MASARYKOVA UNIVERZITA
Přírodovědecká fakulta
Ústav matematiky a statistiky

Bakalářská práce

Brno 2019 **Ngoc Phuong Vu**

MASARYKOVA
UNIVERZITA
PŘÍRODOVĚDECKÁ FAKULTA
ÚSTAV MATEMATIKY A STATISTIKY

Bagplot

Bakalářská práce

Ngoc Phuong Vu

Vedoucí práce: RNDr. Radim Navrátil, Ph.D. **Brno 2019**

Bibliografický záznam

Autor:	Ngoc Phuong Vu Přírodovědecká fakulta, Masarykova univerzita Ústav matematiky a statistiky

Název práce:	Bagplot
Studijní program:	Matematika
Studijní obor:	Statistika a analýza dat
Vedoucí práce:	RNDr. Radim Navrátil, Ph.D.
Akademický rok:	2018/2019
Počet stran:	vii + 33
Klíčová slova:	Bagplot; Konstrukce bagplotu; Vizualizace dat; Zobecnění krabicového grafu; Poloprostorová hloubka; Tukeyho medián; Hloubkové kontury; Krabicový graf; Kvantily; Odlehlost

Bibliographic Entry

Author:	Ngoc Phuong Vu Faculty of Science, Masaryk University Department of Mathematics and Statistics
Title of Thesis:	Bagplot
Degree Programme:	Mathematics
Field of Study:	Statistics and data analysis
Supervisor:	RNDr. Radim Navrátil, Ph.D.
Academic Year:	2018/2019
Number of Pages:	vii + 33
Keywords:	Bagplot; Bagplot construction; Data visualization; Generalisation of the boxplot; Halfspace depth; Tukey median; Depth contours; Boxplot; Quantiles; Outlyingness

Abstrakt

V této bakalářské práci se věnujeme bagplotu, grafickému nástroji pro vizualizaci dvourozměrných dat, který je zobecněním krabicového grafu. Nejprve si popíšeme známý krabicový graf, uvedeme si jeho různé varianty a ukážeme si jejich výhody. Dále si nadefinujeme pojmy jako poloprostorová hloubka, hloubkové kontury nebo Tukeyho medián, potřebné k sestrojení bagplotu. Uvedeme si konstrukci grafu a poté si na několika příkladech ilustrujeme jeho využití. K vykreslení grafů v této práci používáme programovací jazyk R.

Abstract

This thesis is devoted to the bagplot, a graphical tool for visualizing two-dimensional data, which is a generalisation of the boxplot. First, we will have a look at the boxplot, as well as some of its variants and demonstrate their advantages. Next, we define notions necessary to construct the bagplot, such as the halfspace depth, depth contours or the Tukey median. We describe the construction of the graph and then illustrate its use on several examples. For plotting the graphs in this paper we use the programming language R.

MASARYKOVA UNIVERZITA
Přírodovědecká fakulta

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

Akademický rok: 2018/2019

Ústav: Ústav matematiky a statistiky

Student: Ngoc Phuong Vu

Program: Matematika

Obor: Statistika a analýza dat

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

Název práce: Bagplot

Název práce anglicky: Bagplot

Oficiální zadání:
Bagplot je nástroj exploratorní analýzy dat pro vizualizaci dvourozměrných dat, je zobecněním jednorozměrného boxplotu. Úkolem studenta bude popsat jednorozměrný boxplot a jeho zobecnění do dvourozměrného prostoru jakožto bagplot. Nejprve odvodí jeho teoretické vlastnosti a poté ilustruje jeho použití na vhodných příkladech ve statistickém softwaru.

Literatura:
Handbook of data visualization. Edited by Chun-houh Chen - Wolfgang Härdle - Antony Unwin. Berlin: Springer, 2008. xiii, 936. ISBN 9783540330370.

ROUSSEEUW, Peter J. a Annick M. LEROY. *Robust regression and outlier detection*. Hoboken, N.J.: Wiley-Interscience, 2003. xiv, 329. ISBN 0471488550.

ROUSSEEUW, P. J., I. RUTS a J. W. TUKEY. *The bagplot: a bivariate boxplot. The American Statistician*, 1999, roč. 53, s. 382–387.

Understanding robust and exploratory data analysis. Edited by David C. Hoaglin - Frederick Mosteller - John W. Tukey. Wiley classics library ed. New York: Wiley, 2000. xx, 447. ISBN 0471384917.

Jazyk závěrečné práce:

Vedoucí práce: RNDr. Radim Navrátil, Ph.D.

Datum zadání práce: 2. 10. 2018

V Brně dne: 31. 10. 2018

Souhlasím se zadáním (podpis, datum):

15.11.2018

Ngoc Phuong Vu	RNDr. Radim Navrátil, Ph.D.	prof. RNDr. Jan Slovák, DrSc.
student	**vedoucí práce**	**ředitel Ústavu matematiky a statistiky**

Poděkování

Na tomto místě bych chtěl hlavně poděkovat mému vedoucímu RNDr. Radimovi Navrátilovi, Ph.D. za cenné rady a připomínky, bez kterých bych se neobešel při tvorbě této bakalářské práce. Zároveň bych mu chtěl poděkovat za trpělivost, ochotu a čas, který věnoval průběžným opravám této práce.

Dále děkuji Jaroslavu Paličkovi za společně strávený čas plný zábavy a bláznivých nápadů během mého studentského života.

Velké poděkování patří mé přítelkyni Anežke za podporu a povzbuzení, nejen v průběhu psaní této práce. Také ji děkuji za prožité radostné a bezstarostné chvíle naplněné smíchem.

V neposlední řadě bych chtěl poděkovat svým rodičům za jejich nekonečnou podporu v průběhu celého studia.

Prohlášení

Prohlašuji, že jsem svoji bakalářskou práci vypracoval samostatně s využitím informačních zdrojů, které jsou v práci citovány.

Brno 15. května 2019 .....

Ngoc Phuong Vu

Obsah

Úvod

Průzkumová analýza dat (také známá pod zkratkou EDA - z anglického názvu exploratory data analysis) je základní krok pro jakoukoliv výzkumnou analýzu. Jedná se o souhrn metod zkoumající distribuci nasbíraných dat (obvykle se ověřuje, zda data pochází ze souboru s normálním rozdělením), odhalující odlehlá pozorování (tj. hodnoty, které se výrazně odlišují od ostatních), a v neposlední řadě také zobrazující případné vztahy mezi proměnnými. Všechny tyto získané poznatky pak poslouží k posouzení kvality dat a následně k vytvoření vhodných statistických modelů. Hlavními nástroji průzkumové analýzy bývají zpravidla grafické metody (např. histogram, krabicový graf, Paretův diagram, korelační diagram aj.), které umožňují vhled do zkoumaného souboru a za pomoci nichž lze získat informace o struktuře dat. Ačkoliv je poměrně snadné tyto grafy vykreslit, k jejich správné interpretaci je zapotřebí určitá zkušenost.

Průzkumovou analýzu dat popsal John W. Tukey v knize [16] v roce 1977. Ne příliš známým nástrojem této analýzy, popsaný v článku [10] z roku 1997, je dvourozměrný graf *bagplot*, kterým se tato práce zabývá. Je zobecněním jednorozměrného boxplotu, který je díky své jednoduchosti a přehlednosti často využívaný při rozboru dat. Podobně jako u boxplotu, výhodou bagplotu je snadná detekce odlehlých hodnot, názorné zobrazení rozsahu souboru, korelace mezi dvěma proměnnými a šikmosti dat.

Práce je rozdělena na dvě kapitoly, v té první si představíme populární krabicový graf. Zavedeme si všechny potřebné pojmy k jeho sestrojení a na vhodných příkladech si ukážeme jeho výhody. Popíšeme si několik dalších variant boxplotu, které navíc zobrazují například rozsah souboru dat nebo konfidenční intervaly pro medián a zároveň si předvedeme, jak tyto grafy sestrojit v programovacím jazyce R.

Ve druhé kapitole, kde čerpáme hlavně z článku [12] od autorů Peter J. Rousseeuw, Ida Ruts a John W. Tukey, se dostaneme k samotnému bagplotu. Hlavním cílem zde bude zobecnit pojem kvantil, na kterém je založen krabicový graf a poté si popíšeme jeho konstrukci. Podobně jako v předchozí kapitole si zde uvedeme několik příkladů, na kterých si ilustrujeme výhody použití bagplotu.

Pro snazší porozumění definovaných pojmů jsou v této práci vykresleny obrázky za pomocí programu `Geogebra`. Ačkoliv v dnešní době existuje mnoho statistických softwarů, schopných sestrojení diagnostických grafů, jako jsou například `MATLAB`[1], `STATISTICA`[2], `SAS`[3], zaměřil jsem hlavně na volně dostupný jazyk R[4].

[1] Dostupné z: https://www.mathworks.com/products/matlab.html
[2] Dostupné z: https://www.statistica.pro
[3] Dostupné z: https://www.sas.com/en_us/software/stat.html
[4] Dostupné z: https://www.r-project.org/

Kapitola 1

Boxplot

Boxplot (také **krabicový graf** nebo **krabicový diagram**) je jedním z nástrojů pro vizualizaci jednorozměrných dat, který přehledně zobrazuje informace o mediánu, horním a dolním kvartilu, maximálních a minimálních hodnotách, odlehlých hodnotách a dalších vlastnostech souboru zkoumaných dat. Boxplot použil jako první John W. Tukey v roce 1970, ale nebyl známý až do roku 1977, kdy byl formálně představen v knize [16]. Díky své jednoduchosti a kompaktnosti je často používaný k předběžné analýze dat.

Boxploty mohou být vykresleny vodorovně nebo svisle. Hlavní část grafu tvoří obdélník, neboli "krabička", uvnitř kterého se nachází linie reprezentující medián. Tato krabička reprezentuje 50 % prostředních hodnot uspořádaného souboru. Ze střední části obdélníka vycházejí linie, tzv. "vousy", které dosahují buď krajních hodnot, tedy minimální nebo maximální hodnoty (tato konvence ovšem není praktická, jelikož při jejím využití přicházíme o důležitou vlastnost boxplotu, a tou je detekce odlehlých hodnot, které se pak v grafu zobrazují jako samostatné body), popřípadě jiných hodnot, které si později popíšeme.

I přes jeho jednoduchost je krabicový graf hojně využívaným nástrojem hlavně k rychlému a snadnému posouzení rozložení dat, zjištění případné asymetrie a jednoznačnému zobrazení odlehlých hodnot. Oproti ostatním grafům, jako je například histogram, je krabicový graf kompaktnější a používá se při porovnávání variability několika souborů. Výhodou zmíněného histogramu je zase detailnější popis distribuce dat.

Některé boxploty mohou obsahovat znak (obvykle "+") reprezentující aritmetický průměr dat. Existují různé varianty tohoto grafu, které navíc zobrazují počet pozorování v jednotlivých skupinách dat souboru nebo odhad hustoty ve vybraných bodech. Krabicové grafy jsou neparametrické, zobrazují tedy rozdíly mezi datovými soubory bez předpokladu normálního rozdělení dat. Ačkoliv není těžké je ručně sestrojit, při velkém objemu dat je výhodnější použít statistický software k jeho vykreslení, jako jsou například `MATLAB`, `Microsoft Excel`[1], `STATISTICA` nebo programovací jazyk  .

[1] Dostupné z https://www.microsoft.com/Microsoft/Excel

1.1 Základní pojmy

Boxplot je založen na 5 hodnotách. Prvním z nich je medián nacházející se ve středu grafu, uvnitř krabičky, která je ohraničena 1. a 3. kvartilem. Dále obsahuje graf vnější hradby, které mohou být určeny minimální a maximální hodnotou, obvyklejší je však zvolit 1.5-násobek mezikvartilového rozpětí. K nadefinování všech těchto pojmů nejprve potřebujeme znát pojem kvantil.

1.1.1 Kvantily

Zadefinujme si nejprve teoretický kvantil z distribuční funkce.

Definice 1. Nechť F je distribuční funkcí a $\alpha \in (0,1)$. Potom funkce

$$F^{-1}(\alpha) = Q(\alpha) = \inf\{x \in \mathbb{R} : F(x) \geq \alpha\}$$

se nazývá **kvantilová funkce** a číslo

$$x_\alpha = Q(\alpha)$$

se nazývá **α-kvantilem** rozdělení s distribuční funkcí $F(x)$.

α-kvantil x_α je tedy taková hodnota náhodné veličiny, která udává $\alpha\%$ pravděpodobnost výskytu hodnot menších nebo rovno x_α.

Protože však zpravidla kreslíme boxplot na základě reálných dat, budeme potřebovat výběrový kvantil.

