2.4) jsme tedy matici \mathbb{X} aproximovali $n \times p$ rozměrnou maticí

$$\tilde{\mathbb{X}} = \tilde{\mathbb{G}}\tilde{\mathbb{H}}^\top,$$

kde matice $\tilde{\mathbb{G}}$ a $\tilde{\mathbb{H}}$ mají rozměry $n \times 2$ a $p \times 2$ a jsou tvořeny prvními dvěma sloupci matic \mathbb{G} a \mathbb{H} .

Biplot pak sestává z řádků matic $\tilde{\mathbb{G}}$ a $\tilde{\mathbb{H}}$, přičemž řádky $\tilde{\mathbb{G}}$ představují tzv. *skóry* a řádky $\tilde{\mathbb{H}}$ tzv. *zátěže* prvních dvou hlavních komponent. Řádky $\tilde{\mathbb{G}}$ jsou na biplotu vyznačeny jako body a řádky $\tilde{\mathbb{H}}$ jako šipky. Ty vycházejí z bodu, jehož souřadnice odpovídají aritmetickému průměru sloupců matice $\tilde{\mathbb{G}}$, a v případě, že pracujeme s centrovanými daty (tj. $\overline{\mathbf{X}} = \mathbf{0}_p$), tak z počátku soustavy souřadnic. K těmto charakteristikám se ještě vrátíme v části s grafickými ilustracemi, která následuje níže.

Nyní se podíváme na další vlastnosti biplotu, podrobnější informace jsou k dispozici v knize Jolliffe (2002, str. 90-96). Bez újmy na obecnosti budeme předpokládat, že data, která máme k dispozici, jsou již centrovaná. Pak má výběrová kovarianční matice tvar

$$\mathbb{S} = \frac{1}{n} \mathbb{X}^\top \mathbb{X}.$$

Vyjádření pozorování ve formě rozkladu (2.6) je nejednoznačné, neboť k němu vede více možných voleb α . Z hlediska lepší interpretace biplotu je zvykem položit $\alpha = 1$. Při této volbě dostáváme $\mathbb{G} = \sqrt{n}\mathbb{U}$, $\mathbb{H} = \frac{1}{\sqrt{n}}\mathbb{V}\mathbb{L}$, a využitím ortonormality matice \mathbb{U} také

$$\begin{aligned} \mathbb{X}^\top \mathbb{X} &= (\mathbb{G}\mathbb{H}^\top)^\top \mathbb{G}\mathbb{H}^\top = \mathbb{H}\mathbb{G}^\top \mathbb{G}\mathbb{H}^\top = n\mathbb{H}\mathbb{H}^\top \\ &= n\mathbb{S}. \end{aligned}$$

Odtud tedy vidíme, že součin $\mathbf{h}_j^\top \mathbf{h}_k$ představuje výběrovou kovarianci mezi j -tou a k -tou veličinou a $\tilde{\mathbf{h}}_j^\top \tilde{\mathbf{h}}_k$ ji aproximuje. Dále kosinus úhlu α , který svírají vektory $\tilde{\mathbf{h}}_j$ a $\tilde{\mathbf{h}}_k$, $j \neq k$, je přibližně roven výběrovému korelačnímu koeficientu mezi veličinami X_j a X_k , tj.

$$\cos \alpha = \frac{\tilde{\mathbf{h}}_j^\top \tilde{\mathbf{h}}_k}{\|\tilde{\mathbf{h}}_j\| \|\tilde{\mathbf{h}}_k\|} \approx r_{X_j X_k} \in [-1, 1].$$

Euklidovská vzdálenost mezi $\tilde{\mathbf{g}}_j$ a $\tilde{\mathbf{g}}_k$ je rovna Mahalanobisově vzdálenosti mezi vektory pozorování \mathbf{x}_j a \mathbf{x}_k , tedy

$$\|\tilde{\mathbf{g}}_j - \tilde{\mathbf{g}}_k\| = \|\mathbf{x}_j - \mathbf{x}_k\|_M.$$

V následující části uvedeme ukázkou použití biplotu na příkladu s reálnými daty. Jak již bylo zmíněno výše, šipky na grafu představují zkoumané náhodné veličiny. Délka šipky zhruba odpovídá rozptylu. Tedy čím je šipka delší, tím větší vliv bude mít náhodná veličina na uspořádání dat. Dále kosinus mezi šipkami přibližně vyjadřuje hodnotu korelačního koeficientu mezi příslušnými náhodnými veličinami. Čím je úhel mezi šipkami menší, tím více vztah mezi veličinami odpovídá lineární závislosti.

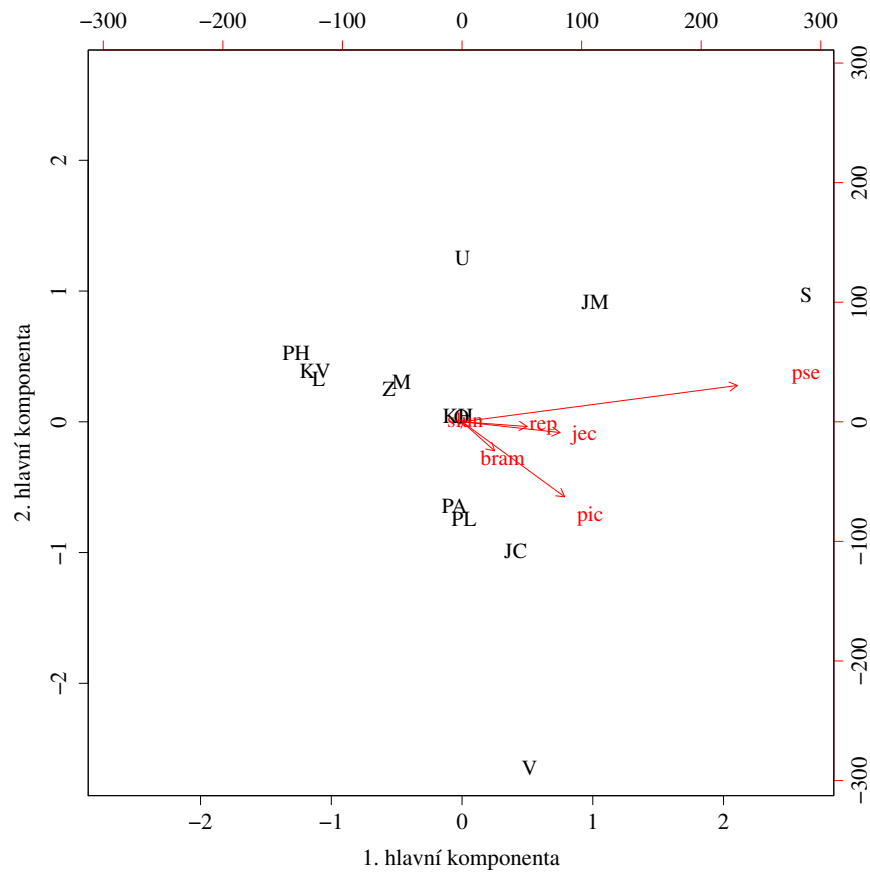
Příklad – zemědělství. Ukážeme si, jakým způsobem interpretovat biplot pro původní i standardizovaná data. Vycházíme z datového souboru, který udává hektarové výnosy sklizně hlavních zemědělských plodin v jednotlivých krajích České republiky za rok 2014 (původní i standardizovaná data jsou včetně značení přiřazených plodinám a jednotlivým krajům uvedena v přílohách A.1 a A.2). Zkoumanými veličinami jsou produkce pšenice, ječmene, brambor, řepky, slunečnice a píce. Dohromady tedy pracujeme s $6 \times 14 = 84$ pozorováními.

Nejprve se věnujme ukázce na původních datech. Pomocí funkce `prcomp` v softwaru R získáme základní představu o hlavních komponentách daného datového souboru. Část výstupu nalezneme v tabulce 2.1. První dva řádky uvádějí pořadí a přibližnou hodnotu směrodatné odchylky šesti hlavních komponent. Na třetím řádku je uvedeno, jak velká procentuální míra variability je příslušnou komponentou vysvětlena, poslední řádek pak odpovídá celkové procentuální variabilitě vysvětlené několika prvními komponentami. Již první komponenta vysvětluje přibližně 91% celkové variability, což se zdá jako velice dobrý výsledek. Pokud se však podíváme na strukturu dat, zjistíme, že mezi plodinami silně dominují výnosy pšenice, což nejspíš bude mít značný vliv i na podobu hlavních komponent. První hlavní komponentu lze charakterizovat jako průměr výnosů zemědělských plodin, kde největší váhu mají výnosy sklizně pšenice. Naproti tomu druhou hlavní komponentu lze popsat jako „kontrast“ – průměr výnosů se zápornými koeficienty u všech plodin kromě pšenice a slunečnice. V absolutní hodnotě největší váhu v tomto případě mají pícniny.

Podívejme se nyní na samotný biplot na obrázku 2.1. Rozmístění krajů zleva doprava naznačuje výnosnost v pěstování zemědělských plodin od nejmenší po největší. Nejhuře co do výnosnosti jsou na tom Praha, Karlovarský a Liberecký kraj. Zhruba uprostřed se nachází např. Královéhradecký a Pardubický kraj. Nejvyššími výnosy pak disponuje kraj Středočeský. Ve spodní části biplotu jsou rozmístěny ty kraje, u nichž převládají výnosy sklizně ječmene, brambor, řepky a pícnin (např. Vysočina), v horní části jsou oblasti s převládajícími výnosy sklizně pšenice a slunečnice (např. Ústecký kraj). Nejdelsí červená šipka přísluší pšenci, jež má největší vliv na rozložení bodů na biplotu. Šipky pro pícniny a brambory, resp. ječmen a řepku ukazují stejným směrem a úhel mezi nimi je malý, což odpovídá kladné korelovanosti a většímu vzájemnému vlivu.

Jak jsme již zmínili výše, výnosy pšenice dosahovaly ve srovnání se zbylými plodinami mnohem vyšších hodnot, kvůli čemuž mohl mezi veličinami vzniknout značný nepoměr. Z tohoto důvodu ještě provedeme standardizaci dat, pomocí které srovnáme důležitost všech plodin. V tabulce 2.2 vidíme, že po této úpravě datového souboru vysvětluje první hlavní komponenta 73,95% celkové variability, což je méně než v předchozím případě. První komponenta opět reprezentuje průměr všech veličin nyní s mnohem vyrovnanějšími váhami. Druhá hlavní komponenta má podobu průměru, ve kterém pšenice, ječmen a slunečnice vystupují s kladným znaménkem a brambory, řepka a pícniny se záporným znaménkem.

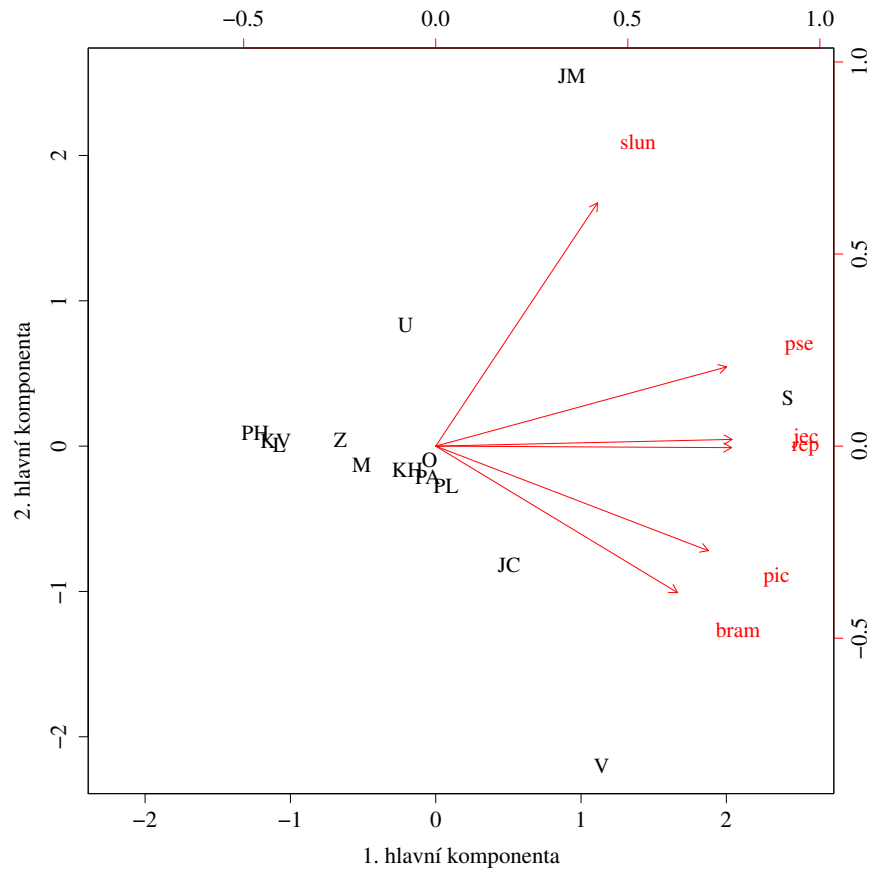
Kladně korelované jsou opět výnosy sklizně ječmene a řepky, podobně také výnosy sklizně brambor a pícnin, jak lze vidět na obrázku 2.2. Naproti tomu úhel, který svírají šipky příslušné bramborám a slunečnici, je téměř pravý, a proto tyto veličiny nejspíše budou nekorelované. Rozložení krajů, co se výnosů týče, je podobné jako v nestandardizovaném případě.



Obrázek 2.1: Biplot pro zemědělská data

k	1	2	3	4	5	6
sd	332,59	93,08	31,08	24,98	15,13	1,65
%	91,34	7,15	0,80	0,52	0,19	0,00
Σ	91,34	98,50	99,29	99,81	100,00	100,00

Tabulka 2.1: Variabilita vysvětlená hlavními komponentami pro zemědělská data



Obrázek 2.2: Biplot pro standardizovaná zemědělská data

k	1	2	3	4	5	6
sd	2,11	1,02	0,57	0,36	0,25	0,08
%	73,95	17,28	5,49	2,12	1,06	0,00
Σ	73,95	91,23	96,72	98,84	99,91	100,00

Tabulka 2.2: Variabilita vysvětlená hlavními komponentami pro standardizovaná zemědělská data

2.3 Volba počtu hlavních komponent – přehled literatury

Hlavní přínos metody hlavních komponent spočívá v tom, že několika prvními hlavními komponentami lze nahradit zkoumané náhodné veličiny, kterých je obvykle mnoho, a tím snížit dimenzi datového souboru na mnohem menší číslo. Vystává však otázka, jakým způsobem definovat tento vhodný počet hlavních komponent, abychom významně zredukovali počet zkoumaných veličin, avšak aby stále ještě nedošlo k markantní ztrátě informace. K problematice volby optimálního počtu lze přistoupit několika způsoby. Např. výpočetně jednodušší pravidla považují za vhodný takový počet několika prvních hlavních komponent, které splňují podmínku týkající se předem stanovené míry vysvětlené variability. U výpočetně složitějších metod se často po adekvátním počtu hlavních komponent požaduje, aby v nějakém smyslu optimalizoval předepsané kritérium. Rovněž je důležité si uvědomit, že většina metod nevede k odhadu kýženého optimálního počtu hlavních komponent ve statistickém slova smyslu.

Otázkou volby vhodného počtu hlavních komponent se rovněž zabývala celá řada autorů a máme k dispozici velké množství přehledových článků a knih zabývajících se tímto tématem. Ucelený přehled pravidel a metod nalezneme v knize Jolliffe (2002, kap. 6.1). Srovnání „heuristických“ pravidel (např. postupy odvozené na základě podílu vysvětlené variability, Kaiserovo-Guttmanovo kritérium, pravidla založená na grafickém výstupu atd.) s jejich obdobou v bootstrapové verzi a se statistickými testy nabízí článek Sobczyk a kol. (2017). Porovnání tohoto druhu metod je rovněž předmětem článku Peres-Neto a kol. (2005). Pohled na problematiku skrze zavedení pravděpodobnostního modelu nabízí článek Tipping a Bishop (1999). Otázkou výběru počtu hlavních komponent na základě dalších pokročilých metod se zabývá Park a Konishi (2017). Předmětem zájmu mnoha studií jsou také postupy založené na výpočetně náročných procedurách. Jmenujme např. metody odvozené na základě křížového ověřování, které shrnuje Bro a kol. (2008), nebo metody založené na základě bayesovského modelu, jež je blíže popsán v Hoyle (2008), Minka (2000), Seghouane a Cichocki (2007), Sobczyk a kol. (2017) a Suarez a Ghosal (2017). V rámci této kapitoly uvedeme přehled základních postupů, které jsou často založené na subjektivní úvaze či analýze grafického výstupu, a shrnutí výpočetně složitějších metod bude náplní kapitoly 5.

2.4 Heuristická pravidla

Heuristická pravidla, jež si nyní představíme, lze stručně shrnout jako „ad hoc pravidla“, která se obvykle snadno aplikují v praxi a jsou výpočetně rychlá. Avšak problém může nastat, pokud je používáme zcela automaticky, neboť jejich aplikace na data se specifickou strukturou může vést k zavádějícím výsledkům. Zaměříme se na metodu založenou na součtovém podílu variability, Kaiserovo-Guttmanovo kritérium, *broken stick* model, *scree* graf a LEV diagram.

Zmíněná pravidla aplikujeme na několik datových souborů z balíčku `SMSdata`, v softwaru R, jež je možné stáhnout na adrese Hlávka, Z. (2012). Konkrétně použijeme datové soubory `athletic` (výsledky atletické soutěže pro 55 zemí a 8

disciplín), `bank2` (hodnoty 6 rozměrových parametrů 100 švýcarských bankovek), `uscomp` (ekonomické parametry 78 firem v USA – 6 ukazatelů) a `uscrime` (záznamy o trestných činech v 50 státech USA – měřeno 7 proměnných). Z uvedených datových souborů jsme použili pouze spojitě proměnné a kvůli různému měřítku jsme provedli standardizaci veličin u všech datových souborů. Výběrové charakteristiky použitých dat jsou k dispozici v příloze A.3. Hlavní komponenty jsme pro každý z datových souborů získali pomocí příkazu `princomp`, který k výpočtu používá spektrální rozklad výběrové kovarianční matice. Příkaz `prcomp`, použitý při analýze hlavních komponent pro zemědělská data, naopak používá singulární rozklad datové matice. Oba přístupy jsou však ekvivalentní, a proto není příliš důležité, který z nich zvolíme.

2.4.1 Součtový podíl variability

Toto pravidlo je velmi intuitivní a již jsme je nastínili na předchozích místech této kapitoly. Zhruba řečeno požadujeme, aby informace obsažená v několika prvních hlavních komponentách dosahovala předem zvolené hodnoty. Výběrové hlavní komponenty jsou konstruovány tak, aby měly co možná největší rozptyl, přičemž rozptyl k -té výběrové komponenty se rovná k -tému vlastnímu číslu $\hat{\lambda}_k$ výběrové kovarianční matice. Navíc dle vztahu (2.3) platí, že součet rozptylů výběrových hlavních komponent se rovná součtu prvků výběrové kovarianční matice na hlavní diagonále. Využitím tohoto poznatku pak definujeme *součtový podíl variability* jako

$$s_m = \frac{\sum_{i=1}^m \hat{\lambda}_i}{\sum_{i=1}^p s_{ii}} = \frac{\sum_{i=1}^m \hat{\lambda}_i}{\sum_{i=1}^p \hat{\lambda}_i},$$

což se v případě použití výběrové korelační matice zjednoduší na tvar

$$s_m = \frac{\sum_{i=1}^m \hat{\lambda}_i}{p}.$$

Optimální počet hlavních komponent, který je na základě tohoto pravidla doporučeno uvažovat, je takové nejmenší číslo $m \in \{1, \dots, p\}$, pro které $s_m > s^*$, kde s^* se obvykle stanoví jako 0,95 (Jackson, 1993, str. 2207).

Otázkou však zůstává, jakým způsobem zvolit hodnotu s^* vzhledem k datovému souboru, jenž máme k dispozici. Obecně se zvyšujícím se počtem pozorování n či sledovaných statistických znaků p volíme spíše nižší s^* . Poznamenejme, že vždy je třeba zohlednit konkrétní strukturu dat. Pokud např. jsou první, resp. první dvě hlavní komponenty primárním zdrojem variability, který je možno podložit racionálními argumenty, položíme $s^* \geq 0,9$. Pokud ovšem pracujeme s velkým množstvím veličin, tj. číslo p je značně velké, pak se doporučuje zvolit $s^* \leq 0,7$. I přes svou snadnou aplikaci je kvůli značné subjektivitě při volbě kýžené hladiny s^* výše popsaná metoda považována za nespolehlivou.

