УНИВЕРЗИТЕТ У БЕОГРАДУ
МАТЕМАТИЧКИ ФАКУЛТЕТ

Катарина Јеремић

Статистичке функције дубине и њихова примена у откривању аутлајера у дводимензионом простору

— мастер рад —

Београд, 2017.

Садржај

1 Статистичке функције дубине **1**
1.1 Примери статистичких функција дубине 2

2 Аутлајери **11**
2.1 Једнодимензионе статистичке методе 12
2.2 Проблем вишедимензионих аутлајера 13

3 Рачунске методе за детекцију аутлајера **15**
3.1 Методе 16
3.1.1 Филцмозеров метод 17
3.1.2 Метод Русова и Ван Зомерена 17
3.1.3 Метод Бекерове и Гадерове 17
3.2 Робусне оцене 18
3.2.1 MCD оцене 19

4 Графичке методе за детекцију аутлајера **24**
4.1 Балон дијаграм 24
4.2 Врећасти дијаграм 25

5 Откривање аутлајера на подацима **30**
5.1 Симулирани подаци 30
5.1.1 Нормална расподела 31
5.1.2 Студентова расподела 37
5.2 Реални подаци 40
5.2.1 Подаци о вину 40
5.2.2 Подаци о аутомобилима 43

Закључак **44**

Литература **45**

Поглавље 1

Статистичке функције дубине

Основни појам у непараметарској детекцији вишедимензионих аутлајера су такозване функције дубине које се могу дефинисати на више начина. Представићемо различите функције дубине у зависности од њених аналитичких карактеристика. Сам појам статистичке функције дубине је први пут коришћен у Тјукијевом [1] раду као алат за визуелно представљање дводимензионих података. Касније је овај појам проширен на вишедимензиони случај. Најпростије речено, функција дубине неке тачке представља меру дубине те тачке у облаку свих података. Већина функција дубине је робусна и афино инваријантна што их чини применљивим у вишедимензионим анализама посебно када подаци садрже аутлајере.

Неке од примена статистичких функција дубине су:

- Одређивање вишедимензионе медијане као и квантила.
- Детекција аутлајера.
- Кластеровање података.
- Вишедимензионе оцене густине.

Уведимо прво појам једнодимензионе статистичке дубине. Ако посматрамо тачку x једнодимензионе случајне величине X са расподелом F, њој се може прићи са леве или са десне стране. Дефинишимо вероватноће p и q, као вероватноће приласка са леве односно десне стране, са $p = P\{X \leq x\} = F(x)$ и $q = P\{X \geq x\} = 1 - F(x-)$.

[1] Џон Тјуки (1915-2000) амерички статистичар и математичар

Дефиниција 1.0.1 *Нека је X једнодимензиона случајна величина са расподелом F. Статистичка дубина d_F тачке x у односу на расподелу F дефинише се као*

$$d_F(x) = \min\{F(x), 1 - F(x-)\},$$

где је $F(x-) = \lim_{t \uparrow x} F(t)$.

Статистичке функције дубине служе и за рангирање елемената узорка, па је медијана тачка са максималном вредношћу дубине. Из дефиниције 1.0.1 се може закључити да је дубина једнодимензионе медијане једнака $\frac{1}{2}$.

За разлику од једнодимензионог, у вишедимензионом простору не постоји природно уређење. Због тога се статистичке функције дубине користе да би се извршило рангирање у таквим просторима. У научној литератури постоје различите врсте статистичких функција дубине и та област је врло популарна. С обзиром да је за откривање аутлајера неопходно уредити податке, коришћење статистичких функција дубине се наметнуло као добро решење.

У вишедимензионом случају ствар је много компликованија. Тачки x се може прићи на бесконачно много начина. Из тог разлога постоји велики број различитих дефиниција вишедимензионих статистичких функција дубине. Неке од њих су базиране на симплексима[2], неке на конвексним омотачима или пак затвореним полупросторима. У наставку ће бити више речи о свакој од њих.

1.1 Примери статистичких функција дубине

Тјуки је 1975. увео појам положајне дубине, познатији као полупросторна дубина [3], као средство за визуелно представљање дводимензионих скупова података. У дводимензионом случају, полупросторна дубина тачке x из скупа података S_n се дефинише као најмањи број тачака које се налазе са једне од страна праве која пролази кроз x. Ова дефиниција се може проширити на вишедимензиони случај.

[2] генерализовани појам троугла или тетраедра у више димензија; енгл- simplex
[3] енгл- halfspace depth

Дефиниција 1.1.1 (Полупросторна функција дубине) *Полупросторна дубина тачке* $x = (x_1, ..., x_p) \in S_n = \{x_i = (x_{i1}, ..., x_{ip}); i = 1, ..., n\} \subset \mathbb{R}^p$ *из* p*-димензионог скупа података* S_n*, се дефинише као најмањи број тачака у затвореном полупростору који садржи тачку* x *у својој граничној равни.*

Уопштење дефиниције полупросторне дубине:

За дату расподелу вероватноће P у p-димензионом реалном простору $\mathbb{R}^p$, Тјукијева функција полупросторне (положајне) дубине $HD(x; P)$ обезбеђује поредак тачака (прецизније, Борелових скупова) $x \in \mathbb{R}^p$ од центра ка крајевима. Тиме се обезбеђује рангирање тачака базирано на расподели вероватноће P сходно следећој дефиницији

$$HD(x; P) = \inf\{P(H) : x \in \text{затвореном полупростору} H\} \tag{1}$$

Поред Тјукијеве функције дубине, постоје и друге функције дубине. Уведимо појам Махаланобисове дубине која је једна од најчешће коришћених функција дубине приликом детекције аутлајера. Постоји велики број метода за откривање аутлајера који су базирани на овој функцији дубине. У посебном одељку овог рада поменуте методе ће бити детаљније објашњене.

Дефиниција 1.1.2 (Махаланобисова дубина) *Махаланобисова дубина тачке* $x \in S_n \subset \mathbb{R}^p$ *из* p*-димензионог скупа података* S_n *се дефинише као:*

$$MD(x; S_n) = [1 + (x - \bar{x})^T S^{-1} (x - \bar{x})]^{-1} \tag{2}$$

где су $\bar{x}$ *и* S *вектор средње вредности и коваријациона матрица од* S_n*.*

Ова функција није робусна, јер је сачињена од неробусних мера као што су узорачка средња вредност и узорачка коваријациона матрица. Друга мана ове функције је то што зависи од постојања другог момента.