Mějme náhodný výběr $X_1, X_2, \ldots, X_n$. Datový soubor získaný náhodným výběrem lze znázornit pomocí číselných charakteristik, které nazýváme výběrové charakteristiky. Dělíme je na:

1. Míry polohy - souhrnná statistika, která představuje střední nebo typickou hodnotu vzorku dat

2. Míry variability - určují rozptyl kolem své střední nebo typické hodnoty

Mezi známé míry polohy patří výběrový průměr, modus a **výběrové kvantily**. Výběrový α-kvantil $\overline{x}_\alpha$ je obecně definován jako hodnota rozdělující datový soubor na dvě části - první část obsahuje hodnoty menší než daný kvantil $\overline{x}_\alpha$ a druhá část obsahuje hodnoty větší nebo rovny hodnotě $\overline{x}_\alpha$. První část tedy obsahuje aspoň $\alpha \cdot 100\,\%$ dat a druhá část obsahuje alespoň $(1-\alpha)\cdot 100\,\%$ dat.

Některé kvantily mají speciální názvy, např. *percentil* $\overline{x}_{0.01}$ je hodnota, pod kterou leží 1 % všech hodnot, *decil* $\overline{x}_{0.1}$ je 10. percentil a *kvartil* $\overline{x}_{0.25}$ je 25. percentil.

Nejčastěji používané kvantily jsou $\overline{x}_{0.25}$ nazývaný **dolní kvartil** (také 1. kvartil), dále $\overline{x}_{0.5}$ nazývaný jako **medián**, a v neposlední řadě $\overline{x}_{0.75}$ nazývaný **horní kvartil** nebo také 3. kvartil.

Postup při určování výběrových kvantilů:

1. Uspořádáme hodnoty datového souboru podle velikosti $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$.

2. Pokud $n\alpha$ je celé číslo k, pak $\bar{x}_\alpha = \frac{x_{(k)}+x_{(k+1)}}{2}$,
   pokud k není celé číslo, pak $\bar{x}_\alpha = x_{\lceil (k) \rceil}$,

přitom $\lceil k \rceil$ značí celou horní část čísla k.

Poznámka. Je-li n liché, je výběrový medián $\bar{x}_{0.5}$ prostřední hodnota seřazeného datového souboru $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, tedy

$$\bar{x}_{0.5} = x_{(\frac{n+1}{2})}.$$

V případě, že je n sudé, medián se vypočítá jako aritmetický průměr dvou prostředních hodnot, tj.

$$\bar{x}_{0.5} = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}.$$

Definice 2. Mezikvartilové rozpětí IQR (z anglického názvu interquartile range) definujeme vztahem:

$$IQR = \bar{x}_{0.75} - \bar{x}_{0.25} \tag{1.1.1}$$

a je to tedy rozdíl horního a dolního výběrového kvartilu.

1.1.2 Odlehlé a extrémní hodnoty

Vykreslením boxplotu lze snadno identifikovat odlehlá pozorování, popřípadě chybné hodnoty, které pak při analýze dat mohou vést k mylným závěrům. Definice odlehlých (a extrémních) hodnot není jednoznačná, neboť obor hodnot náhodné veličiny závisí na charakteru dat. Jako odlehlou hodnotu lze považovat takovou, která nezapadá do pravděpodobnostního chování souboru dat. Tukey ve své knize [16] navrhl následující definici:

Hodnotu x považujeme za odlehlou, jestliže platí

$$x > \bar{x}_{0.75} + k \cdot IQR$$

nebo

$$x < \bar{x}_{0.25} - k \cdot IQR.$$

Hodnota x je extrémní, jestliže platí

$$x > \bar{x}_{0.75} + 2k \cdot IQR$$

nebo

$$x < \bar{x}_{0.25} - 2k \cdot IQR.$$

Obvykle za k volíme hodnotu 1.5 (viz. [16]), odlehlé hodnoty pak leží v intervalu

$$[\bar{x}_{0.25} - 3 \cdot IQR, \bar{x}_{0.25} - 1.5 \cdot IQR]$$

nebo v intervalu

$$[\bar{x}_{0.75} + 1,5 \cdot IQR, \bar{x}_{0.75} + 3 \cdot IQR]$$

a extrémní hodnoty jsou v tomto případě menší než $\bar{x}_{0.25} - 3 \cdot IQR$ nebo větší než $\bar{x}_{0.75} + + 3 \cdot IQR$.

```
> n1 <- table(x < dolni.hradba)
> n1

 FALSE   TRUE
996523   3477
> n2 <- table(x > horni.hradba)
> n2

 FALSE   TRUE
996502   3498
> n1[2] + n2[2]
TRUE
6975
```

Obrázek 1.1: Součet hodnot ležících pod dolní hradbou a nad horní hradbou z vektoru x. $n1$ zde značí počet hodnot ležících pod dolní hradbou a $n2$ počet hodnot nad horní hradbou. Vidíme, že tyto hodnoty tvoří celkem 0.6975 % z celkového počtu pozorování.

Platí, že pokud máme data z normálního rozdělení, pravděpodobnost výskytu odlehlých hodnot je zhruba 0.7 %. Pro ukázku si nagenerujme v vektor x obsahující 1 000 000 hodnot z normálního rozdělení. Zjistíme hodnoty dolní ($\bar{x}_{0.25} - 1.5 \cdot IQR$) a horní ($\bar{x}_{0.75} + + 1.5 \cdot IQR$) hradby. Nakonec sečteme počet hodnot ležících pod dolní hradbou a nad horní hradbou. Výsledek lze vidět na obr. 1.1.

Poznámka. Při vykreslování krabicového grafu v jazyku se odlehlé a extrémní hodnoty nerozlišují, označují se jako tzv. outliery.

1.2 Konstrukce boxplotu

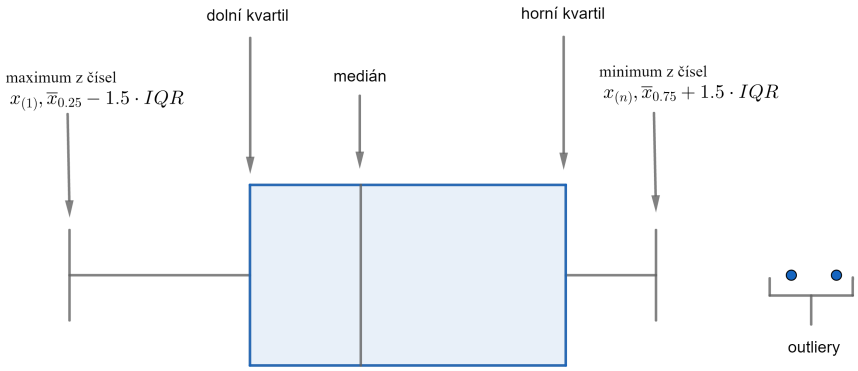
Při sestrojení boxplotu z datového souboru $X = \{x_1, \ldots, x_n\}$, kde $x_i \in \mathbb{R}$ pro $i = 1, \ldots, n$, postupujeme následujícím způsobem:

1. Data seřadíme podle velikosti od nejmenší hodnoty po největší $x_{(1)} \leq x_{(2)} \leq \ldots \leq \leq x_{(n)}$.
2. Nalezneme medián $\bar{x}_{0.5}$, dolní kvartil $\bar{x}_{0.25}$a horní kvartil $\bar{x}_{0.75}$, a na základě těchto hodnot sestrojíme krabičku.
3. Spočítáme mezikvartilové rozpětí IQR, jeho 1.5-násobek a poté určíme dolní hradbu jako $\max\{x_{(1)}, \bar{x}_{0.25} - 1.5IQR\}$ a horní hradbu jako $\min\{x_{(n)}, \bar{x}_{0.75} + 1.5IQR\}$.
4. Hodnoty ležící za hradbami vykreslíme jako jednotlivé body.

Výsledný graf je znázorněn na obr. 1.2.

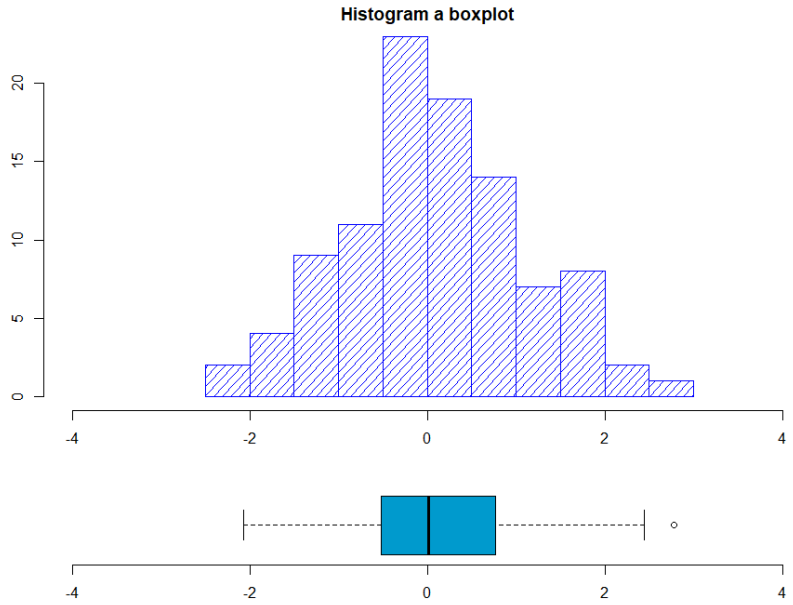
1.3 Porovnání boxplotu s histogramem

Histogram v porovnáním s boxplotem zobrazuje více informací ohledně distribuce dat, naopak na krabicovém grafu je zřetelně vidět medián a odlehlé hodnoty. Z obou grafů lze



Obrázek 1.2: Boxplot

vyčíst symetrii, popřípadě asymetrii dat. Na obrázku 1.3 vidíme oba grafy pro nagenerovaná data ($n = 100$) z normálního rozdělení se střední hodnotou $\mu = 0$ a směrodatnou odchylkou $\sigma = 1$ (v  za pomocí příkazu `rnorm()`).



Obrázek 1.3: Porovnání histogramu s boxplotem.

1.4 Boxploty v R a příklady na jejich využití

Krabicové grafy jsou častým způsobem grafické vizualizace dat a dají se proto vykreslit v téměř každém statistickém softwaru, jako jsou `Matlab, STATISTICA, Microsoft Excel` nebo `SAS`. My se zaměříme na zmiňovaný volně dostupný programovací jazyk R.

Základní syntax pro vykreslení boxplotu v jazyku R je ve tvaru `boxplot(`x`)`, kde `x` je libovolný číselný vektor, pro který chceme nakreslit graf. Chceme-li nakreslit více boxplotů vedle sebe do stejného grafu, je možné do argumentu vložit seznam (list) nebo tzv. data frame, který má ve svých složkách číselné vektory.

Možné volitelné parametry jako určení rozsahu vous nebo horizontální vykreslení lze najít použitím příkazu `?boxplot`.

Příklad 1. Ukážeme si, jak v jazyce R vykreslit krabicové grafy, máme-li k dispozici číselné vektory. Pro jednoduchost si vygenerujeme náhodné vektory, které se řídí normálním rozdělením $N(\mu, \sigma^2)$. K tomu využijeme příkaz `rnorm(n, mean = , sd = )`, kde n je počet čísel, které chceme nagenerovat, argument `mean` značí střední hodnotu μ a `sd` značí směrodatnou odchylku σ. Nagenerujeme si 3 vektory délky $10, 100$ a 1000 z normálního rozdělení $N(0,1)$ za pomocí příkazů

```
a <- rnorm(10)
b <- rnorm(100)
c <- rnorm(1000)
```

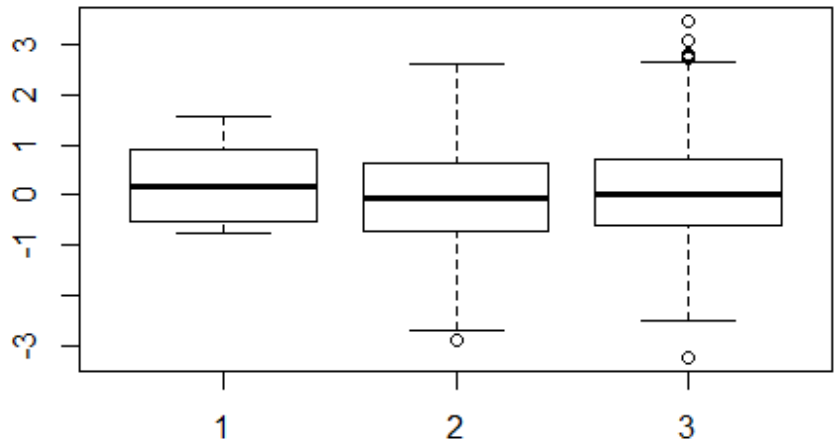
a následně si vykreslíme boxplot s těmito vektory příkazem

```
boxplot(list(a,b,c)).
```

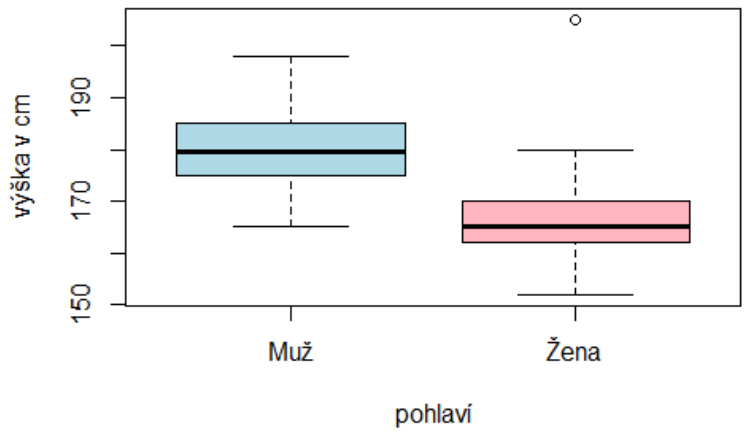
Grafy jednotlivých vektorů lze vidět na obr. 1.4.