2.4.2 Kaiserovo-Guttmanovo kritérium

Kritérium vychází z následující myšlenky: pokud jsou všechny složky náhodného vektoru \mathbf{X} navzájem nezávislé, pak příslušné hlavní komponenty splývají s těmito složkami a v případě použití korelační matice mají také jednotkové rozptyly. Pokud tedy má v obecném případě komponenta menší rozptyl než jedna,

obsahuje ve srovnání s původní veličinou méně informace, a proto se tato komponenta nepovažuje za dostatečně přínosnou.

Kaiserovo-Guttmanovo kritérium doporučuje zvolit takový počet hlavních komponent (odvozených na základě výběrové korelační matice), jejichž rozptyl $\hat{\lambda}_i$ přesáhne hodnotu $\lambda^* = 1$. Kromě intuitivní představy popsané výše se rovněž můžeme setkat s exaktnějším zdůvodněním vycházejícím z modelu faktorové analýzy (Jolliffe, 2002, kap. 6.1.2). Na základě simulačních studií se dále usoudilo, že vhodnější volbou je $\lambda^* = 0,7$ (viz tamtéž). V případě odvození hlavních komponent pomocí výběrové kovarianční matice je možné pravidlo použít tak, že jako bod zlomu stanovíme aritmetický průměr vlastních čísel

$$\bar{\lambda} = \frac{\sum_{i=1}^p \hat{\lambda}_i}{p},$$

resp. nižší hodnotu

$$\lambda^* = 0,7\bar{\lambda}.$$

2.4.3 Broken stick model

Na tomto místě zmiňme další pravidlo vycházející z rozboru samotných hodnot vlastních čísel. Jedná se o tzv. *broken stick* model (nebo též model „zlomené hůlky“), který představuje soubor rovnoměrně rozdělených částí, na které je náhodně rozdělen úsek délky jedna. Kritérium volby počtu hlavních komponent je založené na porovnání se střední hodnotou jednotlivých částí, o jejímž předpise vypovídá následující tvrzení.

Tvrzení 5. *Uvažujme úsek délky jedna, který je rozdělen na p částí tak, že body zlomu jsou nezávislé stejně rozdělené náhodné veličiny s rovnoměrným rozdělením na intervalu $[0, 1]$. Pak je střední hodnota m -té nejdelší části rovna*

$$\frac{1}{p} \sum_{i=m}^p \frac{1}{i}.$$

Důkaz. Důkaz je uveden ve článku Frontier (1976, str. 68). Tvrzení dokážeme jiným způsobem, a to pro střední hodnotu nejdelší části. Označme $X_1 < \dots < X_{p-1}$ náhodné veličiny reprezentující pozice dělení na jednotlivé části, ze kterých se úsek skládá. Počáteční bod úseku splývající s nulou značíme X_0 , koncový bod označujeme X_p a ten ztotožňujeme s jedničkou. Dále použijme následující značení pro délky každé z částí: $V_1 = X_1 - X_0, \dots, V_p = X_p - X_{p-1}$, přičemž platí omezení na celkový součet délek částí:

$$\sum_{i=1}^p V_i = 1.$$

Uvažujme pravděpodobnost, že jednotlivé délky v k -tici sestavené z V_1, \dots, V_p , prvky této k -tice označme jako V_1^*, \dots, V_k^* , překračují hodnoty $a_1, \dots, a_k > 0$, pro něž platí $\sum_{j=1}^k a_j < 1$. Pak podle vzorce (6.4.3) v knize David a Nagaraja (2003) lze ukázat, že se tato pravděpodobnost rovná

$$P \{V_1^* > a_1, \dots, V_k^* > a_k\} = (1 - a_1 - \dots - a_k)^{p-1}. \quad (2.7)$$

Interpretace pro $p = 2$ je následující – pravděpodobnost, že délka náhodně vybrané části (např. první zleva V_1 ; délky dvou částí jsou stejně rozdělené závislé náhodné veličiny) překročí a_1 , je $1 - a_1$, což je doplněk do jedničky. Pokud budeme uvažovat zvyšující se $p > 2$, potom bude pravděpodobnost $(1 - a_1)^{p-1}$ klesat, neboť bude pravděpodobnější, že je interval rozdělen na více částí, a pravděpodobnost překročení a_1 se bude snižovat. Uvedme ještě přehled užitých výrazů:

- p značí počet částí,
- $p - 1$ je počet bodů dělení,
- V_i značí náhodnou veličinu představující délku i -té části,
- $V_{(1)} = \max \{V_1, \dots, V_p\}$.

Využitím zmíněného poznatku se již dostáváme k určení pravděpodobnosti, že část s nejdelší délkou $V_{(1)}$ je větší než $x \in (0, 1)$:

$$\begin{aligned} \mathbb{P} \{V_{(1)} > x\} &= \mathbb{P} \{V_1 > x \vee V_2 > x \vee \dots \vee V_p > x\} \\ &= \sum_{k=1}^p (-1)^{k-1} \binom{p}{k} (1 - kx)^{p-1}, \end{aligned}$$

kde se uvažují pouze ty sčítance, pro které $kx < 1$. Druhá rovnost vychází z principu inkluze a exkluze (viz Matoušek a Nešetřil, 2009, str. 101, Věta 3.6.2). Nyní již můžeme vyjádřit střední hodnotu nejdelší části, přičemž ve výpočtu níže použijeme symbol $F_{V_{(1)}}(x)$ pro označení distribuční funkce veličiny $V_{(1)}$.

$$\begin{aligned} \mathbb{E} V_{(1)} &= \int_0^\infty (1 - F_{V_{(1)}}(x)) dx = \int_0^\infty \mathbb{P} \{V_{(1)} > x\} dx \\ &= \sum_{k=1}^p (-1)^{k-1} \binom{p}{k} \int_0^{1/k} (1 - kx)^{p-1} dx \\ &= \sum_{k=1}^p (-1)^{k-1} \binom{p}{k} \frac{1}{pk} = \frac{1}{p} \sum_{k=1}^p (-1)^{k-1} \frac{1}{k} \binom{p}{k} \\ &= \frac{1}{p} \sum_{i=1}^p \frac{1}{i}. \end{aligned}$$

Na druhé řádce odshora výše uvedeného výpočtu se uplatnilo omezení $kx < 1$ a poslední rovnost byla získána na základě obecně známé rovnosti pro harmonická čísla (viz poznámka za důkazem tvrzení), jejíž důkaz je uveden v příloze A.4. □

Poznamenejme, že výraz

$$H_p = \sum_{i=1}^p \frac{1}{i}$$

z Tvrzení 5 se označuje jako p -té harmonické číslo. Přehled středních hodnot délek jednotlivých částí pro vybrané počty dělení úseku p je uveden v tabulce 2.3. Můžeme si např. všimnout, že střední hodnota délky druhé nejdelší části z osmi, na které je úsek o délce jedna rovnoměrně rozlámán, se přibližně rovná 0,21.

$p \backslash m$	1	2	3	4	5	6	7	8	9	10
5	0,4567	0,2567	0,1567	0,09	0,04					
6	0,4083	0,2417	0,1583	0,1028	0,0611	0,0278				
7	0,3704	0,2276	0,1561	0,1085	0,0728	0,0442	0,0204			
8	0,3397	0,2147	0,1522	0,1106	0,0793	0,0543	0,0335	0,0156		
9	0,3143	0,2032	0,1477	0,1106	0,0828	0,0606	0,0421	0,0262	0,0123	
10	0,2929	0,1929	0,1429	0,1096	0,0846	0,0646	0,0479	0,0336	0,0211	0,01

Tabulka 2.3: Broken-stick model – střední hodnoty délek m -tých nejdelších částí úseku rozděleného na p kusů

Pravidlo volby optimálního počtu hlavních komponent je založeno na porovnání s modelem broken stick. Představme si, že se celková variabilita (součet vlastních čísel korelační nebo kovarianční matice) náhodně přiřadí jednotlivým hlavním komponentám. Pak se podíl variability připadající na jednotlivé hlavní komponenty porovná se středními hodnotami odpovídajících částí v broken stick modelu. Hlavní komponentu považujeme podle kritéria za dostatečně informativní, pokud její *podíl variability*, tj.

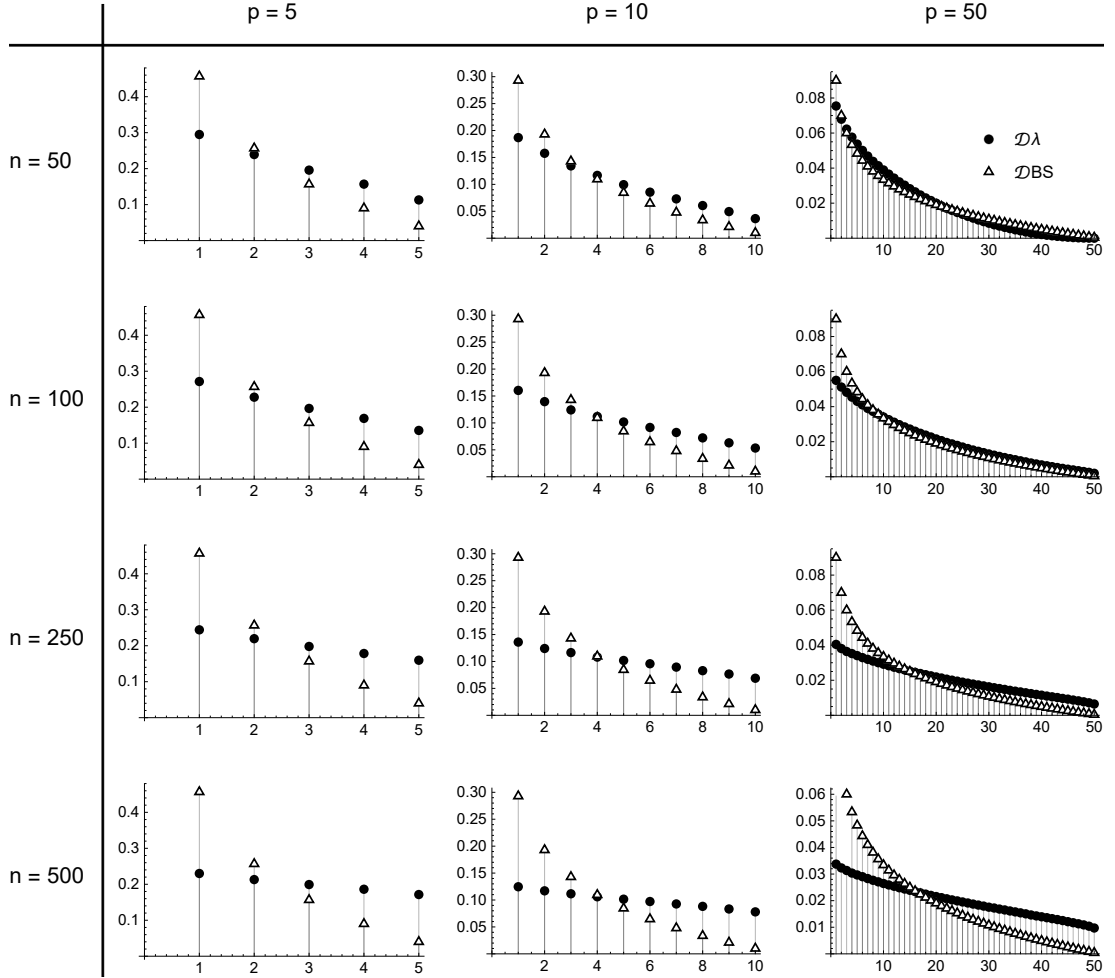
$$\frac{\hat{\lambda}_m}{\sum_{i=1}^p \hat{\lambda}_i}, \quad \text{v případě výběrové kovarianční matice, resp.}$$

$$\frac{\hat{\lambda}_m}{p}, \quad \text{v případě výběrové korelační matice,}$$

přesáhne střední hodnotu m -té nejdelší části v broken stick modelu, což je

$$\frac{1}{p} \sum_{i=m}^p \frac{1}{i} \tag{2.8}$$

(viz Jackson (1993, str. 2207) a Jolliffe (2002, str. 115, kap. 6.1.2)). Počet informativních hlavních komponent je zároveň optimálním počtem hlavních komponent, které bychom měli uvažovat. Uvážíme-li ilustraci speciálního případu na obrázku 2.3, jenž bude popsán v části níže, pak by podle kritéria byly za informativní prohlášeny hlavní komponenty s nižším podílem vysvětlené variability. Je otázkou, zda je tato volba skutečně vhodná, a jestli by nebylo lepší zvolit stejný optimální počet, avšak několika prvních hlavních komponent a striktně ne těch, které překročily střední hodnotu příslušného úseku v modelu. Pokud tedy nastane situace, že za informativní nebude prohlášena již první hlavní komponenta, považujeme doporučení na uvažování jiných, za informativních prohlášených, hlavních komponent za nespolehlivé.



Obrázek 2.3: Srovnání modelů \mathcal{D}_λ a \mathcal{D}_{BS}

2.4.4 Porovnání rozdělení vlastních čísel a broken stick modelu

Odlišnost středních hodnot, a tudíž i rozdělení podílu variability, jenž je vypočten na základě vlastních čísel (vlastní čísla, a tedy i podíl variability jsou seřazeny sestupně) bílé Wishartovy matice (viz definice 4 v kapitole 1), rozdělení značíme \mathcal{D}_λ , a rozdělení vzestupně seřazených úseků z modelu broken stick, značíme \mathcal{D}_{BS} , budeme demonstrovat na numerické simulaci.

Generování pseudonáhodného výběru z \mathcal{D}_{BS} rozdělení:

Ačkoliv ke kýženému porovnání obou rozdělení použijeme teoretické střední hodnoty modelu \mathcal{D}_{BS} , popíšeme efektivní způsob, jakým by bylo možné data z \mathcal{D}_{BS} rozdělení generovat. Náhodný výběr rozsahu n vytvoříme pomocí metody zvané *uniform-spacing*, která je k nalezení v knize Devroye (1986) a říká, že v případě že $\mathbf{Y} = (Y_1, \dots, Y_p)^\top$ je vektor nezávislých stejně rozdělených náhodných veličin s exponenciálním rozdělením se stejnou střední hodnotou, tak potom

$$V_i = \frac{Y_i}{\mathbf{1}_p^\top \mathbf{Y}}, \quad i = 1, \dots, p, \quad (2.9)$$

jsou stejně rozdělené náhodné veličiny. Zřejmě platí

$$V_1 + \dots + V_p = 1, \quad V_i \geq 0, \quad \forall i = 1, \dots, p$$

a lze tak generovat náhodně zvolené dělení intervalu jednotkové délky (simulace broken stick). Výhoda oproti rovnoměrnému vygenerování bodů dělení je skutečnost, že není potřeba body zlomů dvakrát seřadit a až následně mít rovnoměrně rozdělený interval. Poznamenejme, že v případě záměny exponenciálního rozdělení za např. rozdělení rovnoměrné normovaný výsledek (2.9) nevede k rozdělení intervalu s rovnoměrně rozdělenými pozicemi bodů dělení (Shaw, 2010).

Generování pseudonáhodného výběru z \mathcal{D}_λ rozdělení

Náhodný výběr rozsahu n , z něhož sestavíme matici $\mathbb{X}^\top \mathbb{X}$, budeme generovat z rozdělení $\mathcal{N}_p(\mathbf{0}, \mathbb{I}_p)$. Spočteme vlastní čísla této matice a každé z nich podělíme celkovým součtem vlastních čísel (to proto, aby součet prvků vektoru byl roven jedné).

Porovnání rozdělení \mathcal{D}_{BS} a \mathcal{D}_λ

V této numerické studii budeme porovnávat střední hodnoty pořádkových statistik. Pro rozdělení \mathcal{D}_{BS} použijeme teoretické střední hodnoty délek úseků a z náhodných výběrů z rozdělení \mathcal{D}_λ vytvoříme průměry. Tyto hodnoty budeme v grafech vykreslovat postupně zleva doprava, tj. odhad střední hodnoty maxima, popř. teoretická střední hodnota nejdelšího úseku jsou zcela nalevo. Vše provádíme na 1000 simulacích pro odlišná $n = 50, 100, 250, 500$ a $p = 5, 10, 50$. Pro tato data střední hodnoty úseků s nejvyššími délkami z \mathcal{D}_{BS} rozdělení přesahují odhady středních hodnot podílů variabilit, v případě úseků nižších délek je tomu většinou naopak, jak lze sledovat na obrázku 2.3. Zvláště zajímavý je „patologický“ případ $n = p = 50$, kdy došlo k dvojímu překřížení pomyslných křivek spojující teoretické, resp. odhadnuté střední hodnoty.

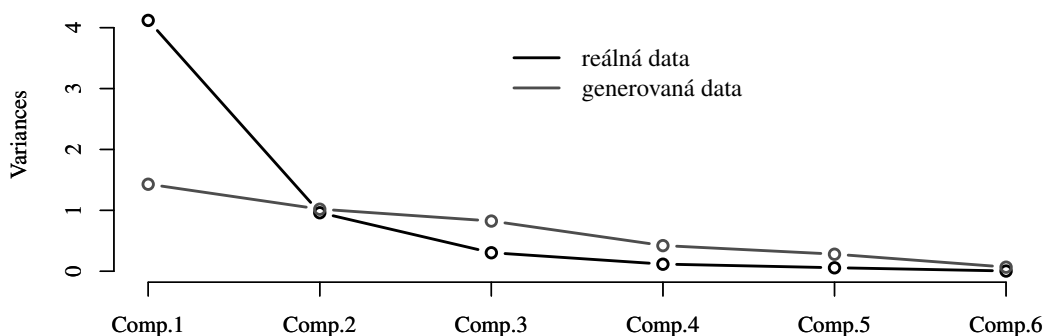
2.4.5 Scree graf

Toto pravidlo vychází z grafu, na jehož vodorovné, resp. svislé ose jsou znázorněna pořadí, resp. vlastní čísla příslušná jednotlivým hlavním komponentám, ať už odvozeným na základě výběrové kovarianční či korelační matice. Vlastní čísla s menšími hodnotami, které reprezentují části variability připadající na jednotlivé hlavní komponenty, by pak měla přibližně ležet na přímce (ne nutně vodorovné). Bod, ve kterém dochází k oddělení této přímky od zbytku vlastních čísel, rozděluje vlastní čísla příslušná interpretovatelným komponentám (nalevo od kýženého bodu) od těch, které nesou pouze menší množství informace. Někteří autoři doporučují rovněž zahrnout komponentu, jež přísluší prvnímu vlastnímu číslu napravo od kýženého oddělujícího bodu.

Nevýhodou tohoto postupu je častá nepřítomnost očividného oddělujícího bodu, případně větší množství potenciálních oddělujících bodů, což do problematiky volby adekvátního počtu hlavních komponent vnáší ještě více subjektivity. Proto existuje celá řada modifikovaných metod, z nichž několik zmíníme. Jedním z alternativních přístupů je výpočet vlastních čísel korelační matice příslušné náhodnému výběru z rozdělení $\mathcal{N}_p(\mathbf{0}, \mathbb{I}_p)$, kde délka náhodných vektorů p a počet realizací n jsou totožné s rozsahem původního datového souboru. Takto získaná

vlastní čísla se zakreslí do klasického scree grafu. Vlastní čísla nalevo od průsečíku dvou vzniklých křivek pak přísluší interpretovatelným komponentám. Takto modifikovaný scree graf jsme aplikovali na standardizovaná zemědělská data a jim odpovídající náhodný výběr o rozsahu $n = 14$ z rozdělení $\mathcal{N}_6(\mathbf{0}, \mathbb{I}_6)$. Podle popsaného pravidla bychom zvolili pouze první hlavní komponentu, jak dokládá obrázek 2.4. Hlavní komponenty pro standardizovaná zemědělská data a data z mnoho-rozměrného normálního rozdělení byly spočítány pomocí příkazu `princomp`. Další variantou scree grafu je *LEV diagram*, který si nyní popíšeme.

Scree graf pro reálná a generovaná data



Obrázek 2.4: Scree graf pro standardizovaná zemědělská data a náhodný výběr o rozsahu $n = 14$ z rozdělení $\mathcal{N}_6(\mathbf{0}, \mathbb{I}_6)$. Na vodorovné ose je pořadí jednotlivých hlavních komponent, na svislé ose jejich rozptyly (příslušná vlastní čísla).