Уведимо затим статистичку функцију дубине преко конвексних слојева. Ова функција се најчешће користи за визуелну детекцију аутлајера преко балон дијаграма о којем ће бити нешто више речи у посебном поглављу.

Дефиниција 1.1.3 (Дубине преко конвексних слојева) *Дубина преко конвексних слојева тачке* $x \in S_n \subset \mathbb{R}^p$ *из* p*-димензионог скупа података* S_n *је једноставно ниво конвексног слоја којем* x *припада.*

Конвексни слој (омотач) скупа тачака X је математички појам који означава најмањи конвексни скуп који садржи X. Конвексни слојеви из претходне дефиниције се дефинишу на следећи начин. Конструишемо најмањи конвексни слој који обухвата све тачке из датог сета података. Тачке на ободу представљају први конвексни слој и одбацују се. Конвексни слој од осталих тачака се опет конструише и затим се поступак понавља одбацивањем ободних тачака за сваки од угњеждених слојева. Последњи слој који се формира представља најдубљу тачку која је медијана датих података. У дводимензионом случају конвексни слојеви су конвексни полигони, самим тим дубина сваке тачке се може графички представити.

Дефиниција 1.1.4 (Симплексна дубина) *Симплексна дубина тачка $x \in S_n \subset \mathbb{R}^p$ из p-димензионог скупа података S_n се дефинише као број затворених симплекса који садрже x и имају $p+1$ темена у S_n.*

Прецизније, симплексна дубина тачке x се може дефинисати као вероватноћа да тачка x припада случајном симплексу из $\mathbb{R}^p$

$$SD(x;P) = P\{x \in S[X_1, ..., X_{d+1}]\}, \tag{3}$$

где је $X_1, ..., X_{d+1}$ случајан узорак из расподеле P, а $S[x_1, ..., x_{d+1}]$ d-димензиони симплекс са теменима $x_1, ..., x_{d+1}$. У дводимензионом случају, симплексна дубина тачка x је број троуглова са теменима из скупа S_n у којима је садржана тачка x.

У раду [16] Роберта Серфлинга и Јиун Зуа из 2000. године су поред класификације статистичких функција дубине дате и жељене аналитичке особине које би свака статистичка функција дубине $SD(x;P)$ требало да испуњава:

- Афина инваријантност. $SD(x;P)$ је независна од координатног система.
- Максимална вредност у центру. Ако расподела вероватноће P има јединствено одређен "центар" (односно, тачку симетрије у односу на неки појам симетрије), тада $SD(x;P)$ има максималну вредност у тој тачки.
- Симетрија. Ако је расподела вероватноће P симетрична око тачке x у складу са неким појмом симетрије, тада је и функција $SD(x;P)$ симетрична.

- Монотоно опадање приликом удаљавања од најдубље тачке. Вредност функције дубине $SD(x;P)$ опада при удаљавању од тачке са највећом вредношћу.
- Тежња нули у бесконачности. $SD(x;P) \to 0, \|x\| \to \infty$.
- Функција $SD(x;P)$ је непрекидна по x са горње стране.
- $SD(x;P)$ је као функција расподеле P.
- Квази-конкавност $SD(x;P)$ као функције променљиве x. Скуп $\{x : SD(x;P) \geq c\}$ је конвексан за сваку реалну вредност c.

У истом раду Зуо и Серфлинг су увели поделу статистичких функција дубине у неколико категорија. Претходно дефинисане функције дубине представљају специјалан случај неког од следећих типова.

Функције типа А : Нека је $h(x; x_1, ..., x_r)$ било која ограничена и ненегативна функција која се може окарактерисати као мера блискости тачке x од тачака $x_1, ..., x_r$. Одговарајућа функција дубине типа А је

$$D(x;P) = E(h(x; X_1, ..., X_r)),$$

где је $X_1, ..., X_r$ случајан узорак из расподеле P.

Узимајући за $r = d+1$ и $h(x; x_1, ..., x_{d+1}) = \mathrm{I}\{x \in S[x_1, ..., x_{d+1}]\}$, добијамо симплексну дубину (3).

Функције типа Б : Нека је $h(x; x_1, ..., x_r)$ било која неограничена и ненегативна функција која представља растојање тачке x од тачака $x_1, ..., x_r$. Одговарајућа функција дубине типа Б је

$$D(x;P) = (1 + E(h(x; X_1, ..., X_r)))^{-1},$$

где је $X_1, ..., X_r$ случајан узорак из расподеле P.

Пример 1 (L^p дубина) *Други начин за мерење дубине је коришћењем L^p норме $\|\cdot\|_p$. Узимајући за $h(x; x_1) = \|x - x_1\|_p$ добијамо одговарајућу функцију дубине типа Б*

$$L^p D(x;F) = (1 + E\|x - X\|_p)^{-1}$$

Функције типа Ц: Нека је $O(x;F)$ мера удаљености тачке x у $\mathbb{R}^d$ у односу на центар тј. најдубљу тачку расподеле F. Функција $O(x;F)$

је обично неограничена, али је зато одговарајућа ограничена функција дубине дефинисана као

$$D(x;F) = (1 + O(x;F))^{-1}$$

Овако дефинисану функцију називамо функцијом дубине типа Ц.

Махаланобис је 1936. увео растојање између две тачке x и y у $\mathbb{R}^d$ са позитивно дефинитном матрицом M димензија $d \times d$ као

$$d_M^2(x,y) = (x-y)^T M^{-1}(x-y)$$

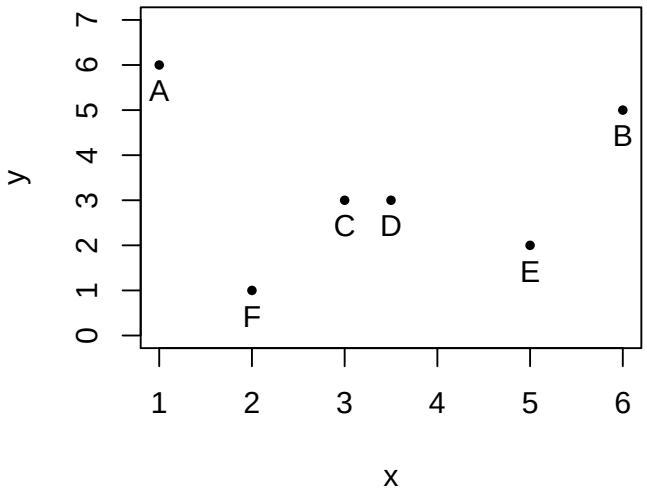
На основу Махаланобисовог растојања можемо дефинисати Махаланобисову дубину типа Ц као

$$MHD(x;F) = (1 + d^2_{\Sigma(F)}(x, \mu(F)))^{-1},$$

где је F расподела вероватноће, а $\mu(F)$ и $\Sigma(F)$ су респективно узорачка средња вредност и узорачка коваријациона матрица. Овако дефинисана функција дубине није робусна због $\mu(F)$.