Příklad 2. Vykreslení krabicových grafů vedle sebe umožňuje rychlé a snadné porovnání charakteristik několika souborů dat. V následujícím obrázku 1.5 jsou zobrazeny boxploty výšek mužů a žen. Data byla získána z průzkumu provedeného na vzorku 98 lidí ve věku od 14 do 62 let prostřednictvím dotazníku na sociální síti.

Na první pohled vidíme, že krabička boxplotu je u mužů umístěn výše než u žen, což se dalo předpokládat. Z grafu lze snadno vyčíst medián který u mužů činí zhruba 180 cm, zatímco u žen je to asi 167 cm. Dále podle polohy mediánu v krabičce a délky obou vous lze vyčíst symetrii, popřípadě asymetrii. U mužů se medián nachází přesně ve středu krabičky a obě vousy jsou zhruba stejně dlouhé, což naznačuje symetrii v datech. U žen je naopak medián položen v dolní polovině krabičky a data jsou tím pádem lehce zešikmená. Na obrázku lze také zpozorovat hodnotu, která se výrazně liší od ostatních. Může se jednat o špatně zadanou hodnotu, popřípadě velmi vysokou respondentku a záleží tedy na osobě provádějící analýzu dat, zda tuto hodnotu použije nebo ji vyloučí ze souboru.



Obrázek 1.4: Boxploty jednotlivých vektorů.



Obrázek 1.5: Výška lidí v závislosti na pohlaví.

1.5 Varianty boxplotu

Původně byl krabicový graf navržen pro ruční počítání, v dnešní době však máme k dispozici počítače a vznikly tak různé varianty boxplotu, které jsou komplexnější, zpravidla obsahující více informací ohledně distribuce dat. My se podíváme na několik nejběžnějších typů boxplotu.

1.5.1 Boxploty s proměnlivou šířkou

Jak nám název napovídá, šířka "krabiček"těchto boxplotů závisí na rozsahu dat každé skupiny. Zpravidla je šířka jednotlivých obdélníků přímo úměrná druhé odmocnině velikosti skupiny. Díky této přidané informaci je, lze snadněji posoudit charakter dat a vyhnout se případné nesprávné interpretaci. Tyto boxploty lze vidět na obrázku 1.6b.

1.5.2 Boxploty se zářezy

Boxploty se zářezy, také nazývané "zubaté boxploty", mají zúženou střední část krabičky a navíc zobrazují konfidenční intervaly okolo mediánu pomocí "zářezů". Délka konfidenčních intervalů je určena tak, aby nepřekrývající se intervaly (zářezy) naznačovaly statisticky významný (obvykle na hladině významnosti 5%) rozdíl mezi mediány skupin dat.

Šířka zářezů je přímo úměrná mezikvartilovému rozpětí IQR vzorku dat a nepřímo úměrná druhé odmocnině počtu pozorování n pro jednotlivé skupiny dat. Velikost zářezů okolo mediánů $\bar{x}_{0.5}$ lze vypočítat pomocí vzorce

$$\bar{x}_{0.5} \pm C \cdot s, \tag{1.5.1}$$

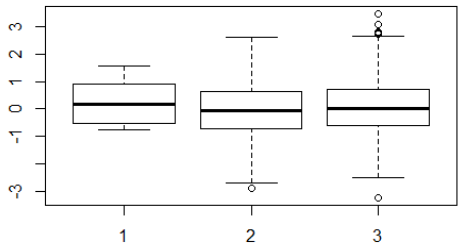
kde C je konstanta, za kterou obvykle volíme 1.7 (důvod pro tuto konvenci je podrobně sepsán v [17]) a s je výběrová směrodatná odchylka mediánu $\bar{x}_{0.5}$ daná

$$s = \frac{1.25 \cdot IQR}{1.35 \cdot \sqrt{n}}. \tag{1.5.2}$$

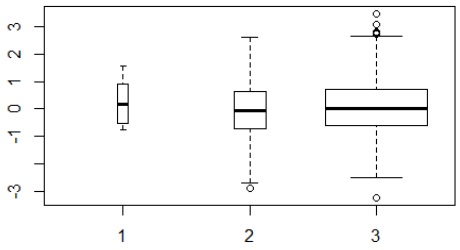
Boxploty se zářezy jsou vykresleny na obrázku 1.6c.

Poznámka. Existuje také varianta boxplotu s proměnlivou šířkou se zářezy, popsaná v [17], která kombinuje dvě výše popsané varianty, viz obrázek 1.6d.

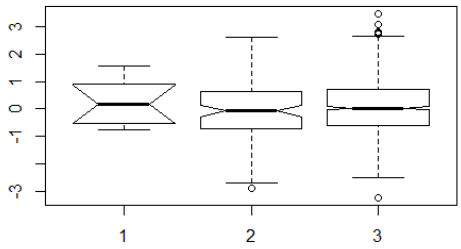
Boxplot s proměnlivou šířkou se dá v  vykreslit pomocí funkce `boxplot()` s argumentem `varwidth = TRUE`. Variantu se zářezy lze získat přidáním argumentu `notch = TRUE`. Obě tyto varianty lze zkombinovat.



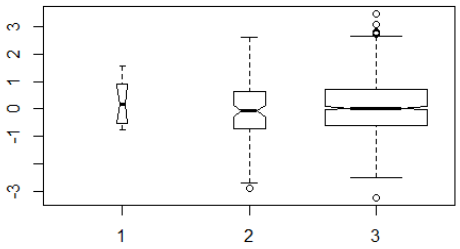
(a) Klasický boxplot.



(b) Boxplot s proměnlivou šířkou.



(c) Boxplot se zářezy.



(d) Boxplot s proměnlivou šířkou se zářezy.

Obrázek 1.6: Grafy 1.6a, 1.6b, 1.6c a 1.6d ukazují rozdíly mezi jednotlivými, výše uvedenými varianty. Data jsou stejná jako v příkladě 1, jedná se tedy o 3 náhodně nagenerované vektory z normálního rozdělení se střední hodnotou $\mu = 0$ a rozptylem $\sigma^2 = 1$, délky $10, 100$ a 1000.

Uvedené typy boxplotu se zpravidla používají při porovnávání více souborů dat. Vykreslením klasického krabicového grafu přicházíme o informace ohledně velikosti jednotlivých souborů, to lze napravit použitím varianty boxplotu s proměnlivou šířkou, na obr. 1.6b. Vidíme, že nám šířky krabic dávají jistou představu o rozsazích datových výběrů. Varianta se zářezy se využívá, chceme-li porovnat mediány mezi několika datovými soubory. Na obr. 1.6c se všechny zářezy překrývají a zamítáme tedy hypotézu o rozdílu mezi mediány na hladině významnosti 5 %.

Jelikož je boxplot diagnostickým grafem a snažíme se při jeho vykreslení zjistit o datech co nejvíce, vyplatí se obě tyto varianty zkombinovat.

Kapitola 2

Bagplot

Bagplot je zobecněním jednorozměrného boxplotu a slouží tedy k vizualizaci dvoudimenzionálních statistických dat. Byl poprvé představen v roce 1999 významnými statistiky P. J. Rousseeuwem, I. Rutsem a J. W. Tukeym v článku *The Bagplot: A Bivariate Boxplot* [12]. Boxplot využívá ke své konstrukci kvantily, které jsou definovány pouze pro jednorozměrné náhodné veličiny. Konstrukce bagplotu je založena na tzv. poloprostové hloubce (definované Johnem W. Tukeym v článku [15]), která zobecňuje kvantily na vícerozměrném prostoru.

Bagplot obsahuje 3 hlavní komponenty:

1. *bag* - polygon obsahující 50% všech "prostředních" hodnot, bývá vykreslen plnou čarou a jeho vnitřek je vybarven tmavší barvou (obdoba krabice)

2. *fence* - získá se trojnásobným zvětšením části *bag*, rozděluje "vnitřní" hodnoty od odlehlých hodnot, na grafu není nijak zaznačen

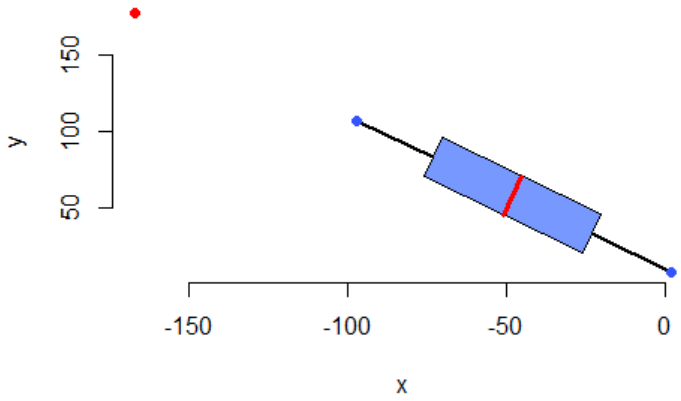
3. *loop* - část grafu obsahující body, které leží mezi *bag* a *fence*, bývá vykreslen světlejší barvou než *bag*

Kromě těchto částí, obsahuje bagplot *hloubkový medián* (analogie jednorozměrného mediánu), který se nachází v centru grafu a je obvykle značen jako křížek. Odlehlé hodnoty jsou zaznačeny červenými hvězdičkami a bývají popsány.

K bagplotu je možné přidat vousy jako tomu bylo u boxplotu. V tomto případě se jedná o úsečky vedoucí z mediánu ke každému bodu z *loopu*. Později si ukážeme, že v případě větších datových souboru se bagplot s vykreslenými vousy stává nepřehledným a je tedy lepší je z grafu vynechat.

Podobně jako v jednorozměrném případě, budeme chtít zobrazit některé charakteristiky dat pomocí bagplotu: pozici hloubkového mediánu, rozsah dat (velikost *bagu*), korelaci mezi proměnnými (orientace *bagu* a *loopu*), šikmost dat (tvar *bagu* a *loopu*), a chvosty (body v blízkosti hranice *loopu* a odlehlých hodnot).

Pro velmi "ploché" vícerozměrné datové soubory se komponenta *bag* stává krabičkou, a máme tedy klasický boxplot. V tomto případě světlejší část *loop* odpovídá vousům jednorozměrného krabicového grafu. Tato situace je znázorněna na obr. 2.1.



Obrázek 2.1: Bagplot pro lineární data.

Bagploty se dají vykreslit za pomocí statistických softwarů `MATLAB`, `S-Plus`[1], popřípadě v jazyce R s použitím balíčku `aplpack`.

[1] Dostupné z http://www.solutionmetrics.com.au/products/splus/default.html

2.1 Zobecnění známých pojmů

Naším prvním úkolem bude jako v případě konstrukce boxplotu si seřadit naše pozorování podle "velikosti". Máme-li k dispozici jednorozměrný datový soubor, není problém si ho uspořádat: pokud soubor obsahuje čísla, seřadíme je podle velikosti a v případě, že obsahuje znaky, lze jim přiřadit číselnou hodnotu a poté je seřadit podle velikosti (např. v souboru {Ano, Ne} "Ano" označíme 1 a "Ne" hodnotou 0, nebo v souboru {základní vzdělání, střední bez maturity, střední s maturitou, absolvent VŠ} lze "základní vzdělání" označit hodnotou 0, "střední bez maturity" hodnotou 1 atd. Výjimku tvoří tzv. nominální data (např. krevní skupiny, barvy), která nelze nijak uspořádat.

Seřazení dat podle velikosti je zároveň jedním z prvních kroků, které učiníme při analýze dat, jelikož nám usnadňuje výpočet základních statistických charakteristik jako je medián nebo modus. Pokud však máme vícerozměrný datový soubor, najít takové uspořádání není tak prosté.

Dalším možným způsobem, jak lze uspořádat číselnou množinu, je řazení z vnějšku směrem dovnitř. Toho docílíme tak, že každému číslu přiřadíme tzv. hloubku, tedy největšímu, resp. nejmenšímu číslu, přiřadíme hloubku 1, druhému největšímu, resp. druhému nejmenšímu číslu, přiřadíme hloubku 2 atd. Výhodou tohoto typu řazení je jednodušší rozšíření do vícerozměrných prostorů. Nejprve si zadefinujeme hloubku bodu pro jednorozměrná data.