2.4.6 LEV diagram

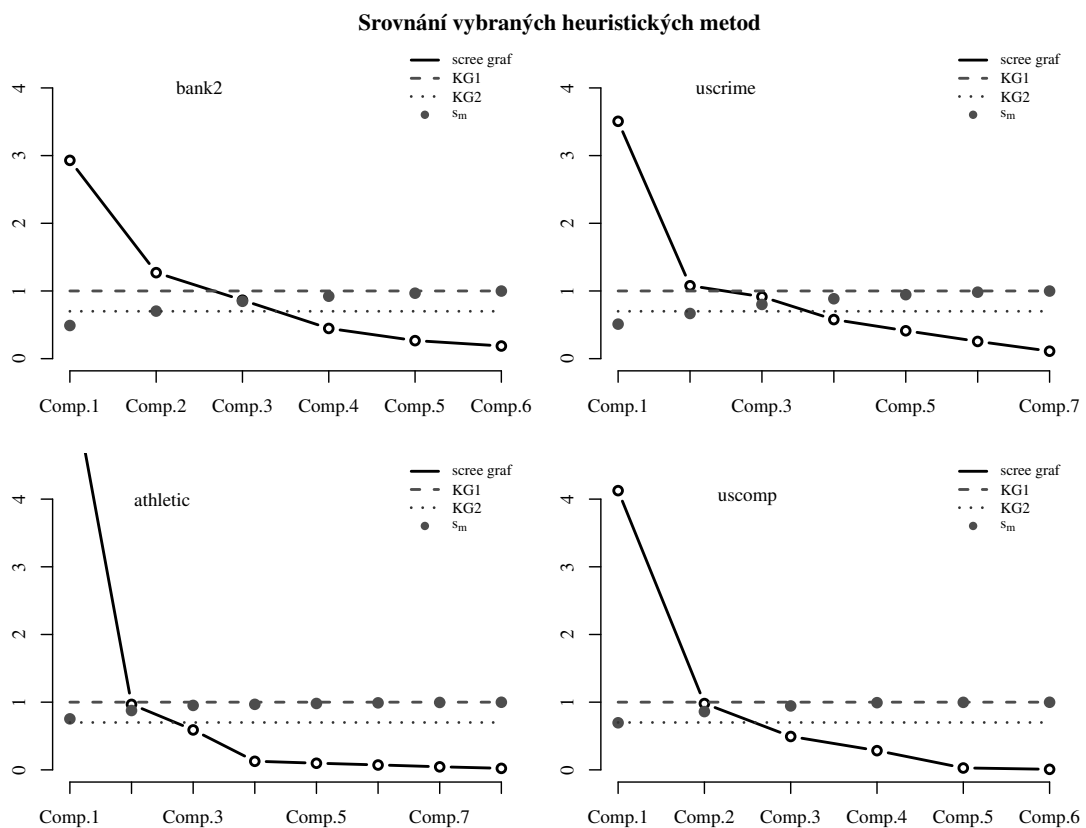
LEV (*Log-Eigenvalue*) diagram zobrazuje na vodorovné ose pořadí hlavní komponenty k a na svislé ose hodnotu $\log \hat{\lambda}_k$, kde $\hat{\lambda}_k$ označuje k -té vlastní číslo výběrové kovarianční, resp. korelační matice. Jedná se tedy o alternativní zobrazení scree grafu zmíněného v předchozí podkapitole. Optimální počet hlavních komponent se pak rovná pořadí komponenty, od které budou logaritmy vlastních čísel ležet přibližně na přímce. Podle článku Craddock a Flood (1969) je vysvětlení takové, že vlastní čísla spojená s vlastními vektory, které odpovídají části matice s chybovými („šum“), nikoliv systematickými složkami („signál“), klesají geometricky (tj. ve smyslu geometrické posloupnosti).

2.5 Shrnutí heuristických pravidel

Na základě přehledu heuristických pravidel pro volbu vhodného počtu hlavních komponent jsme se přesvědčili, že jsou sice výpočetně nenáročná, avšak jejich nevýhodou je značná subjektivita a často také poněkud zavádějící popis. Problematiké je rovněž porovnávání vlastních čísel s konstantami, jejichž původ není

řádně vysvětlen. Také jsme zjistili, že pokud aplikací některého kritéria, např. broken stick modelu, obdržíme doporučený počet hlavních komponent, pak podle popisu kritéria ještě není zcela zřejmé, které hlavní komponenty bychom měli nadále uvažovat. Tato situace se především týká speciálních případů, jakým byla data z obrázku 2.3. Během používání heuristických procedur tedy doporučujeme být obezřetní.

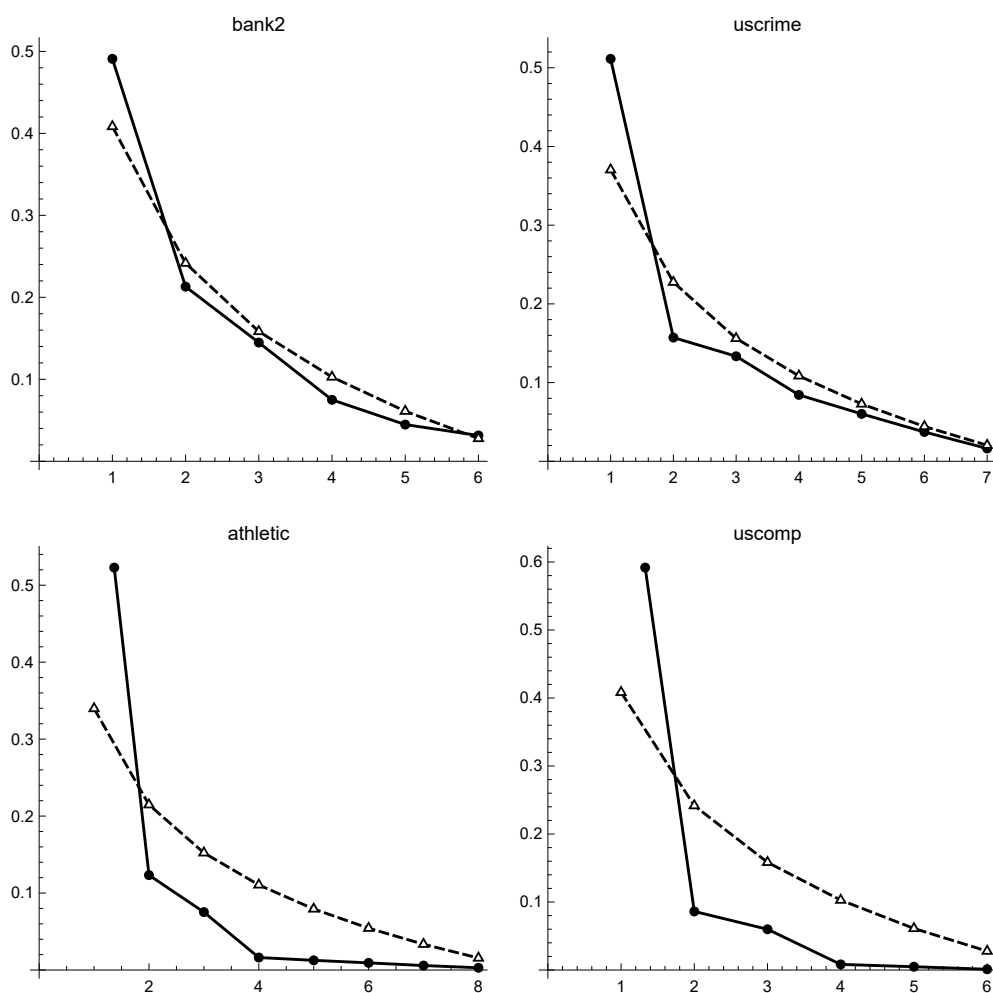
Heuristické přístupy aplikujeme na standardizovaná data z balíčku *SMSdata*. Na základě obrázků 2.5, 2.6 a výsledky z nich shrnující tabulky 2.4 usuzujeme, že nejmenší počet hlavních komponent volí pravidlo založené na broken stick modelu. Procedura zkoumající součtový podíl variability jich naopak radí zvolit nejvíce.



Obrázek 2.5: Srovnání výsledků heuristických metod aplikovaných na data z balíčku *SMSdata* – scree graf, Kaiserovo-Guttmanovo kritérium s volbou porovnávání s konstantou 1 (KG1) a 0,7 (KG2) a součtový podíl variability (znázorněny hodnoty s_m , hranice $s^* = 0,9$ kvůli přehlednosti již nevykreslena). Vlastní čísla byla vypočtena na základě výběrové korelační matice. Na vodorovné ose je uvedeno pořadí jednotlivých hlavních komponent.

	bank2	uscrime	athletic	uscomp
scree graf	2	1	3	2
KG1 ($\hat{\lambda} = 1$)	2	2	2	2
KG2 ($\hat{\lambda} = 0,7$)	3	3	2	2
$s^* = 0,9$	4	5	3	3
broken stick	1	1	1	1

Tabulka 2.4: Doporučené volby počtu hlavních komponent na základě heuristických pravidel



Obrázek 2.6: Srovnání broken stick modelu s podíly variability připadající na jednotlivé komponenty, jejichž pořadí je vyznačeno na vodorovné ose. Vlastní čísla byla spočtena na základě výběrové korelační matice a data pocházejí z balíčku SMSdata.

3. Asymptotické vlastnosti výběrových vlastních čísel

Tato kapitola se blíže věnuje asymptotickému rozdělení vlastních čísel výběrové kovarianční matice a největšího vlastního čísla bílé Wishartovy matice. Zkoumání vlastností vlastních čísel je obecně velmi komplikované, neboť vlastní čísla nelze zapsat jako racionální funkci prvků příslušné matice (Chiani, 2014). Z tohoto důvodu se např. při odvozování asymptotického rozdělení vlastních čísel klade množství omezujících požadavků na původní matici. Nejprve se však podíváme na přesné rozdělení vektoru výběrových vlastních čísel bílé Wishartovy matice (viz definice 4 v kapitole 1). O problematice přesného rozdělení vektoru pozorovaných vlastních čísel takovéto matice pojednává článek Chiani (2014).

3.1 Přesné rozdělení výběrových vlastních čísel

Uvažujme bílou Wishartovu matici $M \sim \mathcal{W}_p(\mathbb{I}_p, n)$, $M \in \mathbb{R}^{p \times p}$, pro kterou platí

$$M = X^T X, \quad X \in \mathbb{R}^{n \times p}, \quad (3.1)$$

kde matice X vznikla jako náhodný výběr z rozdělení $\mathcal{N}_p(\mathbf{0}, \mathbb{I}_p)$ uspořádaný do řádků. Označme $c = \min\{n, p\}$ a $d = \max\{n, p\}$. Pak je podle článku Chiani (2014, str. 70) sdružená hustota vektoru sestupně seřazených vlastních čísel $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_c > 0$ matice M ze vzorce (3.1) dána předpisem

$$f_\lambda(x_1, \dots, x_c) = K \prod_{i=1}^c \exp^{-\frac{x_i}{2}} x_i^\alpha \prod_{j=i+1}^c (x_i - x_j), \quad (3.2)$$

kde $x_1 \geq x_2 \geq \dots \geq x_c > 0$, $\alpha = (d - c - 1)/2$ a K je vhodně zvolená normovací konstanta, jejíž předpis je uveden v článku Chiani (2014, str. 70). Ještě poznamenejme, že ve většině aplikací platí $c = p$ a $d = n$.

Nyní provedeme softwarově symbolický výpočet marginálních hustot přesného rozdělení vlastních čísel bílé Wishartovy matice. Zmíněné teoretické poznatky tedy rozšíříme o simulaci, kterou konfrontujeme s analytickými předpisy pro hustotu, kterou získáme integrací sdružené hustoty (3.2) pomocí symbolického výpočtu v softwaru *Mathematica*. Abychom získali marginální hustoty, budeme integrovat sdruženou hustotu vektoru vlastních čísel přes přebytečné proměnné na definičním oboru. Tuto integraci budeme provádět symbolicky a v této ilustraci se omezíme na $p = 3$. Marginální hustoty vlastních čísel λ_1 , λ_2 a λ_3 získáme pomocí vzorců

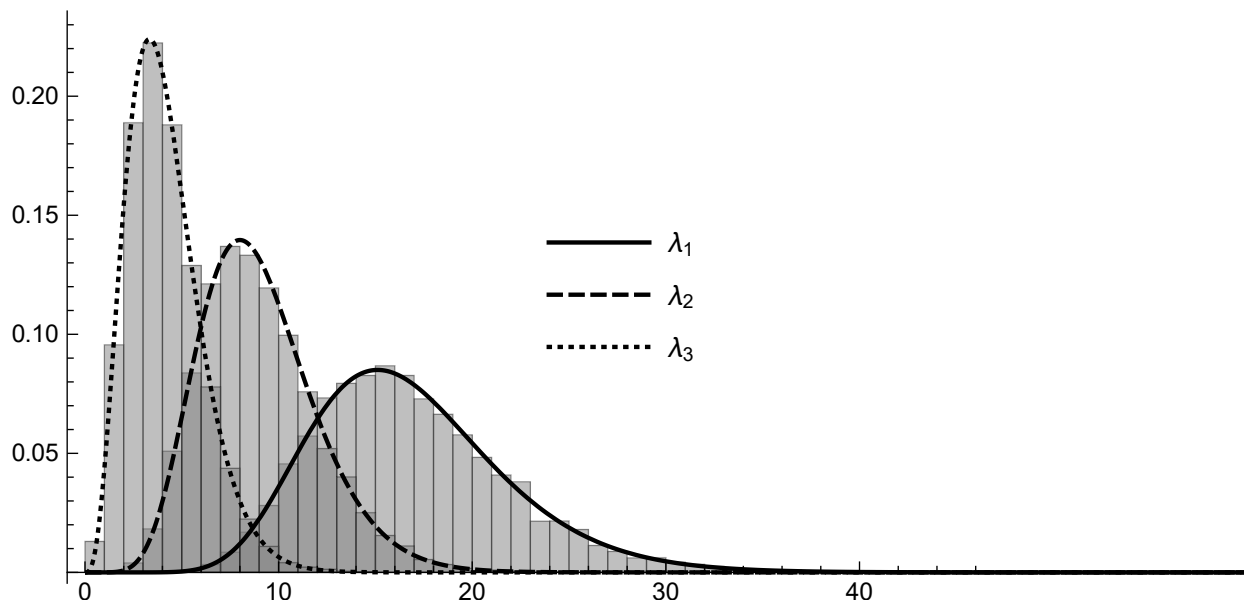
$$f_{\lambda_1}(x) = \int_0^x \int_0^{x_2} f_\lambda(x, x_2, x_3) dx_3 dx_2, \quad (3.3)$$

$$f_{\lambda_2}(x) = \int_x^\infty \int_0^x f_\lambda(x_1, x, x_3) dx_3 dx_1, \quad (3.4)$$

$$f_{\lambda_3}(x) = \int_x^\infty \int_x^{x_1} f_\lambda(x_1, x_2, x) dx_2 dx_1. \quad (3.5)$$

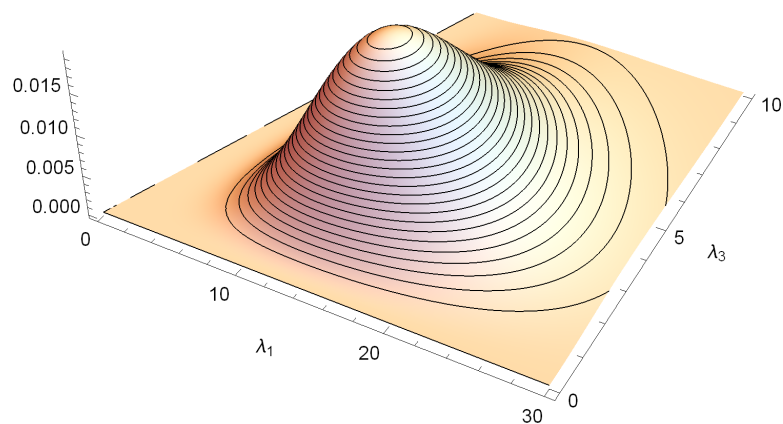
Připomeňme, že sdružená hustota (3.2) uvedená v příloze A.5 udává rozdělení vektoru vlastních čísel, jež jsou sestupně seřazena podle velikosti, což jsme vhodným způsobem zohlednili při volbě integračních mezí ve vzorcích (3.3) – (3.5).

Předpisy pro marginální hustoty mají velmi komplikovaný tvar a jsou rovněž uvedeny v příloze A.5. Analytické předpisy obsahující součin exponenciální funkce a polynomu jsou k nalezení v příloze ve zdrojovém kódu softwaru *Mathematica*. V případě, že vykreslíme histogramy pro 10 000 náhodných výběrů při $p = 3$ a $n = 10$, dostáváme výsledky, které jsou znázorněny na obrázku 3.1. Tvar sružené hustoty největšího a nejmenšího vlastního čísla pozorujeme na obrázku 3.2.



Obrázek 3.1: Marginální hustoty vlastních čísel matice z rozdělení $\mathcal{W}_3(\mathbb{I}_3, 10)$, realizováno pro 10 000 náhodných výběrů

Sružená hustota λ_1 a λ_3



Obrázek 3.2: Sružená hustota největšího a nejmenšího vlastního čísla matice z rozdělení $\mathcal{W}_3(\mathbb{I}_3, 10)$

3.2 Asymptotické vlastnosti výběrových vlastních čísel za předpokladu jejich různosti

Nejprve se podíváme na několik vlastností výběrových hlavních komponent za předpokladu, že původní data pocházejí z normálního rozdělení. Zároveň se podrobněji zaměříme na předpoklad pro příslušná vlastní čísla, která mají být navzájem různá, a na dopady porušení tohoto požadavku. Následující věta pojednává o tom, že za daných předpokladů jsou výběrové hlavní komponenty a vlastní čísla výběrové kovarianční matice v jistém smyslu optimálními odhady svých nevíběrových verzí.

Tvrzení 6. *Uvažujme náhodný výběr z mnohorozměrného normálního rozdělení s kovarianční maticí Σ , jejíž vlastní čísla jsou navzájem různá. Pak výběrové hlavní komponenty a vlastní čísla výběrové kovarianční matice S jsou maximálně věrohodnými odhady jejich teoretických protějšků.*

Důkaz. Viz Mardia a kol. (2003, str. 229, Theorem 8.3.1) □

V případě, že vlastní čísla kovarianční matice nejsou navzájem různá, výše uvedená věta neplatí. Dokládá to např. speciální případ $\Sigma = \sigma^2 \mathbb{I}_p$, kdy p -rozměrná diagonální matice Σ má vlastní číslo σ^2 násobnosti p , ke kterému lze přiřadit jakýkoliv p -rozměrný nenulový vektor jakožto vlastní vektor, z čehož vyplývá nejednoznačnost. Navíc vlastní čísla matice S mohou být obecně navzájem různá, ačkoliv jejich protějšky příslušné matici Σ tuto vlastnost nemají. Za této situace a při splnění dalších požadavků však platí následující obdoba výše uvedené věty.

Tvrzení 7. *Nechť je daný náhodný výběr z mnohorozměrného normálního rozdělení s kovarianční maticí Σ , jejíž $k > 1$ vlastních čísel nabývá hodnoty λ^* . Příslušnou výběrovou kovarianční matici označme jako S . Pak*

- (i) *aritmetický průměr \bar{l} odpovídajících vlastních čísel matice S je maximálně věrohodným odhadem λ^* a*
- (ii) *vlastní vektory matice S odpovídající stejným vlastním číslům matice Σ jsou maximálně věrohodnými odhady odpovídajících vlastních vektorů v teoretické verzi, ačkoliv nejsou jednoznačně určeny.*

Důkaz. Viz Anderson (1963, str. 130, Theorem 2), kde je třeba dosadit $n = N$. □

Nyní shrneme asymptotické vlastnosti výběrových vlastních čísel a vlastních vektorů za předpokladu, že jsou vlastní čísla kovarianční matice navzájem různá.