Пример 2 *Посматрајмо скуп тачака* $A(1,6)$, $B(6,5)$, $C(3,3)$, $D(3.5,3)$, $E(5,2)$ *и* $F(2,1)$. *Покушајмо да одредимо дубину сваке од тачака користећи различите функције дубине као што су симплексна, полупросторна, Махаланобисова и дубина преко конвексних слојева.*

За одређивање симплексне дубине потребно је формирати све могуће троуглове од претходних 6 тачака. Број могућих троуглова је 20 и у следећој табели можемо видети које тачке су садржане у сваком од троуглова.



Слика 1.1: Скуп тачака

	Троугао	*Садржане тачке*
1	$\triangle ABC$	A, B, C
2	$\triangle ABD$	A, B, D
3	$\triangle ABE$	A, B, E
4	$\triangle ABF$	A, B, C, D, F
5	$\triangle ACD$	A, C, D
6	$\triangle ACE$	A, C, D, E
7	$\triangle ACF$	A, C, F
8	$\triangle ADE$	A, D, E
9	$\triangle ADF$	A, C, D, F
10	$\triangle AEF$	A, C, D, E, F
11	$\triangle BCD$	B, C, D
12	$\triangle BCE$	B, C, D, E
13	$\triangle BCF$	B, C, D, F
14	$\triangle BDE$	B, D, E
15	$\triangle BDF$	B, D, F
16	$\triangle BEF$	B, E, F
17	$\triangle CDE$	C, D, E
18	$\triangle CDF$	C, D, F
19	$\triangle CEF$	C, E, F
20	$\triangle DEF$	D, E, F

Из претходне таблице можемо одредити да су дубине тачака A, B, C, D, E, F *једнаке 10, 10, 13, 15, 10, 10 респективно. Ако погледамо полупросторну дубину, потребно је само одредити праве кроз сваку од тачака тако да број тачака са једне од страна буде минималан. Тиме добијамо да су дубине тачака* A, B, C, D, E, F *једнаке 0, 0, 1, 2, 0, 0 респективно. Што се тиче дубине преко конвексних слојева, тачке* A, B, E, F *чине први конвексни пологон па је њихова дубина једнака 1, док је дубине преосталих тачака 2. Одредимо затим Махаланобисову дубину за сваку од тачака. Потребно је прво израчунати узорачку средину и узорачку коваријациону матрицу датог скупа тачака. Једноставним одређиваньем аритметичке средине за сваку од колона* X, Y *добијамо да је* $\mu = \begin{bmatrix} 3.41 \\ 3.33 \end{bmatrix}$. *Узорачку коваријациону матрицу добијамо користећи формулу*

$$\Sigma = \frac{1}{N-1} X^{*T} X^*, \text{ где је } X^* = \begin{bmatrix} X_1 - \bar{X} & Y_1 - \bar{Y} \\ X_2 - \bar{X} & Y_2 - \bar{Y} \\ \vdots & \vdots \\ X_6 - \bar{X} & Y_6 - \bar{Y} \end{bmatrix}$$

Након израчунавања матрице X^* *можемо применити формулу (2) за рачунање Махаланобисове дубине чиме добијамо да су дубине тачака* A, B, C, D, E, F *једнаке 0.22, 0.26, 0.93, 0.97, 0.46, 0.31 респективно.*

	A	B	C	D	E	F
Симплекс дубина	*10*	*10*	*13*	*15*	*10*	*10*
Полупросторна дубина	*0*	*0*	*1*	*2*	*0*	*0*
Дубина преко конвексних слојева	*1*	*1*	*2*	*2*	*1*	*1*
Махаланобисова дубина	*0.22*	*0.26*	*0.93*	*0.97*	*0.46*	*0.31*

Из претходне табеле можемо закључити да методама симплекс, полупросторне и Махаланобисове дубине добијамо да је медијана података тачка D. *У случају дубине преко конвексних слојева нема јединствене медијане, али се може сматрати да је она негде између тачака* C *и* D.

Проблем проналажења и израчунавања вишедимензионе медијане, или најдубље (централне) вредности, није толико очигледан као у једнодимензионом случају. Медијана се у теорији вероватноће и статистици описује као број који раздваја горњу половину узорка, популације

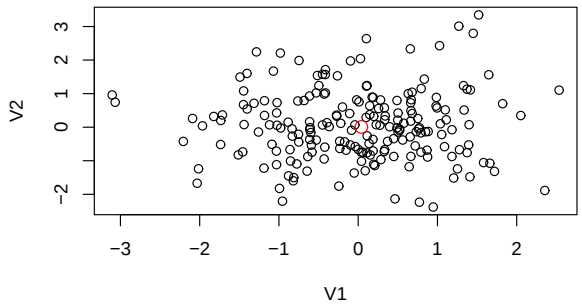
или расподеле вероватноће од доње половине. Медијана коначног низа бројева се може наћи тако што се бројеви поређају по величини, и узме се средњи члан низа. Уколико постоји паран број чланова низа, медијана није јединствена, па се најчешће узима аритметичка средина две вредности које су кандидати за медијану.

Истраживања на тему вишедимензионе медијане су почела још у двадесетом веку. Од тада је било неколико покушаја да се одреди природна дефиниција вишедимензионе медијане. Постоји велики број дефиниција вишедимензионе медијане у зависности која статистичка функција дубине се користи. И у непрекидном и у дискретном случају, вишедимензиона медијана се дефинише као најдубља локација, или прецизније, као тачка θ (или скуп тачака) у којој статистичка функција дубине има максималну вредност.

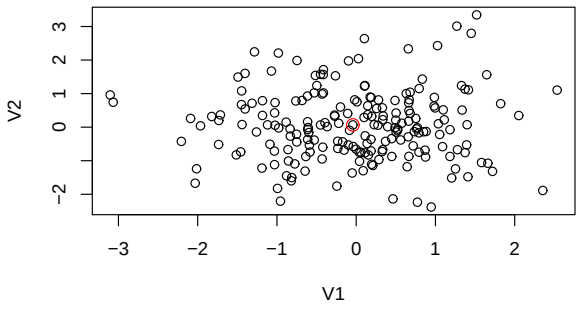
У следећем примеру ћемо показати да за исти скуп података различите статистичке функције дубине дају различиту вредност медијане.