2.1.1 Poloprostorová hloubka a její vlastnosti

Definice 3. Hloubka bodu $z \in \mathbb{R}$ z jednorozměrných dat $X = \{x_1, x_2, \ldots, x_n\}$, kde $x_i \in \mathbb{R}$ pro $i = 1, \ldots, n$, je definována jako

$$depth_1(z;X) = \min(\#\{i; x_i \leq z\}, \#\{i; x_i \geq z\}),$$

kde $\#\{\cdot\}$ je kardinalita množiny $\{\cdot\}$.

Poznámka. Hloubka bodu je tedy minimum z počtu bodů x_i ležících nalevo od bodu z a z počtu bodů ležících napravo od bodu z.

Poznámka. Medián je bod (resp. body) s maximální hloubkou.

$depth_1(x;X)$ je tedy funkce, která každému bodu $x \in \mathbb{R}$ přiřadí jeho hloubku vzhledem k souboru $X = \{x_1, x_2, \ldots, x_n\}$.

John W. Tukey byl první, kdo uvedl definici pojmu hloubky pro vícerozměrná data, známý jako *poloprostorová hloubka* (také *Tukeyho hloubka*). Poloprostorová hloubka nám umožní zobecnit pojem kvantil (jak si ukážeme později), který byl klíčový při konstrukci boxplotu, jelikož právě díky němu jsme mohli definovat pojmy jako medián, 1. a 3. kvartil. Existují celkem 2 definice poloprostorové hloubky, první z nich je založena na distribuční funkci a druhá je založena na datovém výběru. Protože pracujeme s konkrétními daty, uvedeme si zde druhou variantu. Ačkoliv by pro naše účely vystačila definice hloubky pro dvourozměrná data, zadefinujeme si ho zde obecně pro $p-$rozměrný prostor.

Definice 4. Poloprostorová hloubka $depth_p(\mathbf{z};X)$ bodu $\mathbf{z} \in \mathbb{R}^p$ z p-rozměrných dat $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, kde $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T \in \mathbb{R}^p, i = 1, \ldots, n$, jsou jednotlivá pozorování,

se definuje jako nejmenší hloubka bodu $\mathbf{z}$ v libovolné jednorozměrné projekci datového souboru: je-li $\mathbf{u}$ je libovolný vektor z $\mathbb{R}^p$ takový, že $\|\mathbf{u}\| = 1$, pak $\{\mathbf{u}^T\mathbf{x}_i\}_{i=1}^n$ je množina jednorozměrných projekcí souboru X a definujeme

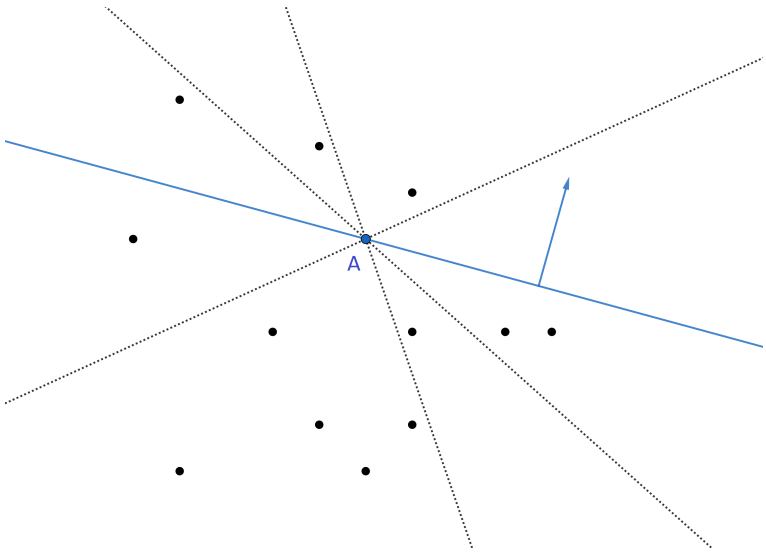
$$depth_p(\mathbf{z};X) = \min_{\|\mathbf{u}\|=1} depth_1(\mathbf{u}^T\mathbf{z};\{\mathbf{u}^T\mathbf{x}_i\}_{i=1}^n), \tag{2.1.1}$$

a ekvivalentně

$$depth_p(\mathbf{z};X) = \min_{\|\mathbf{u}\|=1} \#\{i : \mathbf{u}^T\mathbf{x}_i \geq \mathbf{u}^T\mathbf{z}\}. \tag{2.1.2}$$

Poloprostorovou hloubku lze taky chápat jako nejmenší počet bodů $\mathbf{x}_i$ ležících v uzavřeném poloprostoru s hraniční přímkou procházející přes bod $\mathbf{z}$. V dvourozměrném prostoru bychom vypočítali hloubku postupnou rotací přímky procházející přes bod $\mathbf{z}$ o 180° a spočítali nejmenší počet bodů na polorovinách vytvořených rotující přímkou, viz obr. 2.2.

Máme-li datový soubor $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ s n pozorováními, lze si vypočítat jejich hloubku, na jejímž základě seřadíme jednotlivá pozorování sestupně. Dostaneme uspořádaný soubor $\mathbf{x}_{(1)}, \mathbf{x}_{(2)}, \ldots, \mathbf{x}_{(n)}$, kde $\mathbf{x}_{(1)}$ má největší hloubku, je tedy nejcentrálnější a $\mathbf{x}_{(n)}$ má nejmenší hloubku a je tím pádem nejodlehlejší. Od klasického uspořádání jednorozměrných dat na číselné ose, které začíná nejmenší hodnotou a končí největší, se liší v tom, že začíná "uprostřed" a postupně se "šíří" všemi směry.



Obrázek 2.2: Hloubka bodu v rovině. Na obrázku je celkem 12 bodů z datového souboru, označených černě a bod $\mathbf{A} \in \mathbb{R}^2$, jehož hloubku chceme zjistit. Hledáme polorovinu, ve které se nachází co nejméně bodů ze souborů, na obrázku je hraniční přímka označena modře a odpovídající polorovina naznačena šipkou. Bod $\mathbf{A}$ má tedy hloubku 3. Platí, že všechny krajní body ze souboru mají hloubku 1, body ležící vně konvexního obalu tvořený jednotlivými pozorováními mají hloubku 0 a body náležící do datového souboru mají hloubku aspoň 1.

Funkce $depth_p(\mathbf{z};X)$ nám udává jakousi polohu bodu v datovém souboru. Právě díky ní můžeme zavést důležitý pojem "střed" vícerozměrného datového souboru. V našem případě to bude část grafu zvaná *bag*, která obsahuje 50 % všech hodnot nacházejících se

nejcentrálněji v souboru. Platí, že čím blíž se pozorování nachází ve středu souboru, tím větší bude jeho hloubka a naopak, čím odlehlejší je pozorování, tím menší bude jeho hloubka.

Poznámka. Poloprostorová hloubka byla sice první hloubkovou funkcí, od doby svého vzniku v roce 1975 se však objevily další funkce podobného typu. Příkladem je *simplexová hloubka*, kterou v roce 1990 definovala Regina Y. Liu v článku [6]. Princip této funkce stojí na tom, že v dvourozměrném prostoru libovolné 3 body n-rozměrného datového souboru tvoří trojúhelník (popř. úsečku, ale to pro naše účely není podstatné), dostáváme celkem $\binom{n}{3}$ trojúhelníků a libovolnému bodu $\mathbf{z} \in \mathbb{R}^2$ pak přiřadíme hloubku jako hodnotu odpovídající počtu trojúhelníků, ve kterých leží. Opět platí, že hloubka bude tím větší, čím blíž se bod nachází v centru souboru.

Další hloubkové funkce jsou například *oja depth* (popsaná v [8]), *projekční hloubka* (z ang. *projection depth* definovaná v [3]) nebo *prostorová hloubka* (z ang. *spatial depth*, viz [14]). Hloubková funkce je, volně řečeno, jakákoliv nezáporná reálná funkce, která dává bodům z datového souboru lineární uspořádání.

Algoritmus pro počítání poloprostorové hloubky $depth_2(\mathbf{z};X)$ dvourozměrných dat s lineárně logaritmickou složitostí - $O(n\log(n))$ navrhli Rousseeuw a Ruts v roce 1996 v článku [9]. Ve vyšších dimenzích je výpočet hloubky časově náročnější kvůli její definici zahrnující nekonečně mnoho projekcí. Hlavní myšlenkou existujících algoritmů je použití projekce do prostoru s nižší dimenzí. Složitost výpočtu hloubky libovolného bodu $\mathbf{x} \in \mathbb{R}^p$ vzhledem k p-rozměrnému datovému souboru je potom $O(n^{p-1}\log(n))$.

Výše definovaná *poloprostorová hloubka* je jednou z nejpoužívanějších hloubkových funkcí, jednak díky své intuitivní definici, a jednak protože splňuje všechny požadované vlastnosti hloubkových funkcí zmíněné v článku [6], poté i dokázané v [18] a těmi jsou:

1. Afinní invariance
2. "Maximalita" v centru symetrie (pokud existuje)
3. Monotonie vzhledem k nejhlubšímu bodu - pokud se libovolný bod vzdaluje od nejhlubšího bodu po přímce, jeho hloubka klesá monotónně
4. Konvergence (hloubky) k 0 pro body vzdalující se do nekonečna od nejhlubšího bodu

Nás hlavně zajímá bod 1, jelikož při analýze dat často transformujeme data - například převody jednotek nebo měn. Díky vlastnosti afinní invariance tedy při posunutí nebo nesingulární lineární transformaci jako jsou například zrcadlení, měnění měřítek nebo rotace (popřípadě jejich kombinací) dat, se bude bagplot transformovat odpovídajícím způsobem. Tedy body, které byly v *bagu*, zůstanou v *bagu*, odlehlé hodnoty zůstanou odlehlými apod. Uvedeme si zde několik vlastností poloprostorové hloubky, které byly studovány D. L. Donohem a M. Gaskem (viz [2]).

Lemma 2.1.1. *Poloprostorová hloubka je invariantní vůči afinním zobrazením:*

$$depth_p(\mathbf{Az}+\mathbf{b};\mathbf{A}X+\mathbf{b}) = depth_p(\mathbf{z};X)$$

pro všechna $\mathbf{b} \in \mathbb{R}^p$ *a každé nesingulární lineární zobrazení zadané maticí* $\mathbf{A}$ *rozměru* $p \times p$.

Důkaz. Důkaz je popsán v [2]. □

Hloubka libovolného bodu $\mathbf{z} \in \mathbb{R}^p$ se tedy nemění posunutím, popřípadě lineární transformací bodu $\mathbf{z}$ nebo souboru X. To znamená, že poloprostorová hloubka nezávisí na zvolené souřadnicové soustavě.

Afinní invariance se využívá také v algoritmu výpočtu poloprostorové hloubky v dvourozměrném prostoru, kde princip spočívá v posunutí bodu $\mathbf{z} \in \mathbb{R}^2$, jehož hloubku hledáme, do počátku a poté v převedení z kartézských souřadnic bodů do polárních souřadnic. Výpočet se ještě usnadní projekcí všech zbylých bodů na jednotkovou kružnici se středem v počátku (tedy v bodě $\mathbf{z}$), poté stačí najít nejmenší počet bodů ležící na půlkružnici s hraniční přímkou procházející přes počátek.

2.1.2 Hloubkové oblasti a kontury

Pokud bychom si vypočítali hloubku všech bodů v rovině vzhledem k datovému výběru, zjistili bychom, že mohou nabývat stejných hodnot. To by nemělo být překvapivé, jelikož všechny krajní body datového souboru mají hloubku 1. V jednorozměrném případě, body mající stejný kvantil, leží na témže bodu na číselné ose. Podívejme se nyní, jak to vypadá ve vícerozměrných prostorech.

Definice 5. Nechť $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, kde $\mathbf{x}_i \in \mathbb{R}^p$ pro $i = 1, \ldots, n$, je datový soubor. Uvažujme množinu $D_k = \{\mathbf{z} \in \mathbb{R}^p \mid depth(\mathbf{z}; X) \geq k\}$. D_k nazveme **oblastí hloubky k** a hranice $h(D_k)$ **hloubkovými konturami**.

Podle (2.1.2) je tato množina průnikem všech p-rozměrných poloprostorů obsahujících aspoň $n - 1 + k$ bodů datového souboru X. Hloubkové kontury tedy tvoří konvexní mnohoúhelník, jehož vrcholy jsou buď přímo pozorování $\mathbf{x}_i \in X, i = 1, \ldots, n$ nebo jsou to průsečíky dvou přímek procházejících dvěma pozorováními ze souboru X (převzato z [11]). Konstrukce hloubkových kontur je znázorněna na obr. 2.3. Navíc platí, že každý bod z datového souboru musí náležet aspoň do jedné hloubkové oblasti.