Věta 8. *Nechť Σ značí pozitivně definitní matici, jejíž vlastní čísla jsou navzájem různá. Dále označme matice $M \sim \mathcal{W}_p(\Sigma, n)$ a $U = n^{-1}M$ a uvažujme singulární rozklady $\Sigma = \Gamma\Lambda\Gamma^\top$ a $U = \mathbb{G}\mathbb{L}\mathbb{G}^\top$ s maticemi $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_p)$ a $\mathbb{L} = \text{diag}(\hat{\lambda}_1, \dots, \hat{\lambda}_p)$. Pak za těchto podmínek platí*

$$(i) \sqrt{n}(\hat{\lambda}_i - \lambda_i) \stackrel{as}{\sim} \mathcal{N}(0, 2\lambda_i^2), \quad i = 1, \dots, p,$$

$$(ii) \sqrt{n}(\hat{\gamma}_i - \gamma_i) \stackrel{as}{\sim} \mathcal{N}_p(\mathbf{0}, \mathbb{V}_i), \quad i = 1, \dots, p, \quad \text{kde } \hat{\gamma}_i, \text{ resp. } \gamma_i \text{ značí } i\text{-tý sloupec matice } \mathbb{G}, \text{ resp. } \Gamma \text{ a}$$

$$\mathbb{V}_i = \lambda_i \sum_{k \neq i} \frac{\lambda_k}{(\lambda_k - \lambda_i)^2} \gamma_k \gamma_k^\top.$$

(iii) *Prvky vektoru $(\hat{\lambda}_1, \dots, \hat{\lambda}_p)^\top$ jsou asymptoticky nezávislé s prvky matice \mathbb{G} .*

Důkaz. Viz Mardia a kol. (2003, str. 230, Theorem 8.3.3) pro důkaz části (i) a Anderson (1963) pro důkaz zbylých částí věty. □

Obsah Věty 8 nám mimo jiné říká, že vlastní čísla matice U mají asymptoticky normální rozdělení, jsou asymptoticky nestrannými odhady vlastních čísel matice Λ a platí jejich asymptotická nezávislost. Podobně i vlastní vektory matice U mají asymptoticky normální rozdělení a představují nestranné odhady vlastních vektorů matice Σ .

3.3 Tracyho-Widomovo rozdělení

Při zkoumání asymptotického rozdělení největšího vlastního čísla bílé Wishartovy matice se setkáme s Tracyho-Widomovým rozdělením řádu 1. Jelikož toto rozdělení není příliš známé, uveďme si jeho definici.

Definice 8. *(Tracyho-Widomovo rozdělení) Tracyho-Widomovo rozdělení řádu 1 (značíme jako \mathcal{TW}_1) je definováno distribuční funkcí, jež se řídí následujícím předpisem*

$$F_1(s) = \exp \left\{ -\frac{1}{2} \int_s^\infty q(x) + (x-s)q^2(x) dx \right\}, \quad s \in \mathbb{R}.$$

Ve výše uvedeném symbol $q(x)$ představuje jediné řešení nelineární Painlevého diferenciální rovnice

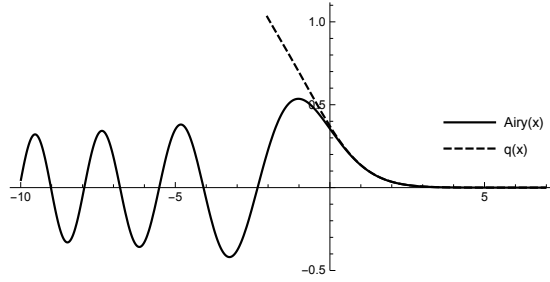
$$q''(x) = xq(x) + 2q^3(x).$$

Navíc se požaduje, aby $q(x)$ splňovalo podmínku

$$q(x) \approx Ai(x) \text{ při } x \rightarrow \infty,$$

kde $Ai(x)$ značí Airyho funkci.

Více podrobností o Airyho funkci najdeme v kapitole 9 knihy Abramowitz a Stegun (2010). Předpis distribuční funkce \mathcal{TW}_1 z definice 8 jsme implementovali v softwaru *Mathematica* a porovnali se zabudovaným předpisem pro tuto distribuční funkci.



Obrázek 3.3: Porovnání Airyho funkce a funkce $q(x)$ z definice 8 implementované v softwaru *Mathematica*

Shrnutí numericky spočtených základních charakteristik Tracyho-Widomova rozdělení je uvedeno v tabulce 3.1 (viz Tracy a Widom, 2009, str. 757).

μ_1	σ_1^2	γ_3	γ_4
-1,207	1,608	0,293	0,165

Tabulka 3.1: Charakteristiky rozdělení \mathcal{TW}_1 , ve sloupcích po řadě zleva doprava: střední hodnota, rozptyl, šikmost, špičatost (jedná se o upravený vzorec s odečtenou špičatostí normálního rozdělení)

3.3.1 Aproximace Tracyho-Widomova rozdělení

Na základě numerické analýzy hustoty Tracyho-Widomova rozdělení řádu 1 bylo zjištěno, že jej lze dobře aproximovat gama rozdělením. V této části práce předvedeme v numerické studii, jak lze tuto aproximaci provést různými způsoby.

Aproximace pomocí posunutého gama rozdělení ($\mathbf{S}\Gamma$)

Distribuční funkci $F_1(x)$ a hustotu $f_1(x)$ rozdělení \mathcal{TW}_1 lze aproximovat pomocí jejich protějšků gama rozdělení (viz definice 5 v kapitole 1), tj.

$$F_1(x) \approx \frac{1}{\Gamma(k)} \gamma\left(k, \frac{x - \mu}{\theta}\right),$$

$$f_1(x) \approx \frac{1}{\Gamma(k)\theta^k} (x - \mu)^{k-1} e^{-\frac{x-\mu}{\theta}},$$

v obou případech se požaduje $x > \mu$. K aproximaci parametrů posunutého gama rozdělení k, θ a μ autoři v literatuře přistupují různě. Např. Chiani (2014, str. 76) používá k aproximaci porovnání tří teoretických charakteristik, což jsou střední hodnota, rozptyl a šikmost posunutého gama rozdělení, tj.

$$\mathbb{E} G = k\theta + \mu, \quad \text{var } G = k\theta^2, \quad \text{Skew } G = \frac{2}{\sqrt{k}},$$

kde $G \sim \mathbf{S}\Gamma(k, \theta, \mu)$. Ty se porovnají s odpovídajícími charakteristikami \mathcal{TW}_1 rozdělení, které značíme μ_1, σ_1^2 a γ_3 jako střední hodnotu, rozptyl a šikmost. Vznikne tak soustava tří rovnic, jejímž řešením je

$$\mu = \mu_1 - \frac{2\sigma_1}{\gamma_3}, \quad k = \frac{4}{\gamma_3^2}, \quad \theta = \frac{\sigma_1\gamma_3}{2}.$$

Aproximace pomocí zobecněného gama rozdělení (SG Γ)

Aproximaci \mathcal{TW}_1 rozdělení můžeme rovněž provést pomocí zobecněného gama rozdělení s parametrem posunutí (polohy), jak je uvedeno v definici 6 v kapitole 1. Distribuční funkci $F_1(x)$ a hustotu $f_1(x)$ rozdělení \mathcal{TW}_1 je pak možné aproximovat jako

$$F_1(x) \approx \frac{1}{\Gamma(k)} \gamma \left(k, \left(\frac{x - \mu}{\theta} \right)^\beta \right),$$

$$f_1(x) \approx \frac{\beta}{\Gamma(k)\theta} \left(\frac{x - \mu}{\theta} \right)^{k\beta-1} e^{-\left(\frac{x-\mu}{\theta}\right)^\beta},$$

pro oba případy se vyžaduje $x > \mu$. Aproximaci rozdělení \mathcal{TW}_1 provedeme porovnáním čtyř teoretických charakteristik obou rozdělení.

Aproximace pomocí posunutého gama rozdělení (S Γ^*) – modifikace výpočtu

Aproximaci lze také provést pomocí posunutého gama rozdělení, které má tři parametry, avšak kromě střední hodnoty, rozptylu a šikmosti zohledníme rovněž špičatost. Z tohoto důvodu budeme minimalizovat funkcionál

$$\min_{k>0, \theta>0, \beta>0, \mu \in \mathbb{R}} \left\{ \sum_{i=1}^4 \left(\frac{m_i(\mathcal{D}) - m_i(\mathcal{TW}_1)}{m_i(\mathcal{TW}_1)} \right)^2 \right\} \quad (3.6)$$

kde $m_i(\mathcal{D})$ označuje i -tou teoretickou charakteristiku (po řadě střední hodnota, rozptyl, šikmost a špičatost) příslušného rozdělení \mathcal{D} , přičemž $m_i(\mathcal{TW}_1)$ získáme přímo ze softwarové implementace. Počáteční podmínky pro parametry nastavíme jako výsledné hodnoty z modelu S Γ a parametr β nastavíme rovný jedné. Newtonova metoda, kterou použijeme, nepřipouští omezující podmínky a v případě nevhodné volby můžeme dojít k parametrům, které jsou záporné a nedávají smysl.

Přednost použití funkcionálu (3.6) spočívá v jeho konvexnosti, a tudíž vhodnosti pro metody založené na gradientním přístupu, jako je Newtonova metoda. Další výhodou je určitá numerická stabilita v charakteristikách. Alternativně bychom mohli jednotlivým charakteristikám přiřadit vhodně stanovené váhy, tento přístup by již vykazoval určitou míru subjektivity.

Určitou představu o přesnosti aproximace si lze vytvořit na základě údajů v tabulce 3.3, kde užíváme značení pro absolutní a relativní chybu aproximace:

$$\Delta(\mathcal{D}) = m_i(\mathcal{D}) - m_i(\mathcal{TW}_1),$$

$$\Delta_r(\mathcal{D}) = \frac{m_i(\mathcal{D}) - m_i(\mathcal{TW}_1)}{m_i(\mathcal{TW}_1)}.$$

Hustota aproximujícího gama rozdělení je téměř v zákrytu hustoty \mathcal{TW}_1 , jak pozorujeme na obrázku 3.4. Výraznější rozdíl je patrný až při volbě logaritmického měřítka na svislé ose.

Shrnutí aproximačních metod

Ve stručné numerické studii ukážeme, jak lze Tracyho-Widomovo rozdělení řádu 1 aproximovat několika způsoby pomocí gama rozdělení, jak bylo popsáno v předcházejícím textu. Budeme porovnávat následující hustoty:

- \mathcal{TW}_1 – Tracyho-Widomovo rozdělení řádu 1 (vnitřní implementace softwaru *Mathematica*),
- \mathbf{ST} – posunuté gama rozdělení z definice 5, kde $\mu = 1$, shodná střední hodnota, rozptyl a šikmost
- \mathbf{SGT} – zobecněné gama rozdělení z definice 6, shodná střední hodnota, rozptyl, šikmost a špičatost
- \mathbf{ST}^* – posunuté gama rozdělení z definice 6, kde $\beta = 1$, získané minimalizací relativní kvadratické chyby střední hodnoty, rozptylu, šikmosti a špičatosti.

Poznamenejme, že v případech \mathbf{ST} a \mathbf{SGT} odpovídal vždy počet parametrů počtu rovnic a bylo možné získat přesné řešení (až na chyby v aritmetické přesnosti). Tedy jsou ve shodě střední hodnota, rozptyl, šikmost a i špičatost v případě \mathbf{SGT} . V případě \mathbf{ST}^* máme ale 4 rovnice a 3 parametry a výsledné řešení nalezneme minimalizací relativní kvadratické chyby podle vzorce (3.6). Jedná se o způsob, jak chybu spravedlivě rozdělit mezi všechny čtyři charakteristiky a přitom žádný z nich neopomenout. Ve všech případech bylo řešení hledáno pomocí modifikované Newtonovy metody (Wolfram Research, Inc., 2018) implementované v softwaru *Mathematica*. Aproximace, co se tvaru hustoty týče, je úspěšná, jak lze sledovat na obrázku 3.4, kde je také zobrazen detail hustoty v modu nebo odchýlení na chvostu v logaritmickém měřítku. Aproximace parametrů hustot \mathbf{ST} , \mathbf{SGT} a \mathbf{ST}^* jsou uvedeny v tabulce 3.2.

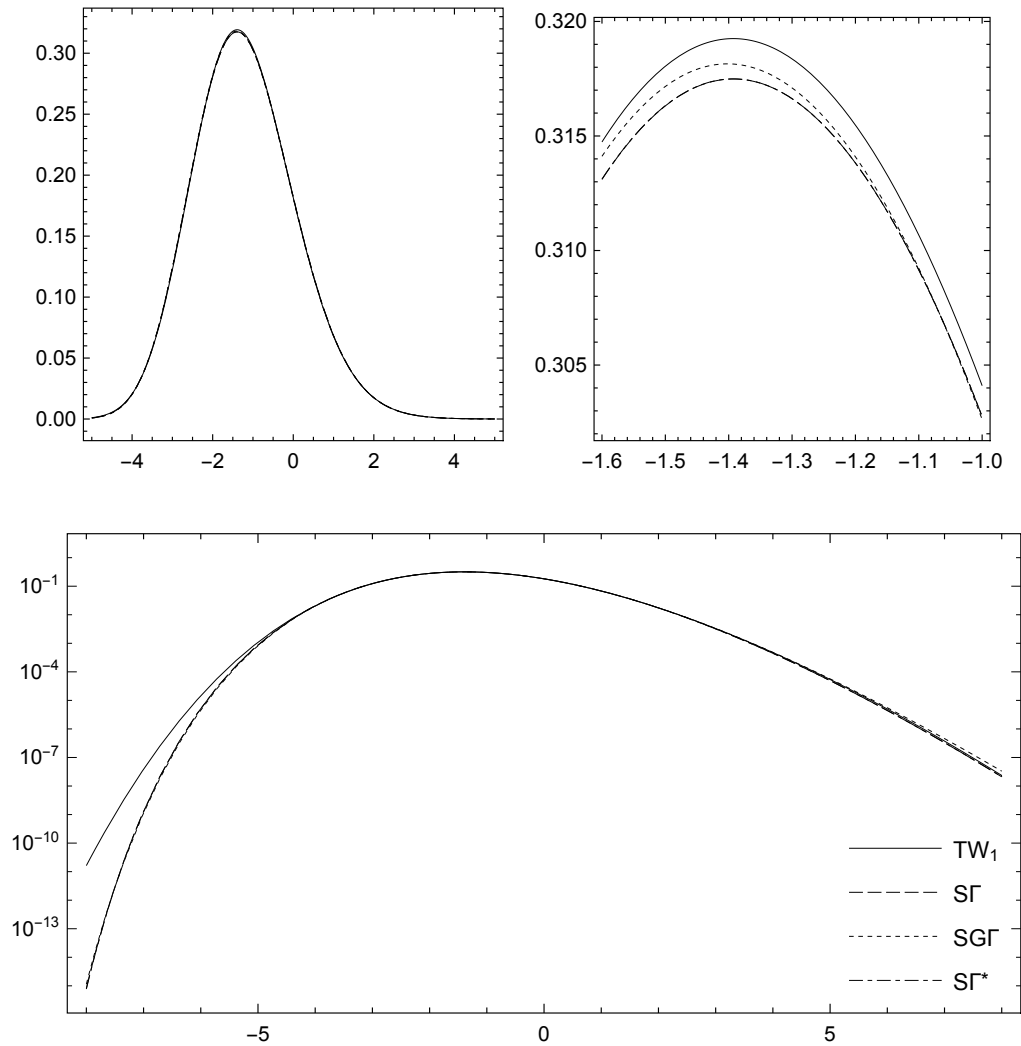
rozdělení	k	θ	β	μ
\mathbf{ST}	46,446	0,186	1	-9,848
\mathbf{ST}^*	46,360	0,186	1	-9,840
\mathbf{SGT}	84,650	0,032	0,783	-10,343

Tabulka 3.2: Parametry gama rozdělení aproximující \mathcal{TW}_1

rozdělení \mathcal{D}	$\mu_{\mathcal{D}}$	$\sigma_{\mathcal{D}}^2$	$\gamma_{3\mathcal{D}}$	$\gamma_{4\mathcal{D}}$
\mathcal{TW}_1	-1,207	1,608	0,293	0,165
ST	-1,207	1,608	0,293	0,129
SGT	-1,208	1,607	0,308	0,151
ST*	-1,207	1,608	0,294	0,129
rozdíl	$\mu_{\mathcal{D}}$	$\sigma_{\mathcal{D}}^2$	$\gamma_{3\mathcal{D}}$	$\gamma_{4\mathcal{D}}$
$\Delta(\mathbf{ST}^*)$	$1,998 \times 10^{-15}$	$-5,773 \times 10^{-14}$	$2,713 \times 10^{-4}$	$-3,582 \times 10^{-2}$
$\Delta_r(\mathbf{ST}^*)$	$1,656 \times 10^{-15}$	$-3,591 \times 10^{-14}$	$9,246 \times 10^{-4}$	$-1,132 \times 10^{-2}$

Tabulka 3.3: Srovnání teoretických charakteristik rozdělení \mathcal{TW}_1 a gama rozdělení, včetně příslušných rozdílů a relativních rozdílů

Jak můžeme sledovat na obrázku 3.4, mezi modely **ST** a **ST*** nelze pozorovat rozdíl pro větší hodnoty hustoty \mathcal{TW}_1 . Po přiblížení oblasti modu pozorujeme, že čtyř-parametrický model **SGT** je v tomto místě blíže \mathcal{TW}_1 než předešlé modely. Na chvostu se odlišují od \mathcal{TW}_1 všechny aproximační modely. Pohledem na obyčejný graf hustoty lze říci, že všechny aproximace jsou velmi úspěšné. V praxi by bylo zřejmě preferované volit model s nejnižším počtem parametrů, tj. **ST**.



Obrázek 3.4: Srovnání grafů hustot rozdělení \mathcal{TW}_1 a aproximujících gama rozdělení. Hustota (vlevo nahoře), hustota v modu (vpravo nahoře) a logaritmické měřítko (dole)

3.4 Rozdělení největšího vlastního čísla bílé Wishartovy matice

V předchozí podkapitole jsme se zabývali asymptotickým rozdělením vlastních čísel výběrové kovarianční matice za předpokladu, že její vlastní čísla jsou navzájem různá. Nyní se podíváme na limitní rozdělení největšího vlastního čísla bílé Wishartovy matice (viz definice 4 v kapitole 1). Připomeňme, že se jedná o náhodnou matici z rozdělení $\mathcal{W}_p(\mathbb{I}_p, n)$. Obecnou Wishartovu matici lze rovněž chápat jako n -násobek výběrové kovarianční matice. Z konstrukce bílé Wishartovy matice je navíc patrné, že zcela dochází k porušení předpokladu různosti vlastních čísel matice reprezentující kovarianční matici náhodného výběru. Taková vlastní čísla jsou totiž všechna rovna jedničkám. Vztahem mezi vlastními

číslly výběrové kovarianční matice a odpovídající Wishartovy matice se zabývá následující tvrzení.

Tvrzení 9. *Předpokládejme, že matice $\mathbb{X} \in \mathbb{R}^{n \times p}$ je vytvořená centrováním náhodným výběrem z rozdělení $\mathcal{N}_p(\mathbf{0}, \Sigma)$. Označme příslušnou výběrovou kovarianční matici \mathbb{S} a Wishartovu matici \mathbb{M} , tj.*

$$\mathbb{S} = \frac{1}{n} \mathbb{X}^\top \mathbb{X}, \quad \mathbb{M} = \mathbb{X}^\top \mathbb{X}.$$

Pak jsou vlastní čísla matice \mathbb{M} rovny n -násobkům vlastních čísel matice \mathbb{S} .

Důkaz. Vlastní čísla matice \mathbb{S} jsou kořeny polynomu s proměnnou λ , který je determinantem matice $\mathbb{S} - \lambda \mathbb{I}_p$, kde \mathbb{I}_p představuje diagonální matici, jež má právě p jedniček na hlavní diagonále. Předpis pro determinant upravíme:

$$|\mathbb{S} - \lambda \mathbb{I}_p| = \left| \frac{1}{n} (\mathbb{M} - n\lambda \mathbb{I}_p) \right| = \left(\frac{1}{n} \right)^p |\mathbb{M} - \tilde{\lambda} \mathbb{I}_p|,$$

kde jsme použili přeznačení $\tilde{\lambda} = n\lambda$. Vlastní čísla nyní získáme položením výše uvedených determinantů jako rovných nule. Odtud je již zřejmé, že se vlastní čísla matice \mathbb{M} rovnají n -násobkům vlastních čísel matice \mathbb{S} . □

Nyní se již podíváme na kýžený poznatek o limitním rozdělení největšího vlastního čísla bílé Wishartovy matice. Následující věta byla převzata z článku Johnstone (2001, str. 300, Theorem 1.1.).