Пример 3 *Посматрајмо дводимезиони скуп података који има 200 елемената узорка из нормалне $N(0, E)$ расподеле. Користећи полупросторну и Махаланобисову функцију дубине можемо одредити дводимензиону медијану. Као што се може закључити са следеће две слике, ове две функције дубине нису показале исту вредност медијане. Полупросторна медијана се налази негде између централних елемената узорка, док је Махаланобисова медијана тачка из датог скупа података.*

Вредност полупросторне медијане је $(0.04038, 0.00708)$*, док Махаланобисова медијана има вредност* $(-0.04178, 0.06788)$*. На сликама 1.2 и 1.3 црвеним кружићем је обележена медијана посматраних података.*



Слика 1.2: Полупросторна медијана



Слика 1.3: Махаланобисова медијана

Поглавље 2

Аутлајери

Аутлајери представљају вредности обележја које неуобичајено много одступају од модела и осталих вредности. Постоји много неформалних дефиниција аутлајера. Једна од тих је да су то елементи узорка који представљају могућа абнормална понашања података. Веома често аутлајери могу бити производ људске грешке при неком статистичком истраживању. Детекција и отклањање аутлајера је један од кључних корака у процесу статистичке анализе. Неки од метода детекције се заснивају на мерама удаљености, кластеровању или пак визуелном представљању података.

Методе за детекцију аутлајера имају велику примену у откривању превара са кредитним картицама, клиничким истраживањима, анализи нерегуларности гласања, чишћењу података, упадима на мреже, предвиђању времена, географским информационим системима и у другим задацима везаним за истраживање података[1].Методе за детекцију аутлајера се могу поделити на једнодимензионе и вишедимензионе методе. Поред ове поделе постоји подела на параметарске и непараметарске методе.

Параметарске методе се користе када је позната основна расподела обележја. Ове методе означавају као аутлајере оне елементе узорка који одступају од претпоставки модела. Параметарске методе су често неодговарајуће за вишедимензионе скупове података без претходног знања о стварној расподели података.

Унутар класе непараметарских метода за детекцију аутлајера постоји одвојен скуп метода заснованих на растојању. Ове методе су засноване на локалним мерама удаљености које представљају уопштење статистичких функција дубине о којима је било речи у претходном поглављу.

[1]енгл- data-mining

Друга класа метода за детекцију аутлајера је заснована на кластер техникама, где кластер мале величине може бити сматран групом аутлајера.

2.1 Једнодимензионе статистичке методе

Већина једнодимензионих метода за детекцију аутлајера се ослања на претпоставку да је позната расподела података. Многи тестови неслагања за откривање једнодимензионих аутлајера даље претпостављају да су познати параметри расподеле, мада је та претпоставка често нарушена у истраживању реалних података.

Централна претпоставка у статистичким методама за откривање аутлајера је постојање модела који дозвољава малом броју елемената узорка да буду из расподела $G_1, G_2, ..., G_k$, које се разликују од циљне расподеле F. За расподелу података се често претпоставља да је нормална расподела $N(\mu, \sigma^2)$. Проблем идентификације аутлајера је тада пренет на проблем идентификовања оних елемената узорка који леже у тзв. аутлајерском региону.

За сваки интервал поверења α, $0 < \alpha < 1$, α-аутлајер регион нормалне $N(\mu, \sigma^2)$ расподеле је дефинисан са:

$$\text{out}(\alpha, \mu, \sigma^2) = \{x : |x - \mu| > z_{(1-\alpha/2)}\sigma\}$$

где је z_q квантил $N(0, 1)$ расподеле. Број x је α-аутлајер у односу на расподелу F ако $x \in out(\alpha, \mu, \sigma^2)$. Иако је традиционално нормална расподела коришћена као цильна расподела, ова дефиниција може бити проширена на сваку унимодалну симетричну расподелу са позитивном функцијом густине, укључујући и вишедимензиони случај.

Пример 4 (Пирсонов критеријум за одбацивање аутлајера) *Опишимо поступак одбацивања аутлајера коришћењем Пирсоновог критеријума.*

- *Прво израчунамо средњу вредност $\bar{x}_n$ и стандардно одступање узорка $\bar{s}_n$.*
- *Одредимо R (које зависи од обима узорка) из Пирсонове таблице и претпоставимо број сумњивих елемената узорка (прва претпоставка је да имамо једно сумњиво опазање).*
- *Израчунамо максимално дозвољено одступање $D_{max} = \bar{s}_n * R$*

- *Ако је потенцијални аутлајер x_i, рачунамо $|x_i - \bar{x}_n|$*
- *Елиминишемо x_i ако је $|x_i - \bar{x}_n| > D_{max}$*
- *Уколико у претходном кораку елиминишемо један елемент узорка, претпостављамо да постоје 2 аутлајера, задржавајући исте вредности за $\bar{x}_n$ и $\bar{s}_n$ и исти обим узорка. Исти поступак се наставља за већи број претпостављених аутлајера, задржавајући исте вредности за $\bar{x}_n$, $\bar{s}_n$ и n.*
- *Поступак се понавља док се не елиминишу сви аутлајери.*

Примера ради, за обим узорак $n = 10$ вредности за R из Пирсонове таблице за 1, 2, 3, 4 и 5 претпостављених аутлајера су 1.878, 1.570, 1.380, 1.237, 1.114.

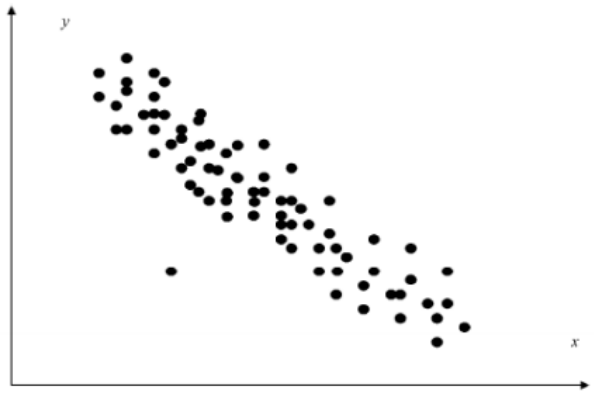
2.2 Проблем вишедимензионих аутлајера

У великом броју случајева, вишедимензиони елементи узорка не могу бити детектовани као аутлајери када се свака променљива посматра независно. Откривање аутлајера је могуће само када је вишедимензиона анализа изведена и интеракције између различитих променљивих су поређене унутар класе података. Једноставан пример је приказан на слици 2.1 који представља податке који имају две мере на дводимензионом простору. Доња лево тачка је вишедимензиони аутлајер, али није једнодимензиони. Када разматрамо сваку меру посебно са респектом на ширење вредности дуж x и y осе, можемо видети да оне падају близу центра једнодимензионе расподеле. Тест за аутлајере мора узети у обзир везе између две променљиве које у овом случају делују необично.