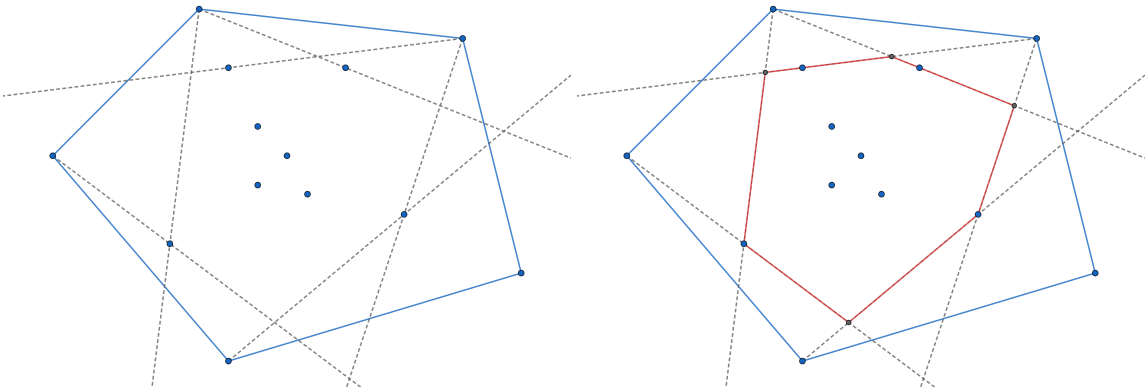
Hloubkové kontury byly předmětem studie v článku [2] a my si zde uvedeme jeden z jeho výsledků.

Lemma 2.1.2. *Hloubkové oblasti tvoří posloupnost vnořených konvexních množin: D_k je konvexní a platí $D_{k+1} \subset D_k$.*

Důkaz. Důkaz je popsán v [2]. □

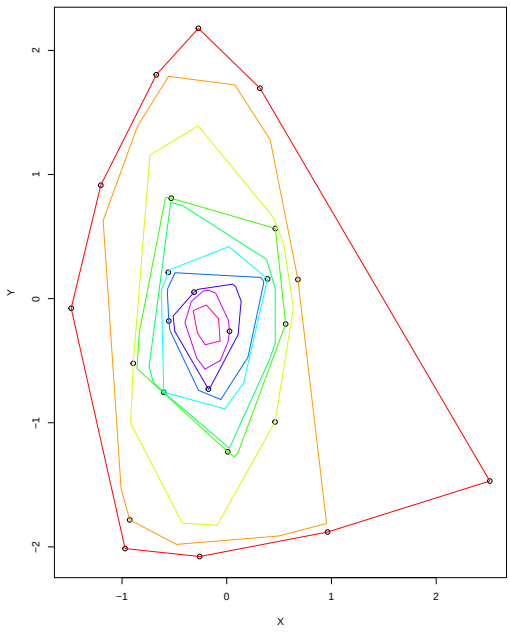
Pro názornou ukázku hloubkových kontur je na obr. 2.4 vykresleno celkem 100 pozorování z dvourozměrného normální rozdělení a jejich příslušné hloubkové kontury. Všimněme si, že odlehlá pozorování ovlivní pouze vnější kontury, v našem případě jenom jednu. Tvary kontur nám také dávají představu o tom, jaký tvar mají naše data. Jelikož máme na obrázku data pocházející z dvourozměrného normálního rozdělení (patří do tzv. eliptických rozdělení, které je zobecňují), mají kontury elipsovitý tvar.

Hloubkové kontury můžeme chápat jako vícerozměrnou analogii kvantilu: podobně jako pozorování, která mají stejnou velikost, mají rovněž stejný kvantil, body z vícerozměrných dat mající stejnou hloubku leží na společné kontuře, a mimo jiné také rozdělují datový soubor na určitý počet částí. Díky hloubkovým oblastem budeme moci zkonstruovat část



(a) (b)

Obrázek 2.3: Ilustrace konstrukce hloubkových kontur. Pozorování z výběru jsou označeny modrými tečkami. Vnější hloubková kontura (kontura hloubky 1) je tvořena konvexním obalem množiny, na obrázku zakreslena plnou modrou čarou. Vrcholy kontury hloubky 2 tvoří buď přímo body z výběru nebo průsečíky přímek procházející přes 2 pozorování, výsledná kontura je na obr. 2.3b vyznačena plnou červenou čarou.



Obrázek 2.4: Hloubkové kontury pro náhodně generovaná data dvourozměrného normálního rozdělení. Obrázek byl vykreslen v za použití funkce `isodepth` z balíčku `depth`.

bag a následně jejím zvětšením i komponentu *fence*. Jak se tyto části sestrojí, si ukážeme v podkapitole 2.2.

První algoritmus pro výpočet hloubkových kontur se složitostí $O(n^2 \log n)$ zvaný *ISODEPTH* navrhli I. Ruts a P. Rousseeuw v roce 1996, viz [13]. Pro větší datové soubory ($n > 1000$) je však jeho použití nepraktické. Rychlejší a účinnější algoritmus se složitostí

$O(n^2)$ byl navržen v článku [7].

2.1.3 Tukeyho medián

Na základě poloprostorové hloubky si můžeme definovat analogii klasického mediánu a tím je mnohorozměrný medián, jenž je v centru bagplotu a kolem kterého se pak vykreslují ostatní části grafu (zejména *bag* a *loop*). Tentokrát se podíváme pouze na dvourozměrný případ.

Definice 6. Tukeyho medián (také **poloprostorový medián**) z datového výběru $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ definujeme jako bod $\mathbf{T}^*$, pro který platí $\mathbf{T}^* = k^*(X)$, kde

$$k^*(X) = \max_{\mathbf{z} \in \mathbb{R}^2} (depth(\mathbf{z}; X)).$$

Poznámka. Podle definice tedy Tukeyho medián nemusí být bod z datového výběru.

Poznámka. Jelikož hloubková funkce $depth_2(\mathbf{z}; X)$ nabývá hodnot od 0 po n, takový bod $\mathbf{z}$ vždy existuje.

Poznámka. Podle definice nemusí být poloprostorový medián určen jednoznačně (existuje-li více bodů s maximální hloubkou). V takových případech volíme za $\mathbf{T}^*$ těžiště nejhlubší oblasti D_{k^*}.

Můžeme také uvažovat pouze body z našeho souboru X a medián bychom pak definovali jako bod $T^\circ = k^+(X)$, kde

$$k^+(X) = \max_{i=1,\ldots,n} (depth(\mathbf{x}_i; X)).$$

Zřejmě platí $1 \leq k^+(X) \leq k^*(X)$. Důvod, proč bychom volili takovou definici mediánu je jeho jednodušší výpočet, jelikož nemusíme uvažovat všechny body $\mathbf{x} \in \mathbb{R}^2$. V případě, že máme "rozumná" data, může nám T° dávat poměrně přesnou aproximaci skutečného mediánu $\mathbf{T}^*$.

Velkou výhodou klasického jednorozměrného mediánu je jeho tzv. robustnost. To znamená, že není citlivý na odlehlé hodnoty, na rozdíl od průměru, na který mají extrémní hodnoty výrazný vliv. Ukažme si to na následujícím příkladu. Předpokládejme, že provádíme měření teploty a získali jsme následující hodnoty:

$$17.15, \quad 17.01, \quad 17.23, \quad 17.14, \quad 17.61$$

a chceme zjistit skutečnou hodnotu. Nejprve si vypočítejme průměr: $\bar{x} = 1/5(17.15 + 17.01 + 17.23 + 17.14 + 17.61) = 17.228$. Předpokládejme, že jsme se přepsali v jedné z hodnot a máme následující data:

$$17.15, \quad 170.1, \quad 17.23, \quad 17.14, \quad 17.61.$$

Vypočítáme-li si nyní průměr, dostaneme $\bar{x} = 47.846$, což zdaleka neodpovídá skutečné hodnotě. Pokud bychom si ovšem vypočítali medián, vyšel by nám $x_{0.5} = 17.23$, medián je tedy robustnějším odhadem.

Dalším možným způsobem, jak lze charakterizovat robustnost je pomocí tzv. *bodu selhání* (z ang. *breakdown point*, definovaný v [1]). Zjednodušeně řečeno, bod selhání statistického odhadu je podíl pozorování, která po nahrazení jinými, nepříznivými hodnotami, nám "pokazí" odhad. V případě výběrového průměru je tato hodnota rovna $1/n$, jelikož při změně pouze jedné hodnoty se průměr vychýlí. Naopak pro výběrový medián platí, že má vysoký bod selhání, přesněji 50 %.

Tuto důležitou vlastnost si zachoval i Tukeyho dvourozměrný medián, jenž má bod selhání větší nebo rovnu $1/3$ (dokázáno v článku [3]). Je možné tedy v datovém souboru $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ nahradit nejvýše $n/3$ pozorování, aniž by se medián $\mathbf{T}^*$ výrazně odchýlil od své původní hodnoty.

Toto pozorování je významné, neboť nechceme aby se náš výsledný graf posunul směrem k extrémním hodnotám.

Výpočet poloprostorového mediánu není zcela jednoduchý proces, a proto si zde uvedeme několik vět, které nám to usnadní. Předpoklad je, aby data byla v tzv. obecné pozici, viz následující definice 7.

Definice 7. Řekneme, že datový soubor $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ je v obecné pozici, jestliže v libovolném $(p-1)$-rozměrném podprostoru leží nejvýše p bodů.

Poznámka. Datový soubor je tedy v obecné pozici, obsahuje-li nejvýše dva body na libovolné přímce nebo nejvýše tři body v libovolné rovině atd.

Věta 2.1.3. *Je-li X v obecné pozici, pak maximální hloubka $k^*(X)$ leží mezi $\lceil n/(p+1) \rceil$ a $\lceil n/2 \rceil$.*

Důkaz. Důkaz je popsán v [2]. □

Pro $p = 2$ tedy platí, že bod $k^*(X) = \max_{\mathbf{z} \in \mathbb{R}^2}(depth(\mathbf{z}; X))$ má hloubku v rozmezí od $\lceil n/3 \rceil$ do $\lceil n/2 \rceil$. Horní hranice $\lceil n/2 \rceil$ se dá ještě snížit na $\lfloor n/2 \rfloor$ pro dvourozměrné případy:

Věta 2.1.4. *Pro dvourozměrný datový soubor X v obecné pozici platí*

$$k^*(X) \leq \lfloor n/2 \rfloor.$$

Důkaz. Důkaz je popsán v [11]. □

Základní myšlenka výpočtu Tukeyho mediánu spočívá v nalezení nejhlubší oblasti D_{k^*} a poté vypočítání těžiště. Věty 2.1.3 a 2.1.4 nám usnadňují výpočet tím, že dávají meze pro největší hloubku a za jejich použití tedy můžeme sestrojit několik oblastí D_k, kde $\lceil n/3 \rceil \leq k \leq \lfloor n/2 \rfloor$, k nalezení k^* za podmínky $D_{k^*} \neq \emptyset$ a $D_{k^*+1} = \emptyset$.

Samozřejmě platí, že čím větší je datový soubor, tím je interval $\left[\lceil n/3 \rceil, \lfloor n/2 \rfloor\right]$ širší a výpočet zabere více času. Detailní popis algoritmu výpočtu poloprostorového mediánu lze najít v článku [11].

2.2 Konstrukce bagplotu

Mějme datový soubor $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ s pozorováními $\mathbf{x}_i = (x_{i1}, x_{i2})^T \in \mathbb{R}^2$, kde $i = 1, \ldots, n$. Bagplot popsaný na začátku této kapitoly zkonstruujeme následujícím způsobem (konstukce je založena na popisu z článku [10]):

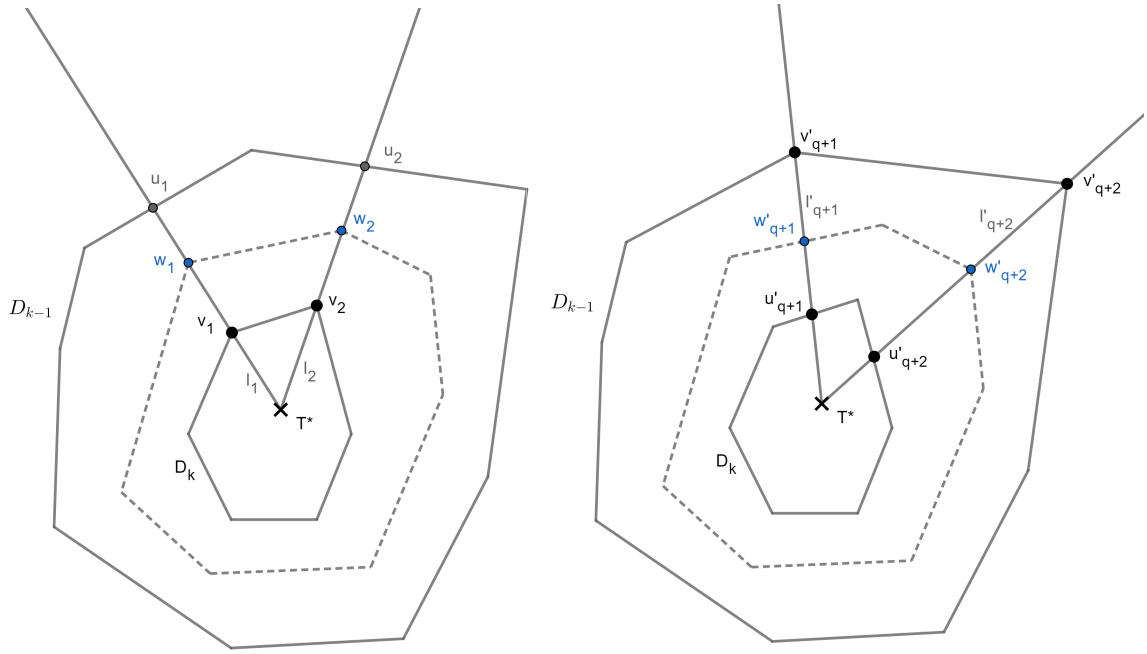
1. Nalezneme **Tukeyho medián** $\mathbf{T}^*$ (bod s největší hloubkou, popř. těžiště nejhlubší oblasti).