Věta 10. *Nechť jsou prvky matice $\mathbb{X} \in \mathbb{R}^{n \times p}$ navzájem nezávislé a stejně rozdělené s normálním rozdělením $\mathcal{N}(0, 1)$. Nechť vlastní čísla matice $\mathbb{X}^\top \mathbb{X}$ jsou $\hat{\lambda}_1 > \hat{\lambda}_2 > \dots > \hat{\lambda}_p$. Dále definujme následující konstanty:*

$$\begin{aligned} \mu_{n,p} &= \left(\sqrt{n+a} + \sqrt{p+b} \right)^2, \\ \sigma_{n,p} &= \left(\sqrt{n+a} + \sqrt{p+b} \right) \left(\frac{1}{\sqrt{n+a}} - \frac{1}{\sqrt{p+b}} \right)^{1/3}, \end{aligned} \tag{3.7}$$

kde $a = -1$ a $b = 0$. Nechť dále

$$\frac{n}{p} \xrightarrow{n, p \rightarrow \infty} \gamma \geq 1.$$

Pak

$$L_1 = \frac{\hat{\lambda}_1 - \mu_{n,p}}{\sigma_{n,p}} \stackrel{as}{\sim} \mathcal{TW}_1, \quad n, p \rightarrow \infty. \tag{3.8}$$

Důkaz. Viz článek Tracy a Widom (1996). □

Výše uvedenou větu je rovněž možné modifikovat pro případ, kdy prvky matice \mathbb{X} pocházejí z rozdělení $\mathcal{N}(0, \sigma^2)$, tj. mají rozptyl $\sigma^2 > 0$, jenž není nutně jednotkový. Označme \mathbb{X} matici skládající se z prvků s jedničkovým rozptylem

jako ve Větě 10 a definujme matici $\tilde{\mathbb{X}}$ vzniklou vynásobením každého prvku \mathbb{X} směrodatnou odchylkou σ , tj.

$$\tilde{\mathbb{X}} = \sigma \mathbb{X}.$$

Podíváme se na vzorec pro výpočet vlastních čísel matice $\tilde{\mathbb{X}}^\top \tilde{\mathbb{X}}$:

$$\left| \tilde{\mathbb{X}}^\top \tilde{\mathbb{X}} - \tilde{\lambda} \mathbb{I}_p \right| = \left| \sigma^2 \left(\mathbb{X}^\top \mathbb{X} - \frac{1}{\sigma^2} \tilde{\lambda} \mathbb{I}_p \right) \right| = (\sigma^2)^p \left| \mathbb{X}^\top \mathbb{X} - \lambda \mathbb{I}_p \right|,$$

kde jsme použili přeznačení symbolu $\tilde{\lambda}$ pro vlastní číslo matice $\tilde{\mathbb{X}}^\top \tilde{\mathbb{X}}$ na

$$\lambda = \frac{1}{\sigma^2} \tilde{\lambda}$$

značící vlastní číslo matice $\mathbb{X}^\top \mathbb{X}$. Získali jsme tedy vztah mezi vlastními čísly matic z rozdělení $\mathcal{W}_p(\mathbb{I}_p, n)$ a $\mathcal{W}_p(\sigma^2 \mathbb{I}_p, n)$, na základě něhož lze příslušně modifikovat asymptotické rozdělení statistiky L_1 jako

$$\tilde{L}_1 = \frac{(\tilde{\lambda}_1/\sigma^2) - \mu_{n,p}}{\sigma_{n,p}} \stackrel{as}{\approx} \mathcal{TW}_1, \quad n, p \rightarrow \infty \quad (3.9)$$

s označením $\tilde{\lambda}_1$ největšího vlastního čísla matice $\tilde{\mathbb{X}}^\top \tilde{\mathbb{X}}$, která byla sestavena z kombinace prvků s rozptyly σ^2 .

Podobně platí, že pokud bychom chtěli získat asymptotický vztah pro rozdělení statistiky L_1 , kde by vystupovalo přímo největší vlastní číslo výběrové kovarianční matice

$$\mathbb{S} = \frac{1}{n} \mathbb{X}^\top \mathbb{X}$$

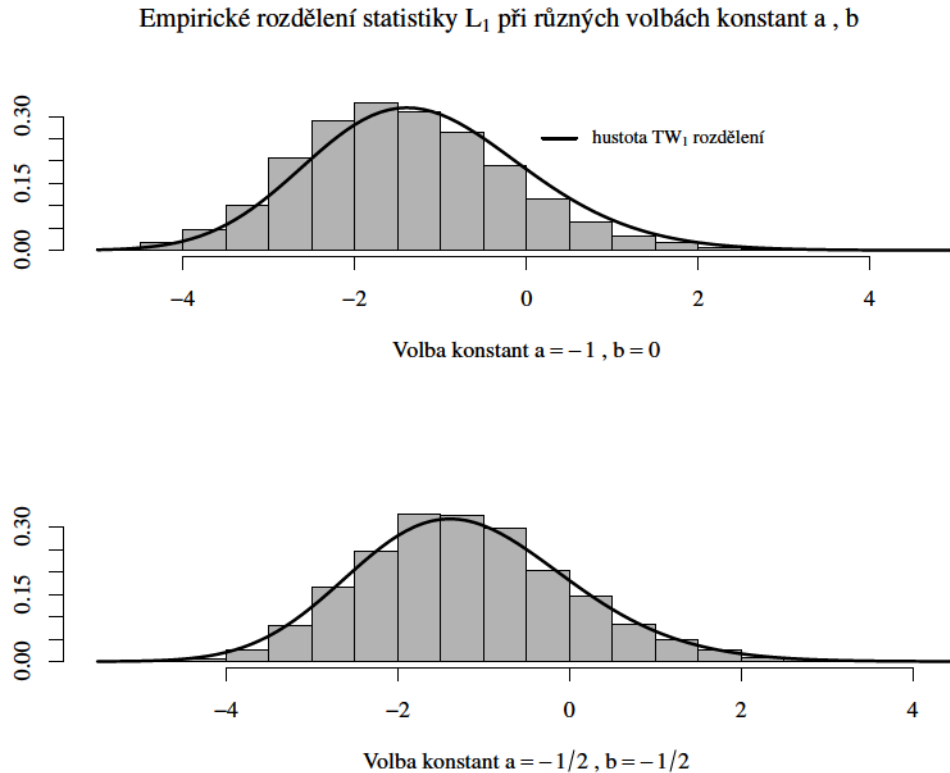
označené jako $\hat{\lambda}_1$, tak by stačilo konstanty $\mu_{n,p}$ a $\sigma_{n,p}$ vydělit n . Jinak řečeno, závěr Věty 10 by měl podobu

$$L_1 = \frac{\hat{\lambda}_1 - \mu_{n,p}/n}{\sigma_{n,p}/n} \stackrel{as}{\approx} \mathcal{TW}_1, \quad n, p \rightarrow \infty.$$

Výsledný asymptotický vztah (3.8) ve Větě (10) rovněž platí, pokud $n < p$, přičemž n, p jsou dostatečně velká, a v tomto případě se musí zaměnit význam n a p ve vzorcích (3.7) (Johnstone (2001, str. 300, Theorem 1.1.)). Ačkoliv Věta 10 platí pro limitní případ, Tracyho-Widomovo rozdělení poskytuje dobrou aproximaci rozdělení statistiky již pro rozměry matic n, p blízké číslu 10 (Vlok a Olivier (2012, str. 1805)). Větu je dále rovněž možné aplikovat na několik zobecněných případů, co se tvaru matice \mathbb{X} týče. Přehled o těchto zobecněních uvádí autoři v článku Saccenti a Camacho (2015, str. 102).

Pozastavme se ještě nad tvarem výrazů $\mu_{n,p}$ a $\sigma_{n,p}$. Na rozdíl od výše uvedené věty doporučují autoři v článcích Chiani (2014) a Ma (2012) volbu konstant $a = b = -\frac{1}{2}$, která pak vede k lepší asymptotické aproximaci rozdělení. Při této volbě se rozdíl mezi rozdělením statistiky L_1 ze vzorce (3.8) a Tracyho-Widomovým rozdělením řádu 1 redukuje na $O((n \wedge p)^{-2/3})$ oproti původnímu $O((n \wedge p)^{-1/3})$, kde $a \wedge b = \min\{a, b\}$ (viz Ma, 2012, str. 323). Z tohoto důvodu budeme i v další části práce používat tuto podobu konstant vedoucí k lepší aproximaci. Představu o „úspěšnosti“ jednotlivých aproximací můžeme získat na základě histogramů na obrázku 3.5. Skutečně se zdá, že hustota \mathcal{TW}_1 rozdělení přiléhá ke znázorněné

distribuci statistiky L_1 mnohem lépe při volbě konstant $a = b = -\frac{1}{2}$ než při předchozí volbě $a = -1, b = 0$.



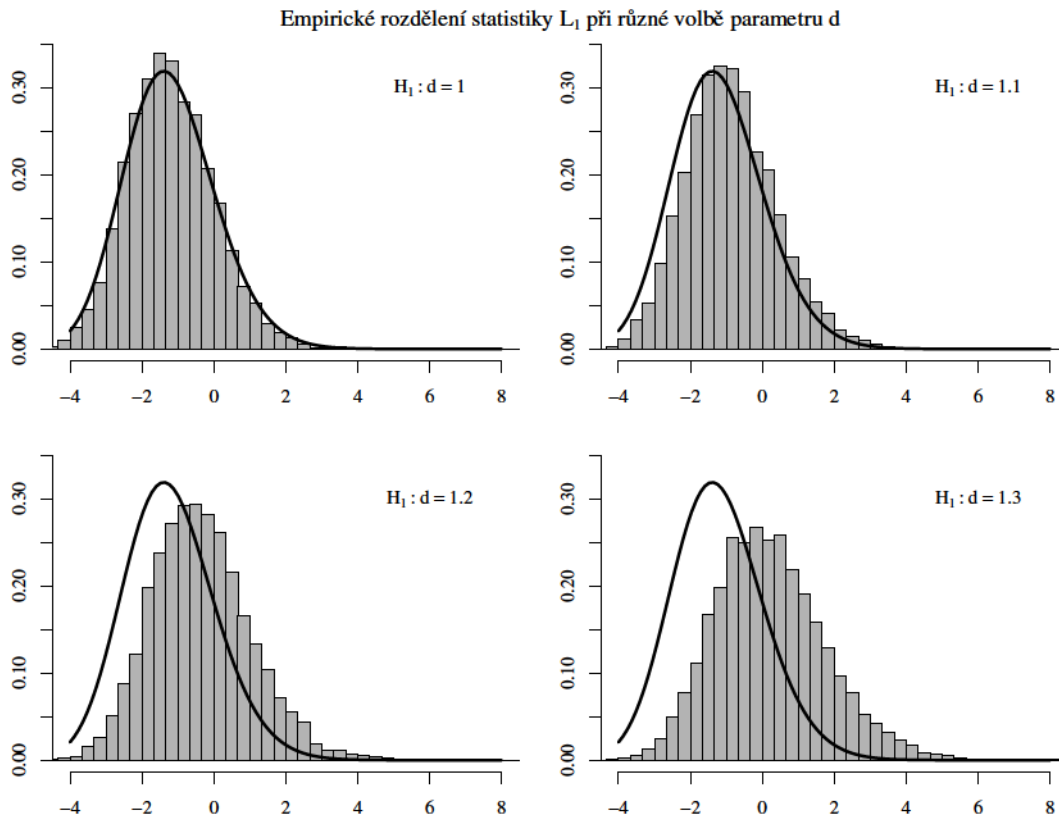
Obrázek 3.5: Srovnání histogramů statistiky L_1 s různými volbami konstant a, b a znázorněnou hustotou \mathcal{TW}_1 rozdělení. Histogramy byly vykresleny na základě nastavení $n = 100, p = 5$ a 10 000 realizací.

Podle jednoho z předpokladů Věty (10) má být matice \mathbb{X} tvořena n realizacemi náhodného vektoru z rozdělení $\mathcal{N}_p(\mathbf{0}, \mathbb{I}_p)$. O tom, že je tento předpoklad skutečně podstatný, se přesvědčíme na základě histogramů na obrázku 3.6. Ty představují distribuci statistiky L_1 , k jejímuž sestavení byly použity realizace náhodného vektoru z rozdělení $\mathcal{N}_p(\mathbf{0}, \Sigma_1)$, kde Σ_1 představuje diagonální matici s vektorem $(d, 1, \dots, 1)^\top$ na hlavní diagonále. Ten je tvořen číslem d na první pozici a $p - 1$ jedničkami na zbylých pozicích. Z obrázku je tedy patrné, že čím více se hodnota d odlišuje od jedničky, tím více se i empirické rozdělení statistiky L_1 vzdaluje od tvaru hustoty \mathcal{TW}_1 rozdělení.

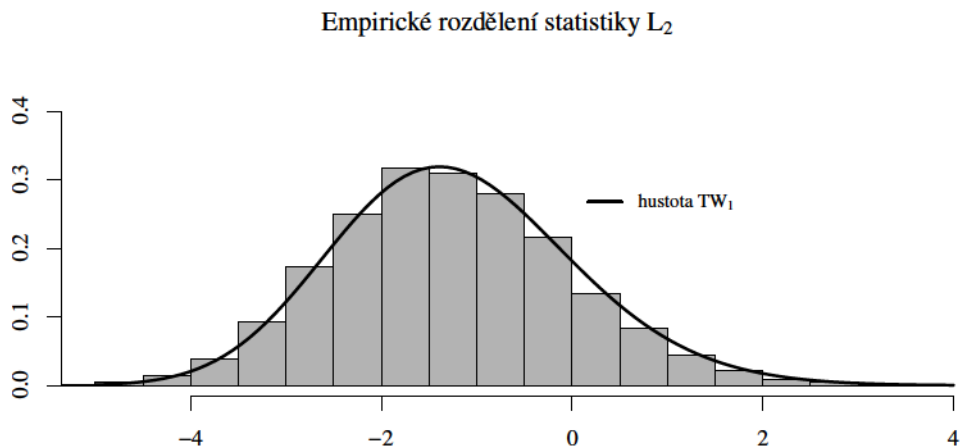
Avšak kvůli asymptotické nezávislosti navzájem různých výběrových vlastních čísel máme „podezření“, že by pro asymptotické rozdělení následující statistiky mohlo platit

$$L_2 = \frac{\hat{\lambda}_2 - \mu_{n,p-1}}{\sigma_{n,p-1}} \stackrel{as}{\sim} \mathcal{TW}_1, \quad n, p \rightarrow \infty,$$

kde $\hat{\lambda}_2$ je druhé největší vlastní číslo matice z rozdělení $\mathcal{W}_p(\Sigma_1, n)$, s maticovým parametrem $\Sigma_1 = \text{diag}(d, 1, \dots, 1)$ (definovaným stejným způsobem jako výše). Na základě porovnání histogramu s hustotou rozdělení \mathcal{TW}_1 na obrázku 3.7 usuzujeme, že by asymptotický vztah skutečně mohl platit.



Obrázek 3.6: Porovnání histogramů statistiky L_1 s různými volbami parametru d kovarianční matice, vykreslena je rovněž hustota \mathcal{TW}_1 rozdělení. Histogramy byly vykresleny na základě nastavení $n = 100$, $p = 25$ a 10 000 realizací.



Obrázek 3.7: Histogram statistiky L_2 s vykreslenou hustotou Tracyho-Widomova rozdělení řádu 1. Použili jsme nastavení $d = 5$, $p = 10$, $n = 1000$ a 10 000 realizací.

Doplňme, že na základě Věty 10 platí vztah:

$$\lim_{n,p \rightarrow \infty} \mathbb{P} \left\{ \frac{\hat{\lambda}_1 - \mu_{n,p}}{\sigma_{n,p}} < q_{TW}(1 - \alpha) \right\} = 1 - \alpha, \quad (3.10)$$

kde $q_{TW}(1 - \alpha)$ je $1 - \alpha$ kvantil rozdělení \mathcal{TW}_1 .

3.5 Asymptotické rozdělení nejmenšího vlastního čísla

V této kapitole jsme se podrobně věnovali asymptotickým vlastnostem největšího vlastního čísla bílé Wishartovy matice. Abychom si vytvořili kompletní představu o asymptotickém rozdělení extrémních hodnot vlastních čísel, zaměřme se ještě krátce na charakteristiky limitního rozdělení nejmenšího vlastního čísla. Nejprve však uveďme definici rozdělení, které bude pro uvedené asymptotické vlastnosti klíčové.

Definice 9. (*Zrcadlené Tracyho-Widomovo rozdělení*) Zrcadlené Tracyho-Widomovo rozdělení řádu 1 (*reflected Tracy-Widom distribution*, v textu značíme \mathcal{RTW}_1) je definováno distribuční funkcí $G_1(x)$, jež splňuje rovnost

$$G_1(x) = 1 - F_1(-x), \quad x \in \mathbb{R},$$

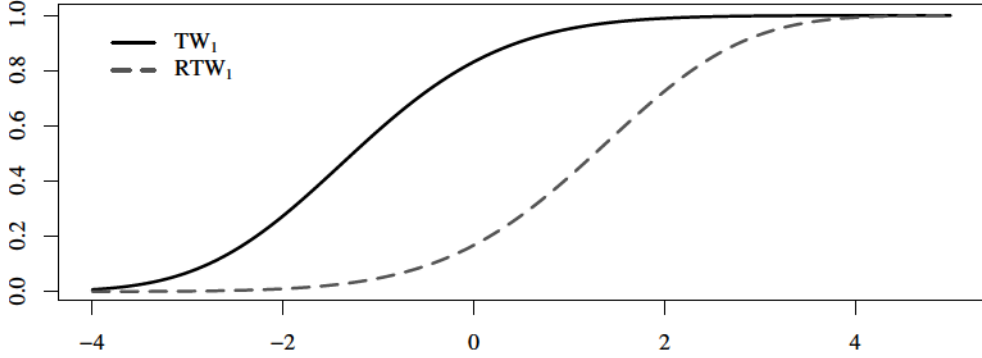
kde $F_1(x)$ představuje distribuční funkci rozdělení \mathcal{TW}_1 .

Označíme-li $f_1(x)$ hustotu rozdělení \mathcal{TW}_1 , pak derivací distribuční funkce dostaneme velmi snadno předpis hustoty \mathcal{RTW}_1 rozdělení $g_1(x)$:

$$g_1(x) = \frac{d}{dx} G_1(x) = \frac{d}{dx} (1 - F_1(-x)) = f_1(-x).$$

Hustota zrcadleného Tracyho-Widomova rozdělení je tedy osově souměrná podle osy y s hustotou své původní verze. Představu o tom, jak vypadá distribuční funkce \mathcal{RTW}_1 rozdělení v porovnání s původní verzí, si lze vytvořit na základě obrázku 3.8.

Distribuční funkce rozdělení \mathcal{TW}_1 a \mathcal{RTW}_1



Obrázek 3.8: Distribuční funkce rozdělení \mathcal{TW}_1 a \mathcal{RTW}_1

Asymptotické rozdělení nejmenšího vlastního čísla nám přiblíží Věta 11, která představuje jistou analogii Věty 10.

Věta 11. *Předpokládejme, že jsou prvky matice $\mathbb{X} \in \mathbb{R}^{n \times p}$ navzájem nezávislé a stejně rozdělené s normálním rozdělením $\mathcal{N}(0, 1)$. Nechť vlastní čísla matice $\mathbb{X}^\top \mathbb{X}$ jsou $\hat{\lambda}_1 > \hat{\lambda}_2 > \dots > \hat{\lambda}_p$. Dále definujme následující konstanty:*

$$\begin{aligned} \mu_{n,p}^- &= \left(\sqrt{n - \frac{1}{2}} - \sqrt{p - \frac{1}{2}} \right)^2, \\ \sigma_{n,p}^- &= \left(\sqrt{n - \frac{1}{2}} - \sqrt{p - \frac{1}{2}} \right) \left(\frac{1}{\sqrt{p - \frac{1}{2}}} - \frac{1}{\sqrt{n - \frac{1}{2}}} \right)^{1/3}, \\ \tau_{n,p}^- &= \frac{\sigma_{n,p}^-}{\mu_{n,p}^-}, \quad v_{n,p}^- = \log(\mu_{n,p}^-) + \frac{1}{8}(\tau_{n,p}^-)^2. \end{aligned} \quad (3.11)$$

Předpokládejme, že

$$\frac{n}{p} \xrightarrow{n, p \rightarrow \infty} \gamma > 1.$$

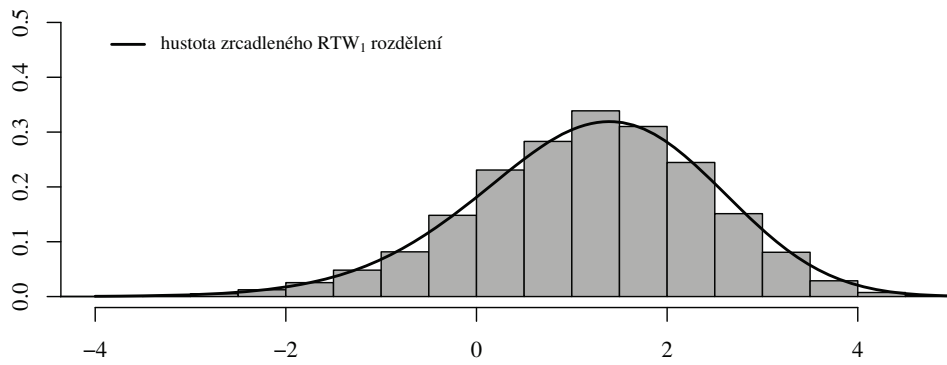
Pak platí

$$R_1 = \frac{\log(\hat{\lambda}_p) - v_{n,p}^-}{\tau_{n,p}^-} \stackrel{as}{\approx} \mathcal{RTW}_1, \quad n, p \rightarrow \infty. \quad (3.12)$$

Důkaz. Viz Ma (2012, str. 325, Theorem 2) pro znění věty. Důkaz pro p sudé viz Ma (2012, str. 341-345, Theorem 2), dle autora článku aproximace též dobře funguje pro lichý případ, což je doloženo simulacemi (Ma, 2012, str. 325). \square

Představu o aproximaci rozdělení statistiky G_1 rozdělením \mathcal{RTW}_1 nabízí histogram na obrázku 3.9 níže. Pozorujeme, že hustota zrcadleného Tracyho–Widomova rozdělení celkem dobře kopíruje tvar histogramu.