У анализи вишедимензионих аутлајера чести су ефекти маскирања или преплављивања. Примери ефекта маскирања су веома чести у реалним подацима, па се из тог разлога приликом детекције аутлајера користе робусне оцене параметара, о којима ће бити речи у наредном поглављу.

Ефекат маскирања:

Ефекат маскирања се дефинише у случају када један аутлајер маскира други аутлајер, па тај аутлајер не може бити детектован у присуству првог. Стога, после брисања маскирајућег аутлајера прави аутлајер је видљив као такав.



Слика 2.1: Дводимензиони дијаграм

Ефекат преплављивања:

Један аутлајер преплављује други елемент узорка ако други може бити сматран аутлајером само у присуству првог. Другим речима, после брисања првог аутлајера други елемент узорка постаје регуларно опажање. Преплављивање се дешава када група удаљених елемената узорка искошава оцене средине и коваријационе матрице ка себи и далеко од других регуларних елемената узорка.

Поглавље 3

Рачунске методе за детекцију аутлајера

У рачунским методама за детекцију аутлајера најчешће се користи Махаланобисова статистичка функција дубине. Стандардни методи за вишедимензионо откривање аутлајера који се заснивају на Махаланобисовом растојању, које се дефинише једнакошћу

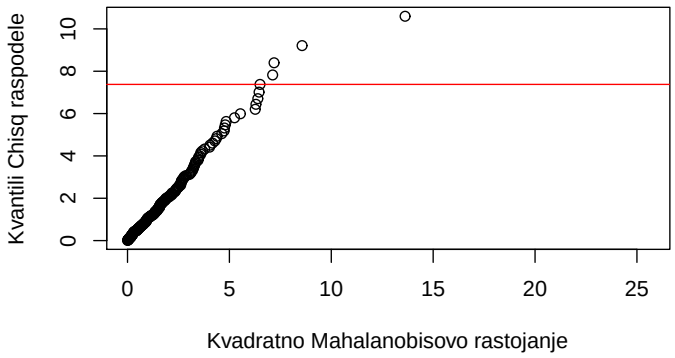
$$MD_i = [(x_i - t)^T C^{-1}(x_i - t)]^{1/2}$$

за p димензиони елемент узорка x_i где је $i = 1, ..., n$, а t и C су респективно узорачка средња вредност и узорачка коваријациона матрица или пак робусне оцене истих. Махаланобисово растојање тачке x_i представља меру удаљености те тачке од центра свих података (медијане). За скуп података који има нормалну расподелу, Махаланобисово растојање има приближно хи-квадрат расподелу са p степени слободе (χ^2_p). Потенцијални аутлајери x_i ће имати велике вредности Махаланобисовог растојања MD_i.

Гарет [2] је увео појам хи-квадрат графика који представља емпиријску функцију расподеле квадрата Махаланобисових растојања у односу на χ^2_p расподелу. Прекид у самом репу расподеле представља идентификатор за аутлајере, јер вредности изнад тог прекида се итеративно бришу док се не добије права линија.

На слици 3.1 су приказани симулирани подаци из нормалне расподеле са 5% контаминираних вредности. Поређењем квантила χ^2_p расподеле са квадратним Махаланобисовим растојањем може се видети прекид у самом репу расподеле што указује на постојање аутлајера. Црвена линија представља квантил $\chi^2_{(2,0.975)}$.

Русов и Ван Зомерен [11] су увели граничну вредност за одређивање да ли је неки елемент узорка аутлајер или не. Ова вредност је



Слика 3.1: Хи-квадрат график

углавном неки квантил χ^2_p расподеле (нпр. 97,5% квантил). Што се тиче самих оцена за узорачку средину и коваријациону матрицу Русов користи MVE [1] оцену. Након неколико година MVE оцена је замењена оценом MCD [2] из разлога бољих статистичких особина и много бржег алгоритма за израчунавање. У посебном одељку овог поглавља биће детаљно објашњена MCD оцена.

Употреба робусних оцена параметара вишедимензионе расподеле често може побољшати перформансе процедуре за детекцију аутлајера. Хади [1] адресира овај проблем и предлаже да се узорачка средња вредност замени медијаном и да се израчуна коваријациона матрица за подскуп оних елемената узорка који имају најмање Махаланобисово растојање. Каусинус и Руиз [6] предлажу робусну оцену за коваријациону матрицу која је заснована на тежинским опажањима према њиховој удаљености од центра.

3.1 Методе

Постоји велики број метода за откривање вишедимензионих аутлајера који користе разне статистичке функције дубине. Наведимо неке од метода које се заснивају на Махаланобисовом растојању.

[1] енгл- minimum volume ellipsoid
[2] енгл- minimum covariance determinant

3.1.1 Филцмозеров метод

Заједничким радом Филзмосера, Герета и Рејмана[9] овај метод је уведен 2005. године. Нека је $G_n(u)$ емпиријска функција расподеле квадрата Махаланобисових растојања MD_i^2 и $G(u)$ функција χ_p^2 расподеле. Ако су подаци из нормалне расподеле тада G_n конвергира ка G. Због тога има смисла поредити репове расподела G_n и G ради откривања потенцијалних аутлајера. Репови χ_p^2 расподеле се могу дефинисати као квантили $\delta = \chi^2_{(p,1-\beta)}$ за малу вредност параметра β (нпр. $\beta = 0,025$) и

$$p_n(\delta) = \sup_{u \geq \delta} (G(u) - G_n(u))^+,$$

где $p_n(\delta)$ мери одступање емпиријске функције расподеле од теоријске функције расподеле на реповима, дефинисаним вредношћу δ. Ако је вредност $p_n(\delta)$ већа од критичне вредности $p_{crit}\ (\delta, n, p)$, онда се она може узети као мера за детекцију аутлајера у скупу података, у супротном мера за детекцију аутлајера је 0.

Критична вредност $p_{crit}\ (\delta, n, p)$ зависи од квантила δ и обима података n и добија се симулацијом на следећи начин. Под претпоставком нормалности података, узорци величине n се симулирају из p-димензионе стандардне нормалне расподеле. Након тога, аутлајер детекција се примењује за сваки од узорака и рачуна се $p_n(\delta)$. Критична вредност се затим дефинише као $(1 - \varepsilon)$ квантил од свих вредности $p_n(\delta)$, за малу вредности ε, нпр. $\varepsilon = 0,05$. Једноставности ради у наставку ћемо овај метод означавати са FGR.