2. Zkonstruujeme komponentu **bag** (část grafu obsahující 50 % nejhlubších bodů):

   - Nalezneme dvě po sobě jdoucí hloubkové oblasti, pro které platí, že jedna obsahuje nejvýše polovinu bodů z datového souboru a druhá jich obsahuje více než polovinu. Označme tyto oblasti D_k a D_{k-1} a počet bodů ležících v nich $\#D_k$ a $\#D_{k-1}$. Platí $\#D_k \leq \lfloor n/2 \rfloor \leq \#D_{k-1}$.
   - Spočítáme parametr λ, který určuje relativní vzdálenost *bagu* od obou oblastí D_k a D_{k-1}, následovně:
   $$\lambda = \frac{\lfloor n/2 \rfloor - \#D_k}{\#D_{k-1} - \#D_k}.$$
   - Vrcholy *bagu* získáme lineární interpolací mezi vrcholy hloubkových oblastí $\#D_k$ a $\#D_{k-1}$ pomocí parametru λ. Nechť $V = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_q\}$ jsou vrcholy hloubkové kontury D_k a označme $l_1, l_2, \ldots, l_q$ přímky procházející vrcholy $\mathbf{v}_i$, kde $i = 1, \ldots, q$, a mediánem $\mathbf{T}^*$. Průsečíky přímek l_i, kde $i = 1, \ldots, q$ s druhou konturou D_{k-1} označme $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_q$. Nechť $W = \{\mathbf{w}_1, \ldots, \mathbf{w}_q, \mathbf{w}_{q+1}, \ldots, \mathbf{w}_r\}$ jsou vrcholy *bagu*. Vrcholy $\mathbf{w}_i$, kde $i = 1, \ldots, q$ vypočítáme následovně:
   $$\mathbf{w}_i = \lambda \mathbf{v}_i + (1 - \lambda)\mathbf{u}_i, \quad \text{kde } i = 1, \ldots, q.$$
   Zbylé vrcholy *bagu* $\mathbf{w}_{q+1}, \ldots, \mathbf{w}_r$ spočítáme analogicky: nechť $V' = \{\mathbf{v}'_{q+1}, \mathbf{v}'_{q+2}, \ldots, \mathbf{v}'_r\}$ jsou vrcholy kontury D_{k-1} a $l'_{q+1}, \ldots, l'_r$ jsou přímky procházející těmito vrcholy a mediánem $\mathbf{T}^*$. Průsečíky těchto přímek s vnitřní konturou D_k označme $\mathbf{u}'_{q+1}, \ldots, \mathbf{u}'_r$, vrcholy $\mathbf{w}_j$, kde $j = q+1, \ldots, r$ jsou pak dány
   $$\mathbf{w}_j = \lambda \mathbf{u}'_j + (1 - \lambda)\mathbf{v}'_j, \quad \text{kde } j = q+1, \ldots, r.$$
   - Sestrojíme *bag* ze všech vrcholů W. Protože byly D_k a D_{k-1} konvexní mnohoúhelníky, bude *bag* opět konvexním mnohoúhelníkem. Konstrukce *bagu* je znázorněna na obr. 2.5.

3. Sestrojíme **fence** (není v grafu zaznačena) trojnásobným[2] zvětšením *bagu* vzhledem k mediánu $\mathbf{T}^*$. Nechť $W = \{\mathbf{w}_1, \ldots, \mathbf{w}_r\}$ jsou vrcholy *bagu* a $p_1, \ldots, p_r$ jsou polopřímky vedoucí z mediánu $\mathbf{T}^*$ přes jednotlivé vrcholy *bagu*. Nechť $k_i = \{\mathbf{x} \in \mathbb{R}^2; \|\mathbf{T}^*\mathbf{x}\| = 3 \cdot \|\mathbf{T}^*\mathbf{w}_i\|\}$, kde $i = 1, \ldots, r$. Vrcholy *fence* $\mathbf{f}_1, \ldots, \mathbf{f}_r$ získáme jako průsečíky polopřímky p_i, kde $i = 1, \ldots, r$ s odpovídající si kružnicí k_i. Tato konstrukce je znázorněna na obr. 2.6. Pozorování za hranicí komponenty *fence* se označují jako odlehlá.

[2] Tato hodnota je zvolena tak, aby v průměru 0.5 % pozorování z dvourozměrného normálního rozdělení byla považována za odlehlá, viz [10].

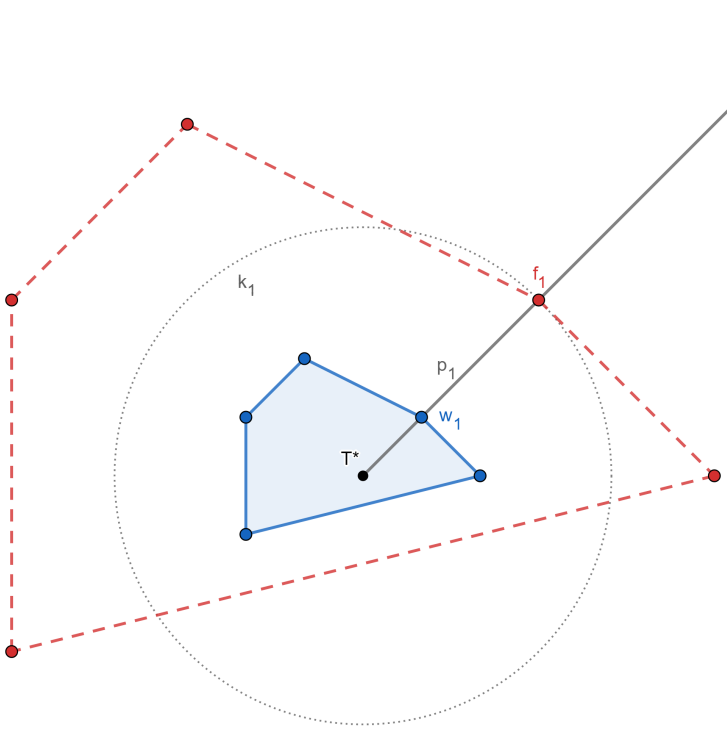


(a) (b)

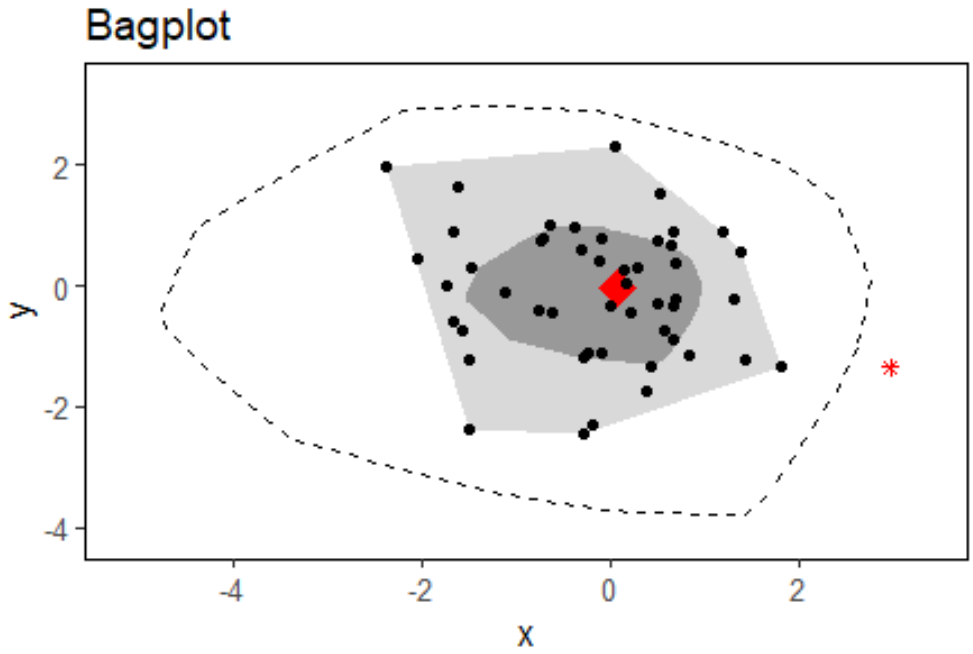
Obrázek 2.5: Sestrojení *bagu*. Na obr. (a) vedeme přímky l_1 a l_2 přes vrcholy $\mathbf{v}_1$ a $\mathbf{v}_2$ kontury D_k a mediánem $\mathbf{T}^*$, vrcholy *bagu* $\mathbf{w}_1$, resp. $\mathbf{w}_2$ dostáváme interpolací mezi vrcholy $\mathbf{v}_1$ a $\mathbf{u}_1$, resp. $\mathbf{v}_2$ a $\mathbf{u}_2$. Na obr. 2.5b zase vedeme přímky přes vrcholy kontury D_{k-1} a postupujeme analogicky. Výsledný polygon *bag* se sestrojí ze všech bodů $\mathbf{w}_k$, kde $k = 1, \ldots q, q+1, \ldots, r$.

4. Poslední část **loop** obsahuje všechny body nacházející se mezi *bagem* a *fence*. Vrcholy této komponenty jsou tedy tvořeny konvexním obalem všech neodlehlých pozorování.

Výsledný graf lze vidět na obr. 2.7.



Obrázek 2.6: Konstrukce *fence*. *Bag* je tvořen 5 body, na obrázku vykreslen modře. Jeho trojnásobným zvětšením vzhledem k mediánu $\mathbf{T}^*$ dostáváme *fence*, na obrázku vyznačen přerušovanou červenou čarou.



Obrázek 2.7: Bagplot z vygenerovaných dat s dvourozměrným normálním rozdělením ($n = 50$). Tukeyho medián je zaznačen červeným kosočtvercem uvnitř *bagu*, vnitřního polygonu vybarveného tmavou šedou obsahující 50 % nejhlubších bodů. Vnější polygon *loop* (vybarvený světlejší barvou) je tvořen všemi body ležící mezi *bagem* a *fence* (na obrázku vykreslený přerušovanou čarou), který vznikl trojnásobným zvětšením *bagu* vzhledem k mediánu. Máme pouze jednu odlehlou hodnotu zaznačenou červenou hvězdičkou.

2.3 Bagploty v R a příklady na jejich využití

Podívejme se nyní na několik příkladů, na kterých si ilustrujeme využití bagplotu. Nejprve si ukažme, jak lze bagplot sestrojit v jazyku R. K tomu je zapotřebí načíst jeden z následujících dvou balíčků: `aplpack` nebo `mrfDepth`. Oba tyto balíčky nabízí funkce `bagplot()` sloužící k vykreslení grafu. V této práci se zaměříme na první zmiňovaný `aplpack`, jelikož nabízí více modifikací výsledného grafu. Výhodou použití balíčku `mrfDepth` je možnost výpočtu částí grafu na základě jiné hloubky než poloprostorové (např. projekční hloubky), popřípadě vykreslení *fence*.

Základní syntax pro konstrukci bagplotu je ve tvaru `bagplot(x, y)`, kde `x`, resp. `y`, jsou x-ové, resp. y-ové, hodnoty pozorování z datového výběru. Příklady volitelných argumentů:

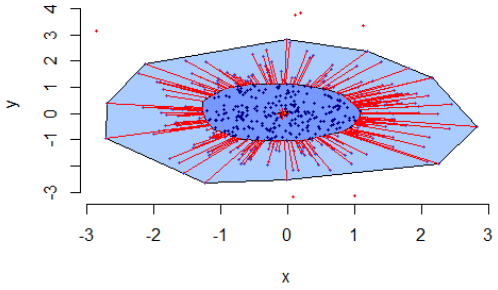
- `factor` - parametr určující velikost *fence* vzhledem k *bagu*, standardně nastaveno na 3
- `show.whiskers` - zobrazí vousy k pozorováním ležící mezi *bagem* a *fence*
- `show.looppoints` - zobrazí body ležící mezi *bagem* a *fence*
- `show.bagpoints` - zobrazí body v *bagu*
- `transparency` - vykreslí graf průhledně

Další možné argumenty lze najít přímo v programu použitím příkazu `?bagplot`.