Empirické rozdělení statistiky R_1



Obrázek 3.9: Histogram statistiky R_1 se znázorněnou hustotou zrcadleného Tracy–Widom rozdělení řádu 1. Použili jsme nastavení $n = 100$, $p = 10$ a 10 000 realizací.

4. Volba počtu hlavních komponent založená na Tracyho-Widomově rozdělení

Požadavek na bílou Wishartovu matici je často velmi omezující, a proto uvažujeme kovarianční matici

$$\Sigma_r = \text{diag}(\lambda_1, \dots, \lambda_r, 1, \dots, 1)_p,$$

která je diagonální, typu $p \times p$, s prvními r prvky většími než jedna a $r - p$ jedničkami na zbylých pozicích na hlavní diagonále. Na rozdíl od předpokladů Věty 10 tedy vycházíme ze situace, kdy vektory řádků matice \mathbb{X} pocházejí z rozdělení $\mathcal{N}_p(\mathbf{0}, \Sigma_r)$ s poněkud obecnějším tvarem kovarianční matice. Nyní si představíme statistické testy na počet nejednotkových vlastních čísel matice Σ_r a popíšeme spojitost s problematikou volby optimálního počtu hlavních komponent.

4.1 Test přítomnosti právě jednoho nejednotkového vlastního čísla

Nejprve test popíšeme v případě, že $r = 1$. Začneme uvedením předpokladů testu. Necht $\mathbb{X}^\top \mathbb{X} \sim \mathcal{W}_p(\Sigma_1, n)$, $\Sigma_1 = \text{diag}(\lambda_1, 1, \dots, 1)_p$, kde $\lambda_1 \geq 1$. Testujeme hypotézu

$$H_0 : \lambda_1 = 1$$

oproti alternativě

$$H_1 : \lambda_1 > 1.$$

Testová statistika má tvar

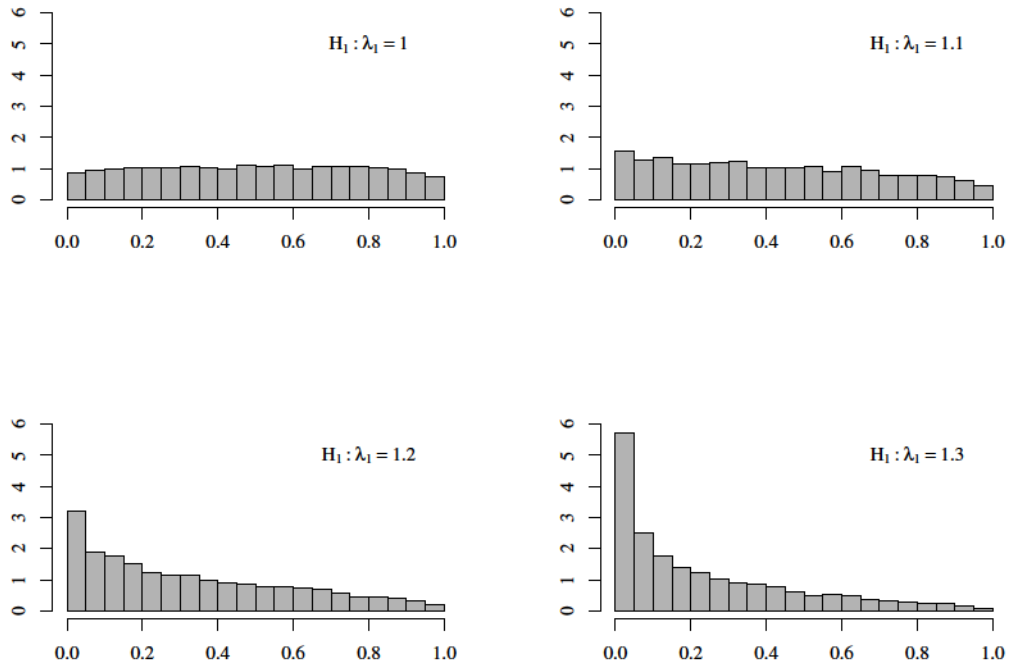
$$T_1 = \frac{\hat{\lambda}_1 - \mu_{n,p}}{\sigma_{n,p}},$$

který odkazuje na vztah (3.10). Nulovou hypotézu zamítneme na hladině α , pokud $T_1 > q_{TW}(1 - \alpha)$, kde $q_{TW}(1 - \alpha)$ značí $1 - \alpha$ kvantil rozdělení \mathcal{TW}_1 . Tedy pro vysoké hodnoty testové statistiky T_1 . P-hodnota testu je

$$1 - F_{TW}(t_1),$$

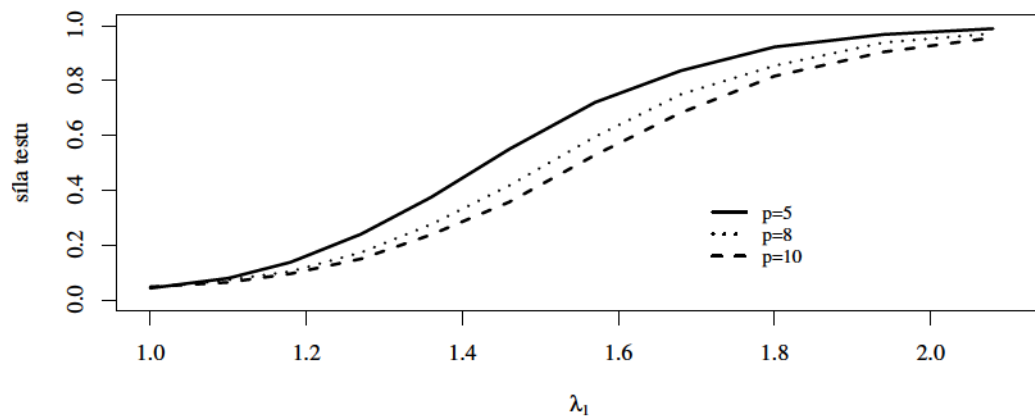
kde symbol $F_{TW}(t_1)$ značí hodnotu distribuční funkce rozdělení \mathcal{TW}_1 v bodě t_1 , což je realizovaná hodnota testové statistiky. Histogramy na obrázku 4.1 představují empirické rozdělení p-hodnoty výše uvedeného testu pro různé volby hodnot λ_1 . Na histogramu si můžeme všimnout, že pokud zvolíme $\lambda_1 = 1$, tj. data přesně odpovídají předpokladům Věty 10, rozdělení p-hodnoty lze přibližně klasifikovat jako rovnoměrné na intervalu $(0, 1)$, což bychom očekávali. Čím více se pak bude hodnota λ_1 vzdalovat od jedničky, tím spíše se bude rozdělení p-hodnoty představě rovnoměrnosti vymykat.

Empirické rozdělení p-hodnoty



Obrázek 4.1: Empirické rozdělení p-hodnoty pro test hypotézy $H_0 : \lambda_1 = 1$ s různými volbami parametru λ_1 kovarianční matice. Histogramy byly vykresleny na základě nastavení $n = 100$, $p = 25$ a 10 000 realizací.

Empirická síla testu



Obrázek 4.2: Znázornění empirické síly testu pro různé volby λ_1 a p , $n = 100$ a s počtem realizací 10 000. Výpočty byly vztaženy ke zvolené hladině $\alpha = 0,05$.

Dalším důležitým nástrojem pro zhodnocení kvality statistického testu je síla testu (sílofunkce), jež vyjadřuje pravděpodobnost, že nulovou hypotézu zamítneme za platnosti alternativy. Z grafů na obrázku 4.2 uvedeném níže, který zobrazuje empirickou sílofunkci pro λ_1 z rozmezí $(1, 2)$ a několik p při pevném $n = 100$, můžeme odvodit následující pozorování. Čím je vyšší p , tím při zvyšujícím se λ_1 síla roste pomaleji, než když je p nižší. Důvod je takový, že se největší vlastní číslo počítá z více odhadovaných vlastních čísel, a tudíž při vyšším p je četnější situace, že je maximální vlastní číslo vyšší. Tudíž se požaduje, aby test byl tolerantnější ohledně nezamítání nulové hypotézy.

4.2 Sekvenční verze testu

Nyní se budeme zabývat obecnějším případem, kdy předpokládáme, že

$$\mathbf{X}^\top \mathbf{X} \sim \mathcal{W}_p(\boldsymbol{\Sigma}_K, n),$$

$$\boldsymbol{\Sigma}_K = \text{diag}(\lambda_1, \dots, \lambda_K, 1, \dots, 1)_p,$$

kde $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_K \geq 1, 1 \leq K < p$. Popíšeme si proceduru pro testování počtu vlastních čísel větších než jedna. Jinými slovy, procedura bude sloužit pro určení počtu nejednotkových (můžeme rovněž říkat „signifikantních“) vlastních čísel K . Testují se tedy dílčí hypotézy

$$H_0 : \lambda_k = 1, \quad k = 1, \dots, p - 1,$$

oproti alternativám

$$H_1 : \lambda_k > 1.$$

Příslušný algoritmus má následující schéma:

1. Pro $k = p - 1, \dots, 1$ spočítáme hodnotu testové statistiky

$$T_k = \frac{\hat{\lambda}_k - \mu_{n, p-k+1}}{\sigma_{n, p-k+1}}$$

2. Zvolíme hladinu α pro dílčí testy a určíme největší index k^* pro prvních k^* testů, jejichž testové statistiky T_1, \dots, T_{k^*} dosáhnou vyšších hodnot, než je hodnota kvantilu $q_{TW}(1 - \alpha)$, tj. dochází k zamítnutí nulových hypotéz. Jinak řečeno, k^* se shoduje s počtem dílčích testů, pro něž p-hodnota

$$1 - F_{TW}(t_i),$$

kde $t_i, 1 \leq i \leq k^*$, jsou realizace testových statistik T_1, \dots, T_{k^*} , je menší než stanovená hladina α .

3. Odhadovaný počet vlastních čísel větších než jedna pak stanovíme jako

$$\hat{K} = k^*.$$

Poznamenejme, že α není celkovou hladinou sekvenčního testu.

Princip, na kterém je procedura založena, dokresluje příklad shrnutý v tabulce 4.1. Jedná se o testování počtu vlastních čísel přesahujících jedničku, jež přísluší matici z rozdělení $\mathcal{W}_7(\Sigma_3, 100)$ s tvarem matice $\Sigma_3 = \text{diag}(8, 5, 4, 1, 1, 1, 1)_7$. Tabulka pro každý z $k = 1, \dots, 6$ dílčích testů shrnuje přibližnou hodnotu výběrového vlastního čísla, konstant vystupujících v testové statistice i testovou statistiku samotnou. Součástí jsou rovněž p-hodnoty, zapsané numericky i symbolicky pro větší přehlednost. Největší index k , pro který je p-hodnota testů menší než stanovená hladina $\alpha = 0,05$, přísluší v pořadí 3. dílčímu testu, a proto je počet nejednotkových vlastních čísel na základě sekvenčního testu stanoven jako 3. Tento výsledek koresponduje se skutečností.

k	$\hat{\lambda}_k$	$\mu_{n,p-k+1}$	$\sigma_{n,p-k+1}$	T_k	p-hodnota
1	782,91	156,86	9,89	63,30	***
2	538,26	151,79	9,95	38,84	***
3	363,88	146,32	10,04	21,67	***
4	119,06	140,32	10,18	-2,09	0,7502
5	111,51	133,54	10,42	-2,11	0,7571
6	89,90	125,43	10,88	-3,27	0,9577

Tabulka 4.1: Test počtu vlastních čísel větších než jedna pro data z mnohorozměrného normálního rozdělení s parametry $n = 100$, $p = 7$, kovarianční maticí $\Sigma_3 = \text{diag}(8, 5, 4, 1, 1, 1, 1)$ a volbou hladiny dílčích testů jako $\alpha = 0,05$.

4.3 Shrnutí testů založených na Tracyho-Widomově rozdělení

V případě, že jsou hlavní komponenty spočteny jako lineární kombinace (téměř) nekorelovaných náhodných veličin, metody pro volbu počtu hlavních komponent se jeví být nespolehlivé (např. poznámka k pravidlu na základě broken stick modelu v kapitole 2) a obecně se v tomto případě optimální počet hlavních komponent definuje jako nula. Tj. bylo by nutné uvažovat všechny hlavní komponenty, resp. všechny původní náhodné veličiny, což by nevedlo k žádané projekci dat do prostoru nižší dimenze.

Testy založené na Tracyho-Widomově rozdělení nabízejí za určitých podmínek alternativní řešení tohoto problému. Detekuje se, jestli se skutečně jedná o problematický případ – v případě prvního testu by se dalo určit, jestli bude alespoň první hlavní komponenta vysvětlovat větší míru variability než zbylé ostatní. Pokud ano, optimálním počtem hlavních komponent by byla jednička – zvolili bychom první hlavní komponentu. Na druhou stranu sekvenční test pomůže detekovat posloupnost několika prvních hlavních komponent, jež vysvětlují více variability než zbylé hlavní komponenty. Optimálním počtem je pak počet těchto několika prvních „dominantních“ hlavních komponent.

Rovněž podotkněme, že předpoklad bílé Wishartovy matice ve Větě 10, a tudíž i požadavek na jednotková vlastní čísla v odvozených testech není zcela striktní.

Pokud bychom jedničky nahradili jinou kladnou konstantou, bylo by třeba testové statistiky upravit pomocí odhadu této konstanty. Tímto případem se zabývají autoři článku Saccanti a Camacho (2015, str. 102). Jiný přístup nabízí využití vztahu (3.9) z kapitoly 3, kdy by se odhady vlastních čísel v testových statistikách nahradily konzistentním odhadem konstanty σ^2 .

4.4 Souvislost s modelem broken stick

Mezi heuristickými metodami jsme v rámci kapitoly 2 zmínili i tzv. broken stick model. Využijeme předchozí poznatky a na základě broken stick modelu odvodíme statistický test pro rovnost délek všech částí. Uvažujeme tedy rovnoměrné dělení intervalu $(0, 1)$ na p částí, přičemž délky těchto částí označíme jako V_1, \dots, V_p a nejdelší z nich je $V_{(1)}$. Část s nejdelší délkou bude rovněž představovat testovou statistiku. Dále testujeme hypotézu

$$H_0 : V_1 = \dots = V_p$$

oproti alternativě

$$H_1 : V_{(1)} > V_i \quad \text{pro nějaké } i \in \{1, \dots, p\}.$$

Hypotézu zamítneme na hladině α , pokud

$$v_{(1)} > q_{BS}(1 - \alpha),$$

kde $v_{(1)}$ je realizovaná hodnota náhodné veličiny $V_{(1)}$ a $q_{BS}(1 - \alpha)$ značí hodnotu $1 - \alpha$ empirického kvantilu rozdělení nejdelší části v broken stick modelu. Hodnoty empirických kvantilů pro dělení úseku na $p = 5$ částí pozorujeme v tabulce 4.2. Pokud bychom tedy zvolili hladinu testu 5% a uvažovali bychom model s dělením na 5 částí, pak by kritickou hodnotou bylo číslo 0,6858.

Pro úplnost jsou v tabulce 4.2 rovněž uvedeny příslušné kvantily \mathcal{TW}_1 rozdělení použité v předchozích testech. Vzhledem ke skutečnosti, že \mathcal{TW}_1 rozdělení přísluší asymptotickým vlastnostem a jeho nosič je reálný, zatímco rozdělení broken stick modelu se vztahuje k přesnému rozdělení, není příliš vhodné kvantily porovnávat přímo. Navíc broken stick rozdělení uvažuje dělení celkové variability normované na určitou konstantu (obvykle délka 1), zatímco \mathcal{TW}_1 obecně popisuje asymptotické rozdělení upraveného největšího vlastního čísla (za jistých předpokladů) bez požadavku, aby se upravená vlastní čísla sečetla na stanovenou konstantu.

Test na základě broken stick modelu lze rovněž chápat jako test rovnosti vlastních čísel, jenž ve srovnání s testy na základě \mathcal{TW}_1 rozdělení nevyžaduje splnění množství omezujících předpokladů. Místo statistik V_i by se dosadil podíl variability

$$\frac{\hat{\lambda}_i}{\sum_{k=1}^p \hat{\lambda}_k}$$

v případě výběrové kovarianční matice (pokud bychom pracovali s korelační maticí, dělili bychom pouze číslem p). Hypotéza a alternativa by měly tvar

$$H_0 : \lambda_1 = \dots = \lambda_p,$$

$$H_1 : \lambda_1 > \lambda_i \quad \text{pro nějaké } i \in \{1, \dots, p\},$$

kde λ_1 opět značí největší vlastní číslo. Rovněž poznamenejme, že na rozdíl od sekvenčního testu testem na základě broken stick modelu nezjistíme počet různých vlastních čísel v případě, že taková vlastní čísla existují.

	$p = 5$		$p = 10$	
$1 - \alpha$	$q_{BS}(1 - \alpha)$	$q_{BS}(1 - q)$	$q_{TW}(1 - \alpha)$	
0,99	0,7828	0,5347	2,0233	
0,95	0,6858	0,4428	0,9793	
0,90	0,6229	0,3985	0,4501	
0,85	0,5839	0,3729	0,1038	
0,80	0,5539	0,3527	-0,1653	
0,75	0,5273	0,3358	-0,3920	
0,70	0,5050	0,3210	-0,5923	
0,65	0,4850	0,3092	-0,7752	
0,60	0,4674	0,2974	-0,9463	
0,55	0,4521	0,2870	-1,1098	
0,50	0,4379	0,2774	-1,2686	
0,45	0,4235	0,2689	-1,4254	
0,40	0,4097	0,2599	-1,5828	
0,30	0,3827	0,2432	-1,9104	
0,20	0,3547	0,2257	-2,2832	
0,10	0,3219	0,2061	-2,7824	

Tabulka 4.2: Empirické hodnoty kvantilů $q_{BS}(1 - \alpha)$ pro příslušná q vypočtené na základě volby $p = 5$ a $p = 10$ dělení úseků a 10 000 realizací, uvedeny jsou rovněž odpovídající hodnoty kvantilů rozdělení \mathcal{TW}_1 .

5. Přehled dalších metod

Zaměříme se na další používané metody volby optimálního počtu hlavních komponent, které využívají pokročilých výpočetních procedur. Zejména se bude jednat o přístupy založené na *křížovém ověřování* a *bayesovských modelech*. Oba typy metod zpravidla předpokládají platnost následujícího modelu:

$$\mathbb{X} = \mathbb{W}^T \mathbb{H}^T + \mathbb{M} + \mathbb{E}, \quad (5.1)$$

kde $\mathbb{X} \in \mathbb{R}^{n \times d}$ je datová matice, $\mathbb{M} \in \mathbb{R}^{n \times d}$ značí matici středních hodnot s hodnotí jedna (jedná se o n -krát zopakovaný řádkový vektor středních hodnot jednotlivých veličin), $\mathbb{H} \in \mathbb{R}^{k \times d}$, $\mathbb{W} \in \mathbb{R}^{d \times n}$ a $\mathbb{E} \in \mathbb{R}^{n \times d}$ je matice chybových složek, jež v sobě zahrnují jak chyby měření, tak chyby modelu. Předpokládáme, že $k \leq \min\{n-1, d-1\}$ a že prvky \mathbb{E} jsou nezávislé stejně rozdělené s normálním rozdělením s nulovou střední hodnotou a rozptylem $v > 0$. Obvykle se také požaduje nezávislost prvků \mathbb{E} s prvky matice \mathbb{W} . Cílem je odhad optimálního počtu hlavních komponent k .