3.1.2 Метод Русова и Ван Зомерена

Русов и Ван Зомерен представљају овај метод у раду [11] 1990. године. Овај метод користи фиксне квантиле $\chi^2_{(p,1-\varphi)}$ као граничне вредности за детекцију аутлајера. Сви елементи узорка чије је квадратно Махланобисово растојање веће од $\chi^2_{(p,1-\varphi)}$ се посматрају као аутлајери. Једноставности ради у наставку ћемо овај метод означавати са RZ.

3.1.3 Метод Бекерове и Гадерове

Овај метод је уведен 1999. године у раду [2]. Појам α аутлајера у односу на вишедимензиону нормалну расподелу $N_p(\mu, \Sigma)$ је елемент скупа

$$\text{out}(\alpha, \mu, \Sigma) := \{x \in \mathbb{R}^p (x - \mu)^T \Sigma^{-1} (x - \mu) > \chi_(p, 1 - \alpha)^2\}$$

који се још назива α аутлајер регија. Велична скупа је измењена у односу на величину n скупа података. То се добија тако што се уводи услов да са вероватноћом $1-\alpha$ ниједан елемент узорка није у аутлајер регији $out(\alpha_n, \mu, \Sigma)$, тако да је $\alpha_n = 1-(1-\alpha)^{1/n}$. Детекција аутлајера се затим одређује помоћу неједнакости

$$\mathrm{OR}(x_1, ..., x_n; \alpha_n) := \{x \in \mathbb{R}^p : (x-\mu)^T C^{-1}(x-\mu) > c(\alpha_n, n, p)\}$$

Критична вредност $c(\alpha_n, n, p)$ се добија симулацијом под претходно наведеним условом, да са вероватноћом $1-\alpha$ ниједан елемент узорка није идентификован као аутлајер. Једноставности ради у наставку ћемо овај метод означавати са BG.

3.2 Робусне оцене

Традиционално, узорачка средина и узорачка дисперзија дају добре оцене за положај и облик података ако они нису контаминирани аутлајерима. Када је база података контаминирана, ови параметри могу одступати и значајно деловати на перформансу детекције аутлајера.

Хампел [5] је увео концепт тачке лома, као мере робусности оцене аутлајера. Тачка лома је дефинисана као најмањи проценат аутлајера који може проузроковати да оцена прими произвољно велике вредности. Стога, што оцена има већу тачку лома робуснија је. Нпр. узорачка средина има тачку лома од $1/n$ пошто једно веће опажање може учинити да узорачка средина и варијанса пређу сваку границу. Према томе, Хампел је предлагао медијану и медијалну апсолутну девијацију (MAD) као робусне оцене очекивања.

Други рад који наглашава проблем робусних оцена је предложен од стране Тјукија 1977. Он је увео боксплот као графички приказ на ком аутлајери могу бити идентификовани. Боксплот је заснован на квартилима расподеле. Први и трећи квартил Q_1 и Q_3, су коришћени за добијање робусних мера за средњу вредност $\hat{\mu}_n = (Q_1+Q_3)/2$ и стандардну девијацију $\hat{\sigma}_n = Q_3 - Q_1$. Друго популарно решење за добијање робусних мера је да се замени средња вредност медијаном и израчуна стандардна девијација заснована на 1-α проценту података, где је уобичајено $\alpha = 5\%$.

Постоји велики број робусних оцена које се могу користити за оцењивање вектора средњих вредности, као и коваријационе матрице приликом детектовања аутлајера. Робусне оцене се најчешће користе када имамо већи проценат аутлајера и посебно приликом рада са реалним подацима (несимулираним). Да би се избегао ефекат маскирања при детекцији

аутлајера, најсигурније је да се уместо узорачке средње вредости и узорачке коваријационе матрице користе њихове робусне оцене. Неке од метода за робусне оцене су M, MCD, Стахел-Донохове оцене итд. У наставку ћемо користити и детаљно објаснити MCD оцене.

3.2.1 MCD оцене

MCD метод је један од првих веома робусних метода за оцењивање узорачке средње вредности, као и узорачке коваријационе матрице. Отпорност на екстремне елементе узорка чини овај метод веома корисним у откривању аутлајера. Иако је већ уведен 1984. године, његова главна употреба је почела тек након увођења ефикаснијег, бржег MCD алгоритма Русова и Ван Дриесена. Од тада, MCD оцена се примењује у многим областима као што су медицина, финансије, анализа слика као и у хемији. Штавише, MCD се такође користи за развијање многих робусних вишедимензионих техника, попут анализе главних компоненти, факторске анализе и регресије.

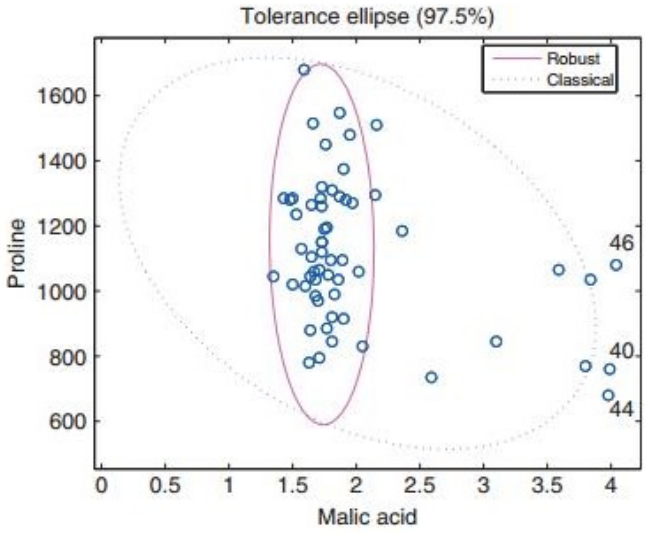
Циљ ове методе је да се нађе h елемената узорка чија коваријациона матрица има најмању детерминанту. MCD оцена узорачке средње вредности је средња вредност тих h елемената узорка, док је MCD оцена узорачке коваријационе матрице управо коваријациона матрица тих h елемената узорка.