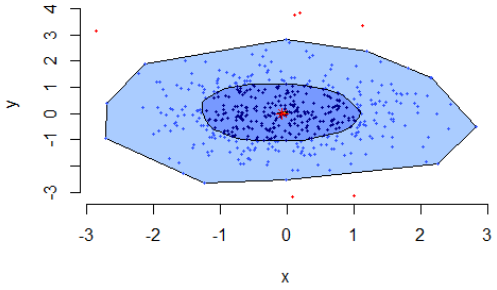
Příklad 3. Ukážeme si, jak vykreslit bagplot z číselných vektorů. Opět si pro jednoduchost vygenerujeme náhodné vektory x, y délky 500, které se řídí normálním rozdělením:

```
x <- rnorm(500)
y <- rnorm(500)
```

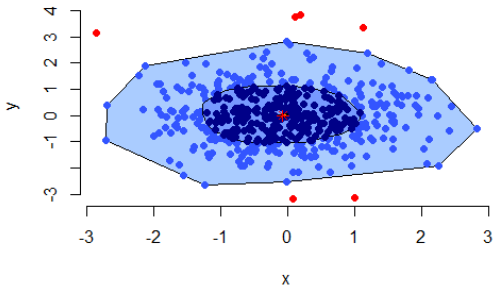
Zkusíme si sestrojit bagplot z těchto vektorů v základní formě bez jakýchkoliv argumentů příkazem `bagplot(x,y)`. Graf lze vidět na obr. 2.8a. Takto vykreslený graf však není přehledný, a to hlavně kvůli vousům. Zkusme je tedy odstranit nastavením argumentu `show.whiskers` na `FALSE`. Výsledek je na obr. 2.8b. Ještě můžeme lépe zviditelnit odlehlé hodnoty přidáním parametru `cex = 1`, viz obr. 2.8c. V případě velkého počtu pozorování je vhodné nevykreslovat vnitřní body a toho docílíme volbou argumentů `show.bagpoints = FALSE` a `show.looppoints = FALSE`, výsledný graf je na obr. 2.8d. Docílíme tak přehlednější vizualizace na úkor výhod korelačního diagramu (též dvourozměrného bodového grafu), jako jsou například rozpoznání shluků nebo děr v datech.



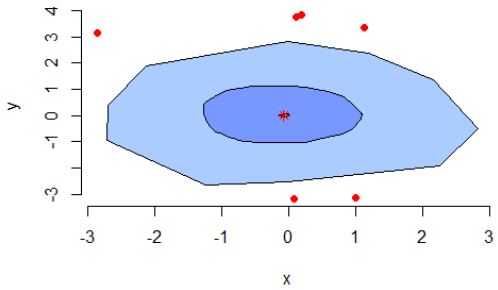
(a) Bagplot s vousy.



(b) Bagplot bez vousů.



(c) Bagplot se zvýrazněnými odlehlými body.

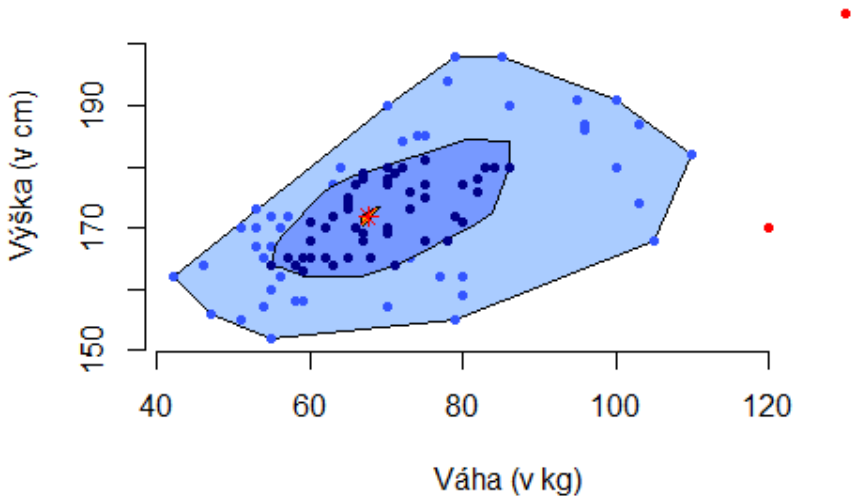


(d) Bagplot bez vnitřních bodů.

Obrázek 2.8: Různé varianty bagplotu.

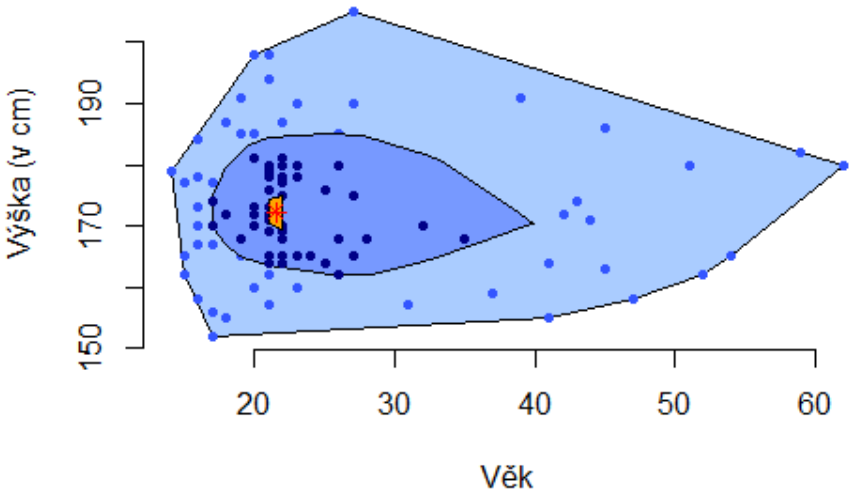
Příklad 4. Na tomto příkladu si ilustrujeme, jaké charakteristiky lze z dat vyčíst sestrojením bagplotu. Použijeme data z příkladu 2, máme tedy údaje o výšce a váze 98 lidí ve věku od 14–62 let. Nutno poznamenat, že vzorek dat je příliš malý a nereprezentuje tak skutečnost.

Na obr. 2.9 je vykreslený bagplot pro výšku a váhu lidí. Medián je na obrázku označen červenou hvězdičkou, vidíme tedy, že průměrný člověk váží necelých 70 kg a měří asi 170 cm. Orientace *bagu* a *loopu* naznačují kladnou korelaci mezi proměnnými. Pozice mediánu (zhruba uprostřed *bagu*) a eliptické tvary *bagu* a *loopu* svědčí o symetričnosti, trochu to kazí pravá část *loopu*, která je širší a data jsou tak lehce zešikmená (zřejmě je to dáno malým rozsahem výběru). Dále na obrázku vidíme dvě odlehlé hodnoty, zaznačené červenými tečkami, a několik pozorování jim blízkých, které natahují *loop*.



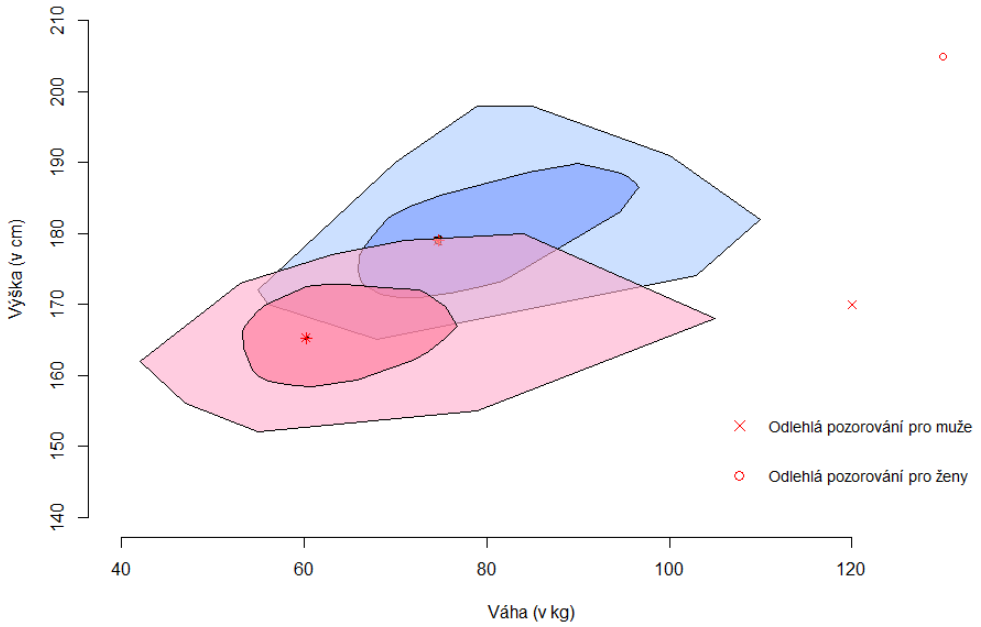
Obrázek 2.9: Výška lidí v závislosti na váze.

Na rozdíl od předchozího grafu 2.9, bagplot vykreslený na obr. 2.10 nenaznačuje žádnou korelaci mezi proměnnými. Data jsou silně zešikmená, lze to poznat jednak z pozice mediánu v *bagu* a jednak z pozice *bagu* v *loopu*, také nám k tomu napomáhají vykreslená pozorování. Graf sice nemá žádné odlehlé hodnoty, ale vidíme, že několik bodů, hlavně v pravé a horní části grafu, natahují hranice *loopu*.



Obrázek 2.10: Výška lidí v závislosti na věku.

Příklad 5. Boxplot si získal svou popularitu kompaktností a rychlým a přehledným zobrazením distribuce více vzorků dat, umožňující jejich porovnání, v jednom grafu. Zkusme si tedy vykreslit překrývající se bagploty do jednoho grafu pomocí argumentu `transparency = TRUE`. Použijeme data z minulého příkladu, tedy údaje o váze a výšce lidí, a vykreslíme si bagplot zvlášť pro muže a zvlášť pro ženy, viz obr. 2.11.



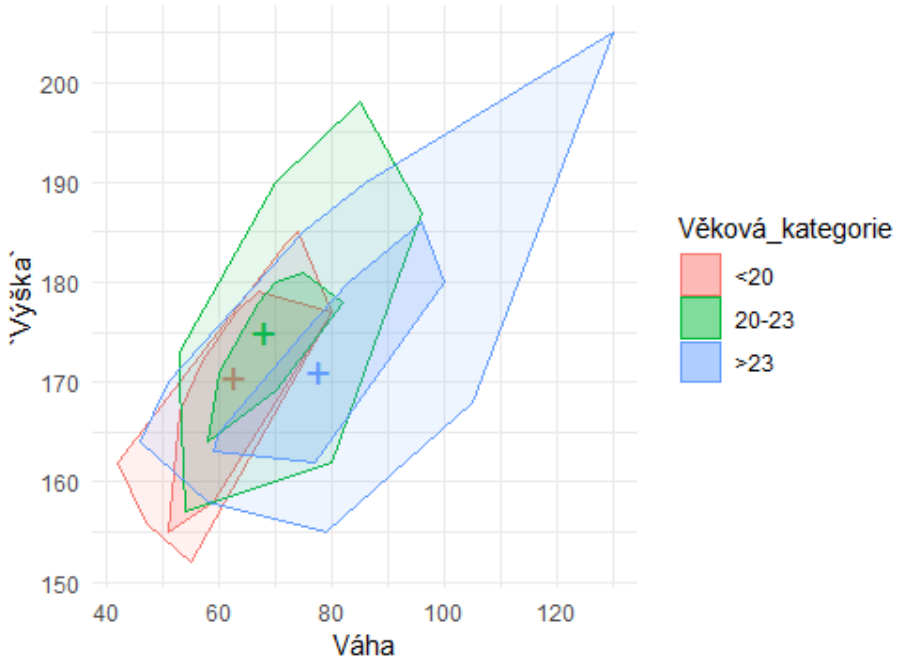
Obrázek 2.11: Bagploty ukazující rozdíl mezi váhou a výškou mezi muži (na grafu zaznačeni modře) a ženy (červeně). Porovnáme-li mediány, snadno z grafu vyčteme, že muži jsou v průměru o 15 cm vyšší a zhruba o 15 kg těžší. Jak *bag*, tak i *loop* jsou u mužů položeny výš a víc vpravo, a naznačují tak vyšší váhu i výšku. Na obrázku jsou také zřetelně vidět odlehlá pozorování pro obě skupiny dat.

Při vykreslení dvou bagplotů lze ještě z obrázku vyčíst potřebné informace, zkusme jich teď vykreslit tři do jednoho grafu. Rozdělíme si náš soubor na tři zhruba stejně velké skupiny na základě věku respondentů (pomocí příkazu `quantile()`). První skupinu tvoří respondenti mladší 20 let, druhou od 20 do 23 let (včetně) a třetí skupinu tvoří osoby starší 23 let. K vykreslení jednotlivých grafů na obrázku 2.12 jsem využil funkci `geom_bag`[3], která pouze vypočítá potřebné souřadnice k sestrojení bagplotu, samotný graf ale nevykreslí. K tomu jsem použil funkci `ggplot` z balíčku `ggplot2`[4], důvodem je jednodušší zápis a taky estetičtější vzhled.