5.1 Křížové ověřování

Metoda založená na křížovém ověřování (*cross-validation*, CV) provádí aproximaci hodnot části datové matice $\mathbb{X} \in \mathbb{R}^{n \times d}$ (může se jednat např. o hodnotu prvku x_{ij} či hodnoty prvků řádku \mathbf{x}_i) na základě její podmatice neobsahující aproximovanou část. Pomocí stanoveného kritéria se pak určí úspěšnost aproximace, přičemž toto kritérium je funkcí uvažovaného počtu komponent. Za optimální je pak obvykle prohlášen ten počet komponent, po jehož zvýšení se již hodnota kritéria výrazně nezlepší.

Mechanismus blíže popíšeme v případě odhadování jednotlivých prvků matice \mathbb{X} (*leave-one-out cross-validation*) pomocí singulárního rozkladu, přičemž budeme vycházet z Jolliffe (2002, str.121). Provedeme aproximaci pozorování x_{ij} . Aplikujeme-li singulární rozklad na matici \mathbb{X} , můžeme podle vzorce (1.1) prvek x_{ij} zapsat jako

$$x_{ij} = \sum_{k=1}^r u_{ik} l_k v_{jk}, \quad i = 1, \dots, n, \quad j = 1, \dots, d, \quad (5.2)$$

kde r označuje hodnost matice \mathbb{X} a u_{ik}, v_{jk} jsou prvky levých a pravých singulárních vektorů. Dále připomeňme, že singulární čísla l_k mají význam odmocněných vlastních čísel matice $\mathbb{X}^T \mathbb{X}$. Hodnotu pozorování x_{ij} lze aproximovat tak, že v rozvoji (5.2) použijeme pouze prvních $m < r$ hlavních komponent, tj.

$$\tilde{x}_{ij}(m) = \sum_{k=1}^m u_{ik} l_k v_{jk}. \quad (5.3)$$

Avšak k výpočtu výrazu na pravé straně vzorce (5.3) již byla použita informace ze všech prvků původní matice \mathbb{X} , a proto by byl takový postup pro křížové ověřování nevhodný. Z tohoto důvodu se aproximace provede jiným způsobem –

členy rozvoje (5.3) spočítáme z vhodné podmnožiny dat matice \mathbb{X} (specifikujeme níže):

$$\hat{x}_{ij}(m) = \sum_{k=1}^m \hat{u}_{ik} \hat{l}_k \hat{v}_{jk}. \quad (5.4)$$

Poznamenejme, že autoři v článku Bro a kol. (2008, str.1244) doporučují zvolit znaménko aproximace ve vzorci (5.4) shodné jako v případě zkrácené verze klasického singulárního rozkladu ze vzorce (5.3). V tomto případě bude mít aproximovaný prvek tvar

$$\hat{x}_{ij}(m) = \text{sgn}(\tilde{x}_{ij}(m)) \sum_{k=1}^m \hat{u}_{ik} \hat{l}_k \hat{v}_{jk}. \quad (5.5)$$

Kritérium metody je odvozené od součtu čtverců rozdílů aproximovaných hodnot od skutečných hodnot prvků (*prediction sum of squares*)

$$\text{PRESS}(m) = \sum_{i=1}^n \sum_{j=1}^d (x_{ij} - \hat{x}_{ij}(m))^2, \quad (5.6)$$

např. se může jednat o průměr předchozí verze (*mean prediction sum of squares*)

$$\text{MPRESS}(m) = \frac{1}{nd} \sum_{i=1}^n \sum_{j=1}^d (x_{ij} - \hat{x}_{ij}(m))^2, \quad (5.7)$$

který je ještě nějakým způsobem upraven. Vhodný počet hlavních komponent je pak zvolen jako takové nejmenší číslo m^* , po jehož zvýšení se již hodnota kritéria významně nesníží.

K výpočtu prvků rozvoje (5.4) se přistupuje různě. Např. v článku Eastment a Krzanowski (1982) uvádějí následující postup. Označme jako $\mathbb{X}_{(-i\bullet)} \in \mathbb{R}^{(n-1) \times d}$ matici, která vznikla z matice \mathbb{X} odstraněním i -tého řádku. Provedeme singulární rozklad této matice

$$\mathbb{X}_{(-i\bullet)} = \mathbb{U}_{(-i\bullet)} \mathbb{L}_{(-i\bullet)} \mathbb{V}_{(-i\bullet)}^\top,$$

kde $\mathbb{U}_{(-i\bullet)} \in \mathbb{R}^{(n-1) \times (n-1)}$, $\mathbb{L}_{(-i\bullet)} \in \mathbb{R}^{(n-1) \times d}$, $\mathbb{V}_{(-i\bullet)} \in \mathbb{R}^{d \times d}$,

a hodnoty \hat{v}_{jk} ze vzorce (5.4) získáme jako odpovídající počet prvků j -tého řádku matice $\mathbb{V}_{(-i\bullet)}$. Podobně použijeme singulární rozklad pro matici $\mathbb{X}_{(-\bullet j)} \in \mathbb{R}^{n \times (d-1)}$ s vynechaným j -tým sloupcem

$$\mathbb{X}_{(-\bullet j)} = \mathbb{U}_{(-\bullet j)} \mathbb{L}_{(-\bullet j)} \mathbb{V}_{(-\bullet j)}^\top,$$

kde $\mathbb{U}_{(-\bullet j)} \in \mathbb{R}^{n \times n}$, $\mathbb{L}_{(-\bullet j)} \in \mathbb{R}^{n \times (d-1)}$, $\mathbb{V}_{(-\bullet j)} \in \mathbb{R}^{(d-1) \times (d-1)}$,

za hodnoty \hat{u}_{jk} ve vzorci (5.4) dosadíme odpovídající počet prvků i -tého řádku matice $\mathbb{U}_{(-\bullet j)}$. Singulární čísla \hat{l}_k čísla získáme kombinací informací z obou singulárních rozkladů jako

$$\hat{l}_k = \sqrt{\hat{l}_k^{(-i\bullet)} \hat{l}_k^{(-\bullet j)}}, \quad k = 1, \dots, m-1,$$

kde $\hat{l}_k^{(-i\bullet)}$, resp. $\hat{l}_k^{(-\bullet j)}$ jsou prvky matice $\mathbb{L}_{(-i\bullet)}$, resp. $\mathbb{L}_{(-\bullet j)}$. Výslednou aproximaci je pak ještě třeba vynásobit znaménkem prvku $\tilde{x}_{ij}(m)$ ze vzorce (5.3), jenž byl vytvořen na základě aproximace původního pozorování za použití úplného datového souboru – dostáváme tedy aproximaci ze vzorce (5.5). Kritérium metody má tvar

$$W(m) = \frac{\text{MPRESS}(m-1) - \text{MPRESS}(m)}{D_{full}(m)} \cdot \frac{D_{res}(m)}{\text{MPRESS}(m)},$$

kde

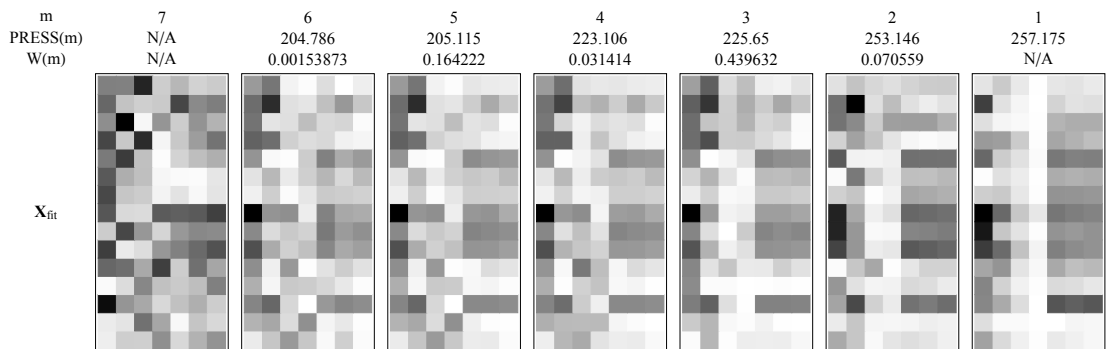
$$\begin{aligned} D_{full}(m) &= n + d - 2m, \quad m = 1, \dots, p-1, \\ D_{res}(m) &= d(n-1) - m(n+d-m-1), \end{aligned}$$

značí počet stupňů volnosti ztracených při aproximaci m -tou hlavní komponentou a počet stupňů volnosti, které zůstávají po aproximaci m -tou hlavní komponentou (viz Abdi a Williams, 2010, str.441). Optimální počet hlavních komponent je pak počet případů, kdy nastala situace $W(m) > 1$, resp. $W(m) > 0,9$.

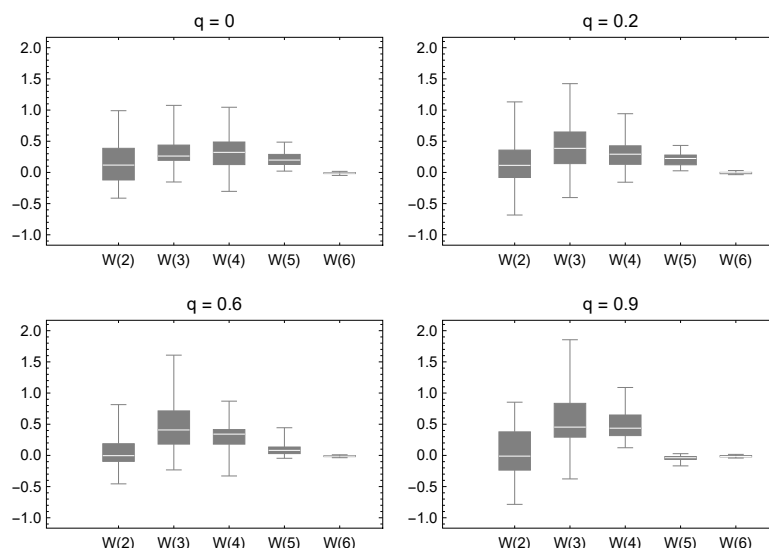
Popsanou metodu ilustrujeme pomocí simulací a aplikace na data z balíčku `SMSdata`. Pro účely simulace definujeme matici \mathbb{Q} s možností volby některých prvků následovně

$$\mathbb{Q} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & q & q & q \\ 0 & 0 & 0 & q & 1 & c & c \\ 0 & 0 & 0 & q & c & 1 & c \\ 0 & 0 & 0 & q & c & c & 1 \end{pmatrix}, \quad (5.8)$$

přičemž prvky $c = 0,9$ zvolíme pevně. Matici \mathbb{X} vytvoříme pomocí generování 50 náhodných vektorů z rozdělení $\mathcal{N}_7(\mathbf{0}, \mathbb{Q})$, kde parametr $q = 0$. Pro klesající počet hlavních komponent m použitých v aproximaci hodnota $\text{PRESS}(m)$ roste, jak dokládá obrázek 5.1 vykreslený pro prvních 15 řádků matice \mathbb{X} . Naopak kritérium $W(m)$ nevykazuje monotónní průběh. Na obrázku 5.2 jsou znázorněny krabicové grafy rozdělení kritéria $W(m)$ pro několik voleb q v matici \mathbb{Q} . Aplikaci metody na datech z balíčku `SMSdata` pak vidíme v tabulce 5.1.



Obrázek 5.1: Aproximace matice na základě zkráceného singulárního rozkladu (5.5); aproximovaná matice uvedena v prvním sloupci, vysvětlivky viz text



Obrázek 5.2: Krabicové grafy rozdělení kritéria $W(m)$ pro různé volby q v matici \mathbb{Q} ze vzorce (5.8); vypočteno na základě nastavení $p = 7$, $n = 50$ a počtu realizací 50

m^*	bank	crime	athlete	comp
0	0,247	-0,453	-0,982	-0,324
-1	0,114	0,047	0,646	1,756
-2	0,085	-0,202	-0,101	-0,102
-3	0,024	0,163	-0,027	-0,028
-4		-0,048	-0,007	

Tabulka 5.1: Data z balíčku `SMSdata` – hodnoty kritéria PRESS pro různé volby m . V levém sloupci uveden počet členů singulárního rozkladu, které byly zanedbány, se záporným znaménkem, tj. $m^* = m - M$, kde M je počet proměnných v jednotlivých datových souborech.

V článku Besse a Ferre (1993) se uvádí, že kritérium z článku Eastment a Krzanowski (1982) je asymptoticky ekvivalentní součtu několika posledních vlastních čísel a že se snižuje s rostoucím počtem uvažovaných hlavních komponent, a tedy vždy vede k volbě tolika hlavních komponent, jako je dimenze náhodného vektoru. Z tohoto důvodu byla v práci Bro a kol. (2008) navržena metoda křížového ověřování využívající EM algoritmu (*expectation-maximization*). Na tuto práci pak navázali autoři článku Josse a Husson (2012), ve kterém uvedli aproximaci kritéria pro volbu počtu hlavních komponent. Oproti popsanému algoritmu autor článku Wold (1978) přistupuje k výpočtu členů rozvoje ze vzorce (5.4) jiným způsobem, a to na základě rozdělení dat z matice \mathbb{X} do několika bloků obsahujících v řádcích i sloupcích přibližně stejný počet prvků. Metodami používanými křížové ověřování pro odhad optimálního počtu hlavních komponent se již zabývala celá řada autorů. Jednotlivé přístupy jsou porovnány v přehledových článcích jako např. Bro a kol. (2008), Diana a Tommasi (2002) a Saccenti a Camacho

(2015). Poznamenejme, že při používání metod založených na křížovém ověřování je třeba brát v úvahu jejich výpočetní složitost.

5.2 Bayesovský přístup

Bayesovský přístup k určení optimálního počtu hlavních komponent vychází z tzv. pravděpodobnostního modelu, který stručně popíšeme. Model (5.1) zapíšeme vektorově:

$$\mathbf{X} = \mathbb{H}\mathbf{W} + \boldsymbol{\mu} + \boldsymbol{\epsilon}, \quad (5.9)$$

kde \mathbf{X} je d -rozměrný náhodný vektor, $\boldsymbol{\mu}$ je d -rozměrný vektor středních hodnot, \mathbb{H} značí matici typu $d \times k$, a dále předpokládáme

$$\mathbf{W} \sim \mathcal{N}_k(\mathbf{0}, \mathbb{I}_k) \quad \text{a} \quad \boldsymbol{\epsilon} \sim \mathcal{N}_d(\mathbf{0}, \mathbb{V}),$$

kde

$$\mathbb{V} = v\mathbb{I}_d, \quad v > 0,$$

a navíc \mathbf{W} a $\boldsymbol{\epsilon}$ jsou navzájem nezávislé. Zápis (5.9) odpovídá modelu faktorové analýzy, kde se ovšem v matici \mathbb{V} nevyžadují stejné prvky na hlavní diagonále. Cílem metody hlavních komponent je vhodně zvolit dimenzi podprostoru k , do něhož se provede projekce dat. Za tímto účelem pro každou z dimenzí $k = 1, \dots, d$ spočítáme podmíněnou hustotu kolekce n realizací vektoru \mathbf{X} , tj. $\mathbb{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^\top$, a tuto hustotu označíme $p(\mathbb{X}|k)$. Výpočet provedeme integrací přes všechny parametry modelu \mathbb{H} , $\boldsymbol{\mu}$, v (výpočty uvedené níže jsou převzaty z článku Minka (2000)). Jelikož zřejmě

$$\mathbf{X}|\mathbb{H}, \boldsymbol{\mu}, v \sim \mathcal{N}_d(\boldsymbol{\mu}, \mathbb{H}\mathbb{H}^\top + v\mathbb{I}_d),$$

kde výraz nalevo značí podmíněné rozdělení \mathbf{X} při daných \mathbb{H} , $\boldsymbol{\mu}$, v , tak

$$\begin{aligned} p(\mathbb{X}|\mathbb{H}, \boldsymbol{\mu}, v) &= \prod_{i=1}^n p(\mathbf{x}_i|\mathbb{H}, \boldsymbol{\mu}, v) \\ &= (2\pi)^{-nd/2} \left| \mathbb{H}\mathbb{H}^\top + v\mathbb{I}_d \right|^{-n/2} \exp \left[-\frac{1}{2} \text{tr} \left((\mathbb{H}\mathbb{H}^\top + v\mathbb{I}_d)^{-1} \mathbb{S} \right) \right], \\ \text{kde } \mathbb{S} &= \sum_{i=1}^n (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^\top. \end{aligned}$$

Využitím vlastností podmíněné hustoty získáváme vztah

$$p(\mathbb{X}|\mathbb{H}, v) = \int_{\mathbb{R}^d} p(\mathbb{X}|\mathbb{H}, \boldsymbol{\mu}, v) p(\boldsymbol{\mu}) d\boldsymbol{\mu} \quad (5.10)$$

a budeme předpokládat, že $\boldsymbol{\mu}$ má neinformativní apriorní rozdělení

$$p(\boldsymbol{\mu}) = 1, \quad \boldsymbol{\mu} \in \mathbb{R}^d.$$

V případě takového nastavení hustoty vektoru středních hodnot je možné vztah (5.10) vyjádřit následovně

$$p(\mathbb{X}|\mathbb{H}, v) = n^{-d/2}(2\pi)^{-(n-1)d/2} |\mathbb{H}\mathbb{H}^\top + v\mathbb{I}_d|^{-(n-1)/2} \cdot \exp\left[-\frac{1}{2} \text{tr}\left((\mathbb{H}\mathbb{H}^\top + v\mathbb{I}_d)^{-1}\hat{\mathbb{S}}\right)\right], \quad (5.11)$$

$$\text{kde } \hat{\mathbb{S}} = \sum_{i=1}^n (\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^\top, \\ \hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i. \quad (5.12)$$

V článku Minka (2000) jsou pak popsány další kroky, kterými lze od vyjádření podmíněné hustoty ve vzorci (5.11) přejít ke kýžené podmíněné hustotě $p(\mathbb{X}|k)$, přičemž se použije rozklad matice \mathbb{H} a Laplaceova aproximace integrace.

Alternativně bychom mohli postupovat podle metody maximální věrohodnosti a odhadnout zbylé parametry modelu jako řešení úlohy

$$\max_{k=1, \dots, d, v > 0} p(\mathbb{X}|\mathbb{H}, v). \quad (5.13)$$

Přehled metod pro volbu počtu hlavních komponent založených na bayesovském přístupu je rovněž uveden v článku Sobczyk a kol. (2017), a to včetně navrženého přístupu pomocí metody PESEL (*penalized semi-integrated likelihood*). Další přístup odvozený na základě bayesovského modelu popisují autoři článku Seghouane a Cichocki (2007).

Závěr

Náplní této práce byla analýza hlavních komponent jakožto prostředku, jak provést projekci dat do prostoru nižší dimenze. Pojednali jsme o problému výběru adekvátního počtu hlavních komponent a představili jsme klasické i výpočetně složité nedávno navržené metody. Zároveň jsme zdůraznili, že problematický je rovněž samotný pojem „vhodný počet hlavních komponent“.

V rámci první části byly představeny základní pojmy doprovázející práci a teoretické vlastnosti hlavních komponent a jejich výběrových verzí. Zároveň jsme popsali teoretickou konstrukci biplotu, tj. grafického nástroje pro zobrazování dat a vztahů mezi náhodnými veličinami vzhledem k prvním dvěma komponentám. Dále jsme podrobně popsali heuristická pravidla pro volbu vhodného počtu hlavních komponent – jednalo se o podíl celkové variability, Kaiserovo-Guttmanovo kritérium, broken stick model, scree graf a LEV model. Výsledky metod byly ilustrovány na reálných datech. Dále jsme upozornili na četná úskalí, jež používání tohoto typu doprovázejí. Důkazem byla např. simulační studie pro empirické rozdělení podílu variability porovnané s broken stick modelem.