Z obrázku 2.12 vidíme, že již při vykreslení tří překrývajících se bagplotů se graf stává nepřehledným. Autor diplomové práce [5] navrhuje při vizualizaci více skupin vykreslit pouze konvexní obal *bagu*, popřípadě pouze konvexní obal *loopu*, a jednotlivé skupiny potom odlišit různými barvami.

[3] Dostupné z https://gist.github.com/benmarwick/00772ccea2dd0b0f1745

[4] Dostupné z https://ggplot2.tidyverse.org



Obrázek 2.12: Bagploty pro jednotlivé věkové kategorie, bez odlehlých hodnot. *Bag* je vždy vykreslen tmavší barvou, *loop* světlejší a medián je označen symbolem "+".

2.4 Odlehlost a její využití

Velkou výhodou bagplotu je snadná detekce odlehlých hodnot bez předpokladu normality. Je však náročný na výpočet a chceme-li pouze odhalit neobvyklá pozorování, existují jednodušší způsoby. Jedna z nich je založena na tzv. odlehlosti (z ang. outlyingness), což lze chápat jako opak hloubky. Platí tedy, že čím odlehlejší je hodnota, tím větší bude mít hodnotu odlehlosti. Následující definice a pozorování jsou převzaty z [2].

Definice 8. Nechť $X = \{x_1, x_2, \ldots, x_n\}$ je datový soubor, **odlehlost** bodu $z \in \mathbb{R}$ vzhledem k výběru X je definována jako

$$r_1(z;X) = \frac{|\, z - \bar{x}_{0.5} \,|}{MAD(X)}, \tag{2.4.1}$$

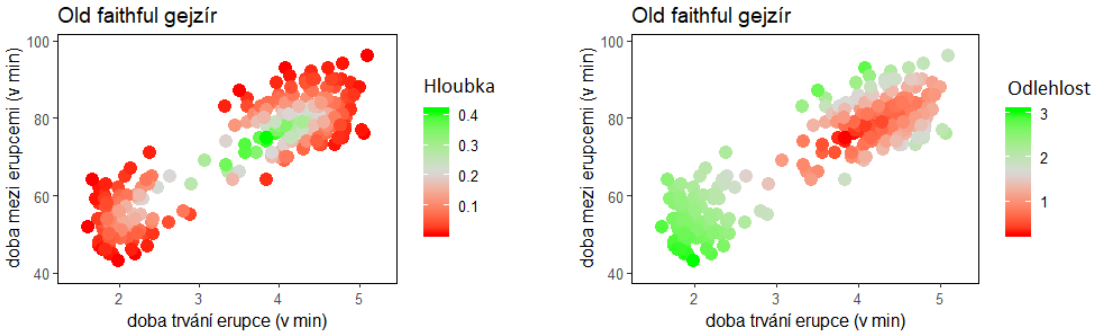
kde $\bar{x}_{0.5}$ značí výběrový medián, a *MAD* (z ang. median absolute deviation) značí mediánovou absolutní odchylku, tedy medián z $\{|\, x_i - \bar{x}_{0.5} \,|\}_{i=1}^{n}$.

MAD je robustní mírou rozptýlenosti a na rozdíl od rozptylu nebo směrodatné odchylky je méně ovlivněna extrémními hodnotami a nevyžaduje normalitu dat.

Analogicky jako v případě vícerozměrné hloubky můžeme definovat odlehlost obecně pro p-rozměrný prostor.

Definice 9. Nechť $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, kde $\mathbf{x}_i = \{x_{i1}, \ldots, x_{ip}\}^T \in \mathbb{R}^p$ pro $i = 1, \ldots, n$, je datový soubor, $\mathbf{u} \in \mathbb{R}^p$, $\|\mathbf{u}\| = 1$. Odlehlost bodu $\mathbf{z} \in \mathbb{R}^p$ vzhledem k výběru X definujeme jako

$$r_p(\mathbf{z};X) = \max_{\|\mathbf{u}\|=1} r_1(\mathbf{u}^T\mathbf{z}; \{\mathbf{u}^T\mathbf{x}_i\}_{i=1}^{n}). \tag{2.4.2}$$



(a) (b)

Obrázek 2.13: Porovnání hloubky s odlehlostí. Na obou grafech jsou vykresleny doby mezi erupcemi v závislosti na době trvání erupcí gejzíru Old Faithful v americkém národním parku Yellowstone. Data pochází z knihovny `datasets` v R. V souboru bylo celkem 299 pozorování zaznamenaných od 1. do 15. srpna roku 1985. Hloubka (v tomto případě navíc dělena rozsahem výběru n) a odlehlost jednotlivých bodů byly vypočteny za pomocí funkcí `depth` a `outlyingness` z balíčku `mrfDepth`.

Vlastnosti odlehlosti jsou analogické vlastnostem hloubky:

1. Odlehlost je invariantní vůči afinním zobrazením:

$$r_p(\mathbf{A}\mathbf{x}+\mathbf{b};\{\mathbf{A}\mathbf{X}_i+\mathbf{b}\}) = r_p(\mathbf{x};\mathbf{X})$$

   pro všechna $\mathbf{b} \in \mathbb{R}^p$ a každé nesingulární lineární zobrazení $\mathbf{A}$.

2. Kontury odlehlosti $O_r = \{\mathbf{x} \in \mathbb{R}^p \mid r_p(\mathbf{x};\mathbf{X}) \geq r\}$ jsou konvexní a vnořené: $O_{r+h} \subset O_r$ pro $h > 0$.

3. Pro symetrická data se minimální odlehlost blíží k 0. Platí, že čím je větší n, tím je vyšší pravděpodobnost konvergence.

4. Pro data z eliptického rozdělení kontury odlehlosti konvergují k eliptickému tvaru dat.

Nevýhodou bagplotu je doba trvání jeho výpočtu, týká se to hlavně mediánu, jehož časová složitost činí $O(n^2(\log n)^2)$. Autoři článku [12] navrhli, aby se při větších souborech vypočítal medián (a následně i *bag*) pouze z menšího, náhodně vybraného vzorku původních dat (se 150 pozorováními). Výpočet ostatních částí grafu se pak provede na celém datovém souboru. Takto lze sestrojit bagplot pro výběr, jehož rozsah se pohybuje v rozmezí několika tisíců bodů.

Bagplot lze také vykreslit za použití odlehlosti a výrazně se tím zkrátí doba jeho výpočtu, tento postup navrhli M. Hubert a S. van der Veeken v článku [4]. Místo mediánu se bere bod s nejnižší odlehlostí, *bag* tvoří 50 % pozorování s nejnižší odlehlostí. Takto sestrojený bagplot ovšem není tak robustní jako bagplot založený na poloprostorové hloubce a je tedy citlivější na větší množství odlehlých pozorování.

Závěr

V této práci jsme si představili dvě metody průzkumové analýzy dat sloužící k rychlému posouzení charakteristik dat jako je distribuce nebo korelace mezi dvěma proměnnými, a také k odhalení odlehlých hodnot, které jsou nutné před samotnou analýzou vyšetřit. V první kapitole jsme si popsali boxplot, který je známým a často používaným nástrojem vizualizace jednorozměrných dat. Jeho výhodou je především jednoduchost a kompaktnost. Využívá se hlavně při porovnávání více skupin dat.

V dnešní době jsou však data často mnohorozměrného charakteru a boxplot je schopen vykreslit pouze vztah mezi kategoriálním a kvantitativním proměnným, ne však mezi dvěma kvantitativními proměnnými. Dvourozměrným zobecněním boxplotu je tzv. bagplot, kterým jsme se zabývali ve druhé kapitole. Nejprve jsme si popsali koncept poloprostorové hloubky zobecňující kvantily, na kterých byl založen boxplot. Popsali jsme si konstrukci bagplotu a ukázali jsme, jak se dá graf sestrojit v R. Na závěr jsme si uvedli několik příkladů, na kterých jsme ilustrovali výhody použití bagplotu.

Mnohé metody detekce odlehlých hodnot vyžadují, aby data byla z normálního rozdělení. Jednou z hlavních předností bagplotu, a také boxplotu, je ten, že nepředpokládá normalitu v datech. Tyto diagnostické grafy lze tedy vykreslit aniž bychom museli napřed něco ověřit a navíc nám poskytují užitečné informace ohledně zkoumaných dat. V případě bagplotu nám navíc část bag dává představu o tom, jaký tvar mají naše data. Použitím statistického softwaru je konstrukce otázka několika minut. Je tedy vhodné si je vykreslit před dalším zkoumáním dat. Cílem práce nebylo pouze grafy srozumitelně popsat, ale také poskytnout stručný návod, jak vykreslené grafy správně interpretovat.

Seznam použité literatury

[1] DONOHO, D. L. a P. J. HUBER. *The notion of breakdown point*. V "A Festschrift for Erich L. Lehmann". Wadsworth, Belmont, CA, 1983, 157–184.

[2] DONOHO, D. L. a M. GASKO. *Multivariate Generalizations of the Median and Trimmed Mean, I*. Technical Report 133, Dept. Statistics, Univ. California, Berkeley, 1987.

[3] DONOHO, D. L. a M. GASKO. *Breakdown Properties of Location Estimates Based on Halfspace Depth and Projected Outlyingness*. The Annals of Statistics. 1992, **20**(4), 1803–1827. DOI: 10.1214/aos/1176348890. ISSN 0090-5364. Dostupné také z: http://projecteuclid.org/euclid.aos/1176348890

[4] HUBERT, M. a S. VAN DER VEEKEN. *Outlier detection for skewed data*. Journal of Chemometrics. 2008, **22**(3), 235–246. DOI: 10.1002/cem.1123. ISSN 08869383. Dostupné také z: http://doi.wiley.com/10.1002/cem.1123

[5] KONTTO, J. P. *Visualizing large epidemiological data sets using depth and density*. 2007. Diplomová práce. University of Helsinki, Faculty of Social Sciences, Statistics.

[6] LIU, R. Y. *On a notion of data depth based on random simplices.* Annals of Statistics. 1990, **18**(1), 405–414. DOI: 10.1214/aos/1176347507.

[7] MILLER, Kim, S. RAMASWAMI, P. ROUSSEEUW, J. A. SELLARES, I. STREINU a A. STRUYF. *Efficient computation of location depth contours by methods of computational geometry.* Statistics and Computing. 2003, **13**(2), 153-162. DOI: 10.1023/A:1023208625954. ISSN 09603174. Dostupné také z: http://link.springer.com/10.1023/A:1023208625954

[8] OJA, H. *Descriptive Statistics for Multivariate Distributions*. Statistics & Probability Letters. 1983, 327–332.

[9] ROUSSEEUW, P. J. a I. RUTS. *Algorithm AS 307: Bivariate Location Depth*. Applied Statistics. 1996, **45**(4). DOI: 10.2307/2986073. ISSN 00359254. Dostupné také z: https://www.jstor.org/stable/10.2307/2986073?origin=crossref

[10] ROUSSEEUW, P. J. a I. RUTS. *The Bagplot: A Bivariate Box-and-Whiskers Plot*. Technical report. Universitaire Instelling Antwerpen, Belgium, 1997.

[11] ROUSSEEUW, P. J. a I. RUTS. *Constructing the bivariate Tukey median*. Statistica Sinica. 1998, **8**(3), 827–839.

[12] ROUSSEEUW, P. J., I. RUTS a J. W. TUKEY. *The Bagplot: A Bivariate Boxplot*. The American Statistician. 1999, **53**(4), 382–387.

[13] RUTS, I. a P. ROUSSEEUW. *Computing Depth Contours of Bivariate Point Clouds*. Computational Statistics & Data Analysis. 1996, **23**, 153–168. DOI: 10.1016/S0167-9473(96)00027-8.

[14] SERFLING, R. *A Depth Function and a Scale Curve Based on Spatial Quantiles*. In: Dodge Y. (eds) Statistical Data Analysis Based on the L1-Norm and Related Methods. Statistics for Industry and Technology. Birkhäuser, Basel, 2002, , 25–38. DOI: https://doi.org/10.1007/978-3-0348-8201-9_3.

[15] TUKEY, J. W. *Mathematics and the Picturing of Data*. Proceedings of the International Congress of Mathematicians. 1975, **2**, 523–531.

[16] TUKEY, J. W. *Exploratory Data Analysis*. Addison-Wesley, 1977.

[17] TUKEY, J. W., R. MCGILL a W. A LARSEN. *Variations of Box Plots*. The American Statistician. 1987,**32**(1), 12–16. DOI: 10.2307/2683468.

[18] ZUO, Y. a R. SERFLING. *General notions of statistical depth function*. The Annals of Statistics. 2000, **28**(2), 461–482.