Následně jsme podrobně pojednali o přesném a asymptotickém rozdělení vlastních čísel bílé Wishartovy matice a provedli čtené grafické ilustrace. Pomocí symbolických výpočtů v softwaru *Mathematica* jsme odvodili předpis pro marginální hustotu vlastních čísel v trojrozměrném případě. Podrobně jsme se věnovali Tracyho-Widomovu rozdělení, které představuje asymptotické rozdělení největšího vlastního čísla bílé Wishartovy matice. Provedli jsme rovněž aproximaci tohoto rozdělení pomocí obecného a posunutého gama rozdělení. Zjištěné poznatky se využily při odvození statistického testu pro počet signifikantních vlastních čísel za určitých předpokladů, který bylo rovněž možné použít k určení vhodného počtu hlavních komponent, a to v případě, kdy se mnohá pravidla jeví jako nespolehlivá. Jak již bylo zmíněno, tyto testy fungují za velmi specifických předpokladů a jsou zajímavé zejména z teoretického hlediska. Na svou aplikaci stále ještě čekají.

Téma volby optimálního počtu hlavních komponent je velice rozsáhlé a má vysoký potenciál, co do počtu nově navržených metod. Dalším námětem, o který by bylo možné tuto diplomovou práci rozšířit, by byla návaznost na testy rovnosti vlastních čísel. Pokud bychom totiž odvodili sdruženou hustotu největšího a nejmenšího vlastního čísla λ_{\max} a λ_{\min} bílé Wishartovy matice, mohli bychom testovat hypotézu

$$H_0 : \lambda_{\max} = \lambda_{\min}$$

oproti alternativě

$$H_1 : \lambda_{\max} > \lambda_{\min}.$$

Zamítnutí nulové hypotézy by znamenalo, že bychom nemuseli dále pracovat s nsníženou dimenzí datového souboru. Jaké komponenty bychom mohli místo původních veličin uvažovat, by bylo ještě nutné podrobit důkladnější analýze.

Jak již bylo zmíněno, v této práci jsme pojednali o Tracyho-Widomově rozdělení a zrcadleném Tracyho-Widomově rozdělení, kterým bylo možné modelovat nejmenší a největší vlastní číslo bílé Wishartovy matice. Pokud jde o nejmenší vlastní číslo, bylo by třeba ještě provést transformaci exponenciální funkcí, abychom odstranili logaritmus. V momentě, kdy umíme modelovat marginální rozdělení

největšího a nejmenšího vlastního čísla, tak by již pro aproximativní řešení postačovalo modelovat závislosti pomocí kopule. Velmi stručně řečeno, např. dvou-dimenzionální kopule je funkce, která dvojici distribučních funkcí přiřadí sdruženou distribuční funkci. Volba vhodné kopule by byla námětem numerické studie. V případě tří-rozměrné bílé Wishartovy matice by bylo ještě možné kopuli pro marginální distribuční funkce tří vlastních čísel této matice konfrontovat s přesným řešením, neboť v rámci kapitoly 3 jsme uvedli přesné vyjádření sdružené hustoty vektoru vlastních čísel, na základě něhož by bylo možné skutečnou distribuční funkci pro tuto trojici vlastních čísel dopočítat. V případě dimenzí vyšších než tři je symbolický výpočet velice náročný a možnost konfrontace by byla otázkou proveditelnosti.

Poslední část práce shrnovala výpočetně složité metody včetně přehledové literatury zabývající se problematikou volby počtu hlavních komponent pomocí výpočetně složitějších metod, než byly procedury doposud zmíněné (konkrétně křížové ověřování a bayesovské modely). Simulační studie a grafické studie byly provedeny v softwarech Wolfram *Mathematica*, verze 11.1.1, a R, verze 3.3.3.

Seznam použité literatury

- ABDI, H. a WILLIAMS, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, **2**(4), 433–459.
- ABRAMOWITZ, M. a STEGUN, I. A. (2010). *Handbook of mathematical functions*. 1. Cambridge University Press, New York. ISBN 978-0-521-14063-8.
- ANDERSON, T. W. (1963). Asymptotic theory for principal component analysis. *The Annals of Mathematical Statistics*, **34**(1), 122–148.
- BESSE, P. a FERRE, L. (1993). Sur l’usage de la validation croisée en analyse en composantes principales. *Revue de statistique appliquée*, **41**(1), 71–76.
- BRO, R., KJELDAHL, K., SMILDE, A. K. a KIERS, H. A. L. (2008). Cross-validation of component models: a critical look at current methods. *Analytical and Bioanalytical Chemistry*, **5**(390), 1241–1251.
- CHIANI, M. (2014). Distribution of the largest eigenvalue for real Wishart and Gaussian random matrices and a simple approximation for the Tracy–Widom distribution. *Journal of Multivariate Analysis*, **129**(8), 69–81.
- CRADDOCK, J. M. a FLOOD, C. R. (1969). Eigenvectors for representing the 500 mb geopotential surface over the Northern Hemisphere. *Quarterly journal of the Royal Meteorological Society*, **95**(2), 576–593.
- DAVID, H. A. a NAGARAJA, H. N. (2003). *Order statistics*. Third edition. John Wiley and Sons, New Jersey. ISBN 0-471-38926-9.
- DEVROYE, L. (1986). *Non-Uniform Random Variate Generation*. Springer, New York. ISBN 978-1-4613-8645-2.
- DIANA, G. a TOMMASI, C. (2002). Cross-validation methods in principal component analysis: a comparison. *Statistical Methods and Applications*, **11**(1), 71–82.
- EASTMENT, H. a KRZANOWSKI, W. (1982). Cross-validatory choice of the number of components from a principal component analysis. *Technometrics*, **24**(1), 73–77.
- FRONTIER, S. (1976). Étude de la décroissance des valeurs propres dans une analyse en composantes principales: Comparaison avec le modèle du bâton brisé. *Journal of experimental marine Biology and Ecology*, **25**(1), 67–75.
- HLÁVKA, Z. (2012). *Data sets for multivariate statistics: exercises and solutions*. Version 1.1, package SMSdata. URL http://www.karlin.mff.cuni.cz/~hlavka/sms2/SMSdata_1.0.zip.
- HOYLE, D. C. (2008). Automatic PCA dimension selection for high dimensional data and small sample sizes. *Journal of Machine Learning Research*, **9**(1), 2733–2759.

- JACKSON, D. A. (1993). Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. *Ecology*, **74**(8), 2204–2214.
- JOHNSON, R. A. a WICHERN, D. W. (1992). *Applied multivariate statistical analysis*. Third edition. Prentice Hall, New Jersey. ISBN 0-13-041773-4.
- JOHNSTONE, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, **29**(2), 295–327.
- JOLLIFFE, I. (2002). *Principal component analysis*. Second edition. Springer, New York. ISBN 0-387-95442-2.
- JOSSE, J. a HUSSON, F. (2012). Selecting the number of components in principal component analysis using cross-validation approximations. *Computational Statistics and Data Analysis*, **56**(6), 1869–1879.
- MA, Z. (2012). Accuracy of the Tracy–Widom limits for the extreme eigenvalues in white Wishart matrices. *Bernoulli*, **18**(1), 322–359.
- MARDIA, K. V., KENT, J. T. a BIBBY, J. M. (2003). *Multivariate analysis*. IV. Series. Academic Press, London. ISBN 0-12-471252-5.
- MATOUŠEK, J. a NEŠETŘIL, J. (2009). *Kapitoly z diskrétní matematiky*. Čtvrté vydání. Karolinum, Praha. ISBN 978-80-246-1740-4.
- MINKA, T. P. (2000). Automatic choice of dimensionality for PCA. *Advances in neural information processing systems*, **13**(1), 598–604.
- PARK, H. a KONISHI, S. (2017). Principal component selection via adaptive regularization method and generalized information criterion. *Statistical Papers*, **58**(1), 147–160.
- PERES-NETO, P. R., JACKSON, D. A. a SOMERS, K. M. (2005). How many principal components? Stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics and Data Analysis*, **49**(4), 974–997.
- R CORE TEAM (2017). *R: a language and environment for statistical computing*. Version 3.3.3, R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>.
- SACCENTI, E. a CAMACHO, J. (2015). Determining the number of components in principal components analysis: a comparison of statistical, crossvalidation and approximated methods. *Chemometrics and Intelligent Laboratory Systems*, **149**(Part A), 99–116.
- SEGHOUANE, A. K. a CICHOCKI, A. (2007). Bayesian estimation of the number of principal components. *Signal Processing*, **87**(3), 562–568.
- SHAW, W. T. (2010). Monte Carlo portfolio optimization for general investor risk-return objectives and arbitrary return distributions: a solution for long-only portfolios. URL <https://arxiv.org/abs/1008.3718>.

- SOBCZYK, P., BOGDAN, M. a JOSSE, J. (2017). Bayesian dimensionality reduction with PCA using penalized semi-integrated likelihood. *Journal of Computational and Graphical Statistics*, **26**(4), 826–839.
- SUAREZ, A. a GHOSAL, S. (2017). Bayesian estimation of principal components for functional data. *Bayesian Analysis*, **12**(2), 311–333.
- TEBBENS, J. D., HNĚTYNKOVÁ, I., PLEŠINGER, M., STRAKOŠ, Z. a TICHÝ, P. (2012). *Analýza metod pro maticové výpočty*. První vydání. Matfyzpress, Praha. ISBN 978-80-7378-201-6.
- TIPPING, M. E. a BISHOP, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B*, **61**(3), 611–622.
- TRACY, C. A. a WIDOM, H. (1996). On orthogonal and symplectic matrix ensembles. *Communications in Mathematical Physics*, **177**(3), 727–754.
- TRACY, C. A. a WIDOM, H. (2009). The distributions of random matrix theory and their applications. *New trends in mathematical physics*, pages 753–765.
- VLOK, J. D. a OLIVIER, J. C. (2012). Analytic approximation to the largest eigenvalue distribution of a white Wishart matrix. *IET communications*, **6**(12), 1804–1811.
- WOLD, S. (1978). Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics*, **20**(4), 397–405.
- WOLFRAM RESEARCH, INC. (2017). *Mathematica*. Version 10.1.1, Student Edition, Champaign, Illinois. URL <http://www.wolfram.com/>.
- WOLFRAM RESEARCH, INC. (2018). *Trust region methods, tutorial*. *Wolfram language documentation*. URL <http://reference.wolfram.com/language/tutorial/UnconstrainedOptimizationTrustRegionMethods.html>.

A. Přílohy

A.1 Zemědělská data

plodiny kraje	pse	jec	bram	rep	slun	pic
PH	29,91	11,30	0,15	7,31	0,00	5,15
S	1166,37	405,16	116,86	287,22	7,17	387,16
JC	457,37	184,53	59,23	135,15	0,05	323,83
PL	350,31	148,66	20,11	107,60	0,99	266,82
KV	67,00	22,34	1,81	20,33	0,00	35,41
U	425,19	134,70	10,21	79,90	4,10	90,25
L	73,87	24,53	3,61	19,13	0,01	44,27
KH	380,98	92,51	18,58	86,97	0,52	204,85
PA	341,92	120,91	23,27	85,69	1,41	255,64
V	425,60	235,05	180,10	126,37	0,01	438,26
JM	715,38	220,54	36,89	116,03	16,58	243,31
O	356,51	224,68	9,45	81,64	0,05	178,66
Z	230,64	66,10	6,27	44,12	0,68	120,46
M	253,21	100,40	18,41	58,75	0,04	114,24

Pozn: pse – pšenice, jec – ječmen, bram – brambory, rep – řepka,
slun – slunečnice, pic – píce

Tabulka A.1: Hektarové výnosy zemědělských plodin na území České republiky za rok 2014

	pse	jec	bram	rep	slun	pic
min	29,91	11,3	0,15	7,31	0,00	5,15
q_1	236,29	72,70	7,06	47,77	0,02	96,25
med	353,41	127,80	18,49	83,66	0,29	191,75
\bar{x}	376,73	142,24	36,07	89,73	2,26	193,45
q_3	425,50	211,54	33,48	113,92	1,31	264,02
max	1166,37	405,16	180,10	287,22	16,58	438,26
sd	290,40	106,85	51,61	69,75	4,60	133,44

Pozn: min – minimum, q_1 – odhad dolního kvartilu, med – odhad mediánu, \bar{x} – výběrový průměr, q_3 – odhad horního kvartilu, max – maximum, sd – odhad směrodatné odchylky

Tabulka A.2: Výběrové charakteristiky hektarových výnosů plodin

značka	kraj	značka	kraj
PH	Praha	KH	Královéhradecký
S	Středočeský	PA	Pardubický
JC	Jihočeský	V	Vysočina
PL	Plzeňský	JM	Jihomoravský
KV	Karlovarský	O	Olomoucký
U	Ústecký	Z	Zlínský
L	Liberecký	M	Moravskoslezský

Tabulka A.3: Přiřazené značení jednotlivým krajům

A.2 Standardizovaná zemědělská data

plodiny kraje	pse	jec	bram	rep	slun	pic
PH	-1,19	-1,23	-0,70	-1,18	-0,49	-1,41
S	2,72	2,46	1,57	2,83	1,07	1,45
JC	0,28	0,40	0,45	0,65	-0,48	0,98
PL	-0,09	0,06	-0,31	0,26	-0,27	0,55
KV	-1,07	-1,12	-0,66	-0,99	-0,49	-1,18
U	0,17	-0,07	-0,50	-0,14	0,40	-0,77
L	-1,04	-1,10	-0,63	-1,01	-0,49	-1,12
KH	0,01	-0,47	-0,34	-0,04	-0,38	0,09
PA	-0,12	-0,20	-0,25	-0,06	-0,18	0,47
V	0,17	0,87	2,79	0,53	-0,49	1,83
JM	1,17	0,73	0,02	0,38	3,11	0,37
O	-0,07	0,77	-0,52	-0,12	-0,48	-0,11
Z	-0,50	-0,71	-0,58	-0,65	-0,34	-0,55
M	-0,43	-0,39	-0,34	-0,44	-0,48	-0,59

Tabulka A.4: Hektarové výnosy zemědělských plodin na území České republiky za rok 2014

	pse	jec	bram	rep	slun	pic
min	-1,19	-1,23	-0,7	-1,18	-0,49	-1,41
q_1	-0,48	-0,65	-0,56	-0,6	-0,49	-0,73
med	-0,08	-0,14	-0,34	-0,09	-0,43	-0,01
\bar{x}	0	0	0	0	0	0
q_3	0,17	0,65	-0,05	0,35	-0,21	0,53
max	2,72	2,46	2,79	2,83	3,11	1,83
sd	1	1	1	1	1	1

Tabulka A.5: Výběrové charakteristiky standardizovaných hektarových výnosů plodin

A.3 Charakteristiky použitých dat z balíčku SMSdata

Použitá data z datového souboru `uscrime`

	land area	popu 1985	robbery	assault	burglary	larceny	autotheft
min	-0,80	-0,84	-1,04	-1,68	-1,79	-1,76	-1,45
q_1	-0,40	-0,70	-0,60	-0,75	-0,69	-0,73	-0,75
med	-0,18	-0,30	-0,27	-0,15	-0,17	-0,03	-0,12
μ	0	0	0	0	0	0	0
q_3	0,12	0,18	0,28	0,82	0,58	0,52	0,73
max	5,87	4,26	3,75	2,31	2,28	2,26	2,56
σ	1	1	1	1	1	1	1

Tabulka A.6: Výběrové charakteristiky standardizovaných veličin z datového souboru `uscrime`

Použitá data z datového souboru `bank2`

	length	height left	height right	inner frame lower	inner frame upper	diagonal diagonal
min	-2,91	-3,11	-2,37	-1,54	-3,67	-2,33
q_1	-0,79	-0,61	-0,63	-0,84	-0,69	-0,85
med	0,01	0,22	0,11	-0,22	-0,06	-0,03
μ	0	0	0	0	0	0
q_3	0,54	0,77	0,66	0,82	0,68	0,88
max	3,73	2,43	2,83	2,27	2,05	1,66
σ	1	1	1	1	1	1

Tabulka A.7: Výběrové charakteristiky standardizovaných veličin z datového souboru `bank2`

Použitá data z datového souboru athletic

	100m	200m	400m	800m	1500m	5000m	10000m	marathon
min	-1,54	-0,74	-1,77	-1,46	-1,21	-1,04	-0,9	-0,91
q_1	-0,57	-0,33	-0,6	-0,6	-0,63	-0,71	-0,71	-0,64
med	-0,17	-0,23	-0,23	-0,05	-0,37	-0,43	-0,45	-0,46
μ	0	0	0	0	0	0	0	0
q_3	0,34	0	0,59	0,34	0,46	0,37	0,48	0,29
max	4,86	5,64	4,46	3,56	3,48	3,56	3,54	3,04
σ	1	1	1	1	1	1	1	1

Tabulka A.8: Výběrové charakteristiky standardizovaných veličin z datového souboru athletic

Použitá data z datového souboru uscomp

	assets	sales	market value	profits	cash flow	employees
min	-0,63	-1,72	-0,29	-1,23	-0,87	-0,57
q_1	-0,53	-0,83	-0,25	-0,22	-0,27	-0,52
med	-0,34	0,01	-0,21	-0,18	-0,22	-0,35
μ	0	0	0	0	0	0
q_3	-0,02	0,85	-0,12	-0,02	-0,06	0,16
max	5,07	1,69	8,12	7,91	7,81	5,59
σ	1	1	1	1	1	1

Tabulka A.9: Výběrové charakteristiky standardizovaných veličin z datového souboru uscomp

A.4 Rovnost pro harmonická čísla

Harmonické číslo H_p lze zapsat buď z definice nebo také podle následujícího tvaru:

$$\begin{aligned} H_p &= \sum_{i=1}^p \frac{1}{i} \\ &= \int_0^1 (1 + x + \dots + x^{p-1}) dx = \int_0^1 \frac{1 - x^p}{1 - x} dx. \end{aligned}$$

Upravujme druhý způsob vyjádření, přičemž zavedme substituci $x = 1 - y$:

$$\begin{aligned} H_p &= \int_0^1 \frac{1 - x^p}{1 - x} dx = \int_0^1 \frac{1 - (1 - y)^p}{y} dy = \\ &= \int_0^1 \left[\sum_{k=1}^p (-1)^{k-1} \binom{p}{k} y^{k-1} \right] dy = \\ &= \sum_{k=1}^p (-1)^{k-1} \binom{p}{k} \int_0^1 y^{k-1} dy = \sum_{k=1}^p (-1)^{k-1} \frac{1}{k} \binom{p}{k}. \end{aligned}$$

Tím je ověřena platnost kýženého vztahu.

A.5 Sdružené a marginální hustoty vlastních čísel bílé Wishartovy matice

$$f_{\lambda}(x_1, x_2, x_3) = -e^{-\frac{x_1}{2} - \frac{x_2}{2} - \frac{x_3}{2}} x_2^3 x_3^4 x_1^5 + e^{-\frac{x_1}{2} - \frac{x_2}{2} - \frac{x_3}{2}} x_2^4 x_3^3 x_1^5 + e^{-\frac{x_1}{2} - \frac{x_2}{2} - \frac{x_3}{2}} x_2^3 x_3^5 x_1^4$$

$$- e^{-\frac{x_1}{2} - \frac{x_2}{2} - \frac{x_3}{2}} x_2^5 x_3^3 x_1^4 - e^{-\frac{x_1}{2} - \frac{x_2}{2} - \frac{x_3}{2}} x_2^4 x_3^5 x_1^3 + e^{-\frac{x_1}{2} - \frac{x_2}{2} - \frac{x_3}{2}} x_2^5 x_3^4 x_1^3$$

$$f_{\lambda_1}(x) = -\frac{e^{-\frac{3x}{2}} x^9}{645120} + \frac{e^{-x} x^8}{40320} - \frac{e^{-\frac{3x}{2}} x^8}{21504} - \frac{e^{-\frac{3x}{2}} x^7}{1536} - \frac{1}{192} e^{-\frac{3x}{2}} x^6 - \frac{19}{768} e^{-\frac{3x}{2}} x^5$$

$$+ \frac{1}{768} e^{-\frac{x}{2}} x^5 - \frac{9}{128} e^{-\frac{3x}{2}} x^4 - \frac{3}{128} e^{-\frac{x}{2}} x^4 - \frac{3}{32} e^{-\frac{3x}{2}} x^3 + \frac{3}{32} e^{-\frac{x}{2}} x^3,$$

$$f_{\lambda_2}(x) = \frac{e^{-x} x^8}{40320}$$

$$f_{\lambda_3}(x) = \frac{e^{-\frac{3x}{2}} x^9}{645120} + \frac{e^{-\frac{3x}{2}} x^8}{21504} + \frac{e^{-\frac{3x}{2}} x^7}{1536} + \frac{1}{192} e^{-\frac{3x}{2}} x^6 + \frac{19}{768} e^{-\frac{3x}{2}} x^5$$

$$+ \frac{9}{128} e^{-\frac{3x}{2}} x^4 + \frac{3}{32} e^{-\frac{3x}{2}} x^3$$