Претпоставимо да су нам подаци задати у облику матрице X димензија $n \times p$, $X = (x_1, x_2, ..., x_n)^t$, где је $x_i = (x_{i1}, x_{i2}, ..., x_{ip})^t$ ити елемент узорка. Број n представља број елемената узорка, а p број променљивих. Ради једноставности и лакшег визуленог приказивања посматрајмо дводимензиони простор. Да бисмо објаснили како се одређују Махаланобисова растојања коришћењем робусних оцена користићемо податке о вину[20], који су јавно доступни. Вински скуп података се састоји од 13 променљивих које представљају мере квалитета 3 различите сорте италијанског вина. Посматраћемо прву групу која садржи 59 вина узимајући у обзир само 2 променљиве пролин (неесенцијална аминокиселина) и јабучну киселину[3]. Графички приказ ових података заједно са класичном и робусном 97.5% елипсом Махаланобисових растојања је приказан на слици 3.2.

Класична елипса је дефинисана скупом свих p-димензионих тачака x за које је Махаланобисово растојање

$$MD = \sqrt{(x - \bar{x})^T S^{-1} (x - \bar{x})}$$

[3]енгл- Malic acid



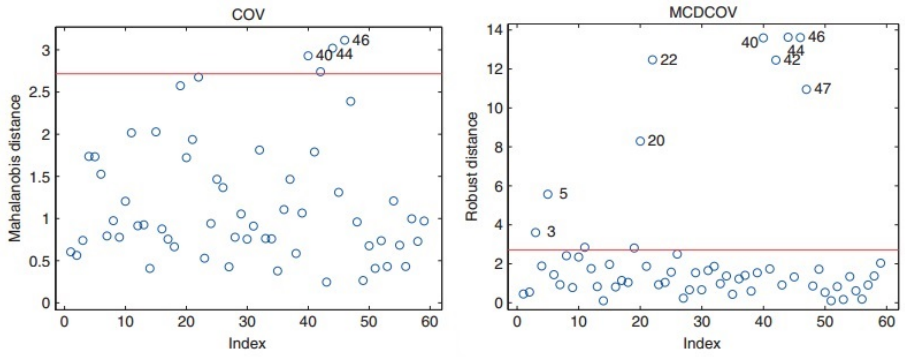
Слика 3.2: Приказ класичне и робусне Махаланобисове елипсе

једнако са $\sqrt{\chi^2_{(p,0.975)}}$, где је $\chi^2_{(p,\alpha)}$ α-квантил χ^2_p расподеле. Махаланобисова растојања $MD(x_i)$ би требало да нам кажу колико далеко је тачка x_i удаљена од центра података (медијане). $\bar{X}$ је узорачка средња вредност, а S узорачка коваријациона матрица података. Као што се може видети са слике 3.2, класична елипса покушава да обухвати све податке, док са робусном елипсом то није случај. Посматрајући Махаланобисова растојања са слике 3.3 можемо видети да се само 3 елемента узорка могу прогласити потенцијалним аутлајерима. Са друге стране робусна елипса је доста мања у поређењу са класичном елипсом и дефинисана је робусним Махаланобисовим растојањем

$$RD(x) = \sqrt{(x - \hat{\mu}_{MCD})^T \hat{\Sigma}^{-1}_{MCD}(x - \hat{\mu}_{MCD})},$$

где су $\hat{\mu}_{MCD}$ и $\hat{\Sigma}_{MCD}$ робусне MCD оцене узорачке средње вредности и узорачке коваријационе матрице респективно. На слици 3.3 робусних Махаланобисових растојања можемо приметити 8 потенцијалних аутлајера и једног мањег аутлајера. Овај пример илуструје ефекат маскирања у детекцији аутлајера. Из тог разлога је много боље користити робусне оцене као што је MCD.

Дефиниција 3.2.1 (MCD оцена) *MCD оцена за оцењивање узорачке средње вредности као и узорачке коваријационе матрице са параметром*



Слика 3.3: Махаланобисово растојање (лево) и робусно растојање (десно)

h, за који важи $[(n+p+1)/2] \leq h \leq n$, се дефинише на следећи начин: $\hat{\mu}_0$ је средња вредност h елемената узорка за које је детерминанта коваријационе матрице минимална, а $\Sigma\mu_0$ је одговарајућа коваријациона матрица тих h елемената помножена фактором c_0.

Приметимо да се MCD оцена може израчунати само ако је $h > p$, јер је у супротном за сваки h-подскуп коваријациона матрица сингуларна. Пошто је $h \geq [(n+2)/2]$, онда је овај услов испуњен за свако $n \geq 2p$.

MCD оцене су намењене првенствено елиптички симетричним унимодалним расподелама. Вишедимензиона расподела са параметрима $\mu \in \mathbb{R}^p$ и позитивно дефинитном матрицом димензије p се назива елиптички симетричном и унимодалном расподелом ако постоји строго опадајућа фукција g таква да се густина расподеле може написати у следећем облику

$$f(x) = \frac{1}{|\Sigma|} g((x-\mu)^T \Sigma^{-1}(x-\mu))$$

MCD оцене су највише робусне ако се за h узме вредност $h = [(n+p+1)/2]$.

Алгоритам за одређивање броја h

Тачне MCD оцене су веома комплексне за израчунавање, посебно из разлога што за одређивање броја h треба проћи кроз $\binom{n}{h}$ подскупова. Из тог разлога је настао ефикаснији метод за одређивање MCD оцена, такозвани брзи MCD метод. Овај метод се заснива на принципу C-корака.

Теорема 3.1 *Посматрајмо скуп података* $X_n = \{x_1, ..., x_n\}$ *са* p *променљивих. Нека је* $H_1 \subset \{1, ..., n\}$, $|H_1| = h$ *и*

$$T_1 := \frac{1}{h} \sum_{i \in H_1} x_i \qquad S_1 := \frac{1}{h} \sum_{i \in H_1} (x_i - T_1)(x_i - T_1)^T$$

Ако је $det(S_1) \neq 0$ *онда дефинишемо релативна растојања*

$$d_1(i) = \sqrt{(x_i - T_1)^T S_1^{-1} (x_i - T_1)}$$

Сада узмимо H_2 *тако да* $\{d_1(i); i \in H_2\} := \{(d_1)_{1:n}, ..., (d_1)_{h:n}\}$ *где* $(d_1)_{1:n} \leq (d_1)_{2:n} \leq ... \leq (d_1)_{n:n}$ *представљају сортирана растојања и затим израчунамо* T_2 *и* S_2 *у односу на* H_2*. Тада је*

$$det(S_2) \leq det(S_1)$$

са једнакошћу ако и само ако је $T_1 = T_2$ *и* $S_1 = S_2$*.*

Претходна теорема захтева да $det(S_1) \neq 0$. У супротном када је $det(S_1) = 0$ тада већ имамо матрицу са минималном детерминантом. Алгоритам се дефинише на следећи начин:

- Задат је h-подскуп H_{old} или пар (T_{old}, S_{old}).
- Рачунамо растојања $d_{old}(i)$ за $i = 1, ..., n$.
- Сортирамо та растојања, чиме добијамо пермутацију π за коју важи

$$d_{old}(\pi(1)) \leq d_{old}(\pi(2)) \leq ... \leq d_{old}(\pi(n))$$

- Затим је ново $H_{new} := \{\pi(1), ..., \pi(h)\}$
- Израчунамо $T_{new} := mean(H_{new})$ и $S_{new} := cov(H_{new})$

За фиксиран број димензије p, овај алгоритам има временску сложеност од $O(n)$. Понављањем C-корака добијамо итерациони процес. У случају када је $det(S_2) = 0$ или $det(S_1) = d(S_2)$ алгоритам се прекида, а у супротном настављамо алгоритам C-корака за $det(S_3)$ и тако даље. Низ $det(S_1) \geq det(S_2) \geq det(S_3)$ је ненегативан и конвергентан. Заправо, пошто постоји коначан број h подскупова, онда мора постојати индекс m за који је $det(S_m) = 0$ или $det(S_{m-1}) = det(S_m)$ чиме је ковергенција постигнута (у пракси је број m скоро увек испод 10). Ово није довољно да би $det(S_m)$ била глобални минимум за MCD оцену, али је свакако неопходан услов.

Из Теореме 3.1 и претходне чињенице се може извући идејни алгоритам за одређивање минимума. Узмимо већи број иницијалних избора за H_1 и применимо алгоритам C-корака док сваки од њих не конвергира. Затим узмимо решење са најмањом детерминантом. Ово свакако отвара нова питања, како генерисати сетове H_1, колико сетова је потребно, како избећи сетове који ће дати исте вредности за детерминанту и тако даље.

Поглавље 4

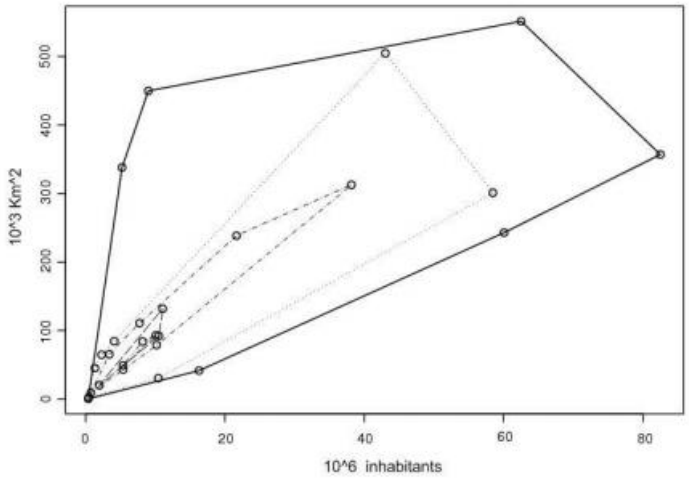
Графичке методе за детекцију аутлајера

Једне од најпознатијих техника за визуелно представљање статистичких функција дубине су балон дијаграм и вреџасти дијаграм. Ове технике се заснивају на различитим функцијама дубине. Једна од најчешћих примена ових графичких метода је детекција аутлајера. Због једноставности и лакоће представљања, фокусираћемо се на дводимензиони случај.

4.1 Балон дијаграм

Откривање аутлајера коришћењем дубине преко конвексних слојева се ради конструкцијом такозваног балон дијаграма. За дати сет података, конвексни омотач представља најмањи конвексни полигон који садржи све тачке из скупа података. Основа овог графика је функција дубине преко конвексних слојева која је дефинисана у дефиницији 1.1.3. Велика предност ове репрезентације је то што је веома интуитивна. Очигледно је да тачка са најширег слоја има мању дубину у поређењу са унутрашњим слојевима. Конструисањем конвексних омотача се може видети ако присуство аутлајера помера вредности у једну страну. За прецизнију детекцију аутлајера употребом конвексних слојева користи се балон дијаграм.

Конструкција балон дијаграма се врши тако што се формира омотач који садржи 50% свих података са центром у најдубљој тачки. Означимо омотач са $CH(X)_{.5}$. Затим се омотач $CH(X)_{.5}$ увећава са коефицијент 1.5. Нека су $V_{.5}$ ивице тог конвексног омотача $CH(X)_{.5}$. Тада се балон



Слика 4.1: Конструкција конвексних слојева

$B_{.5}$ за детекцију аутлајера дефинише као:

$$B_{.5} = \{y_i, \text{таквих да је } y_i = x_i + 1.5(x_i - CHPM), x_i \in V_{.5}\},$$

где је CHPM медијана свих података добијена помоћу дубине дефинисане конвексним слојевима. Овај метод детектује благе аутлајере, што се може изменити повећањем умноженог коефицијента. Сам балон дијаграм се може поистоветити са вишедимензионим боксплотом без вискера. Такође велике разлике у обиму узастопних омотача могу иницирати на постојање аутлајера.

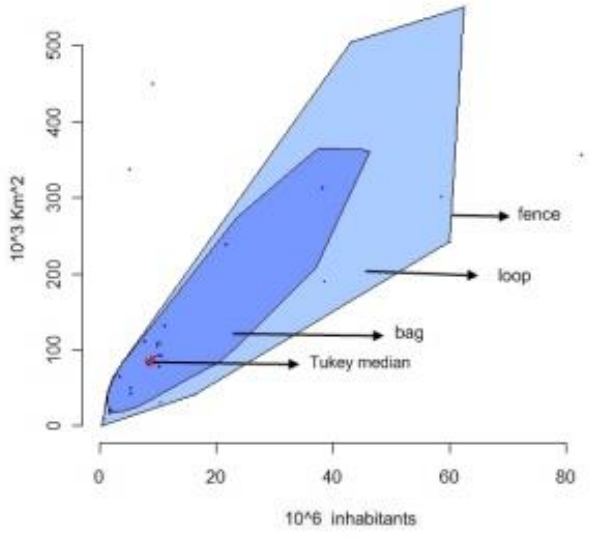
4.2 Врећасти дијаграм

Врећасти дијаграм је представио Русов (1999.) и може се сматрати генерализацијом једнодимензионог боксплота. У самом алгоритму се користи полупросторна функција дубине. Главне компоненте врећастог дијаграма су:

- Врећа[1] која садржи 50% свих елемената узорка и у оквиру те врће се налази Тјукијева медијана, тј. тачка са максималном дубином.
- Ограда[2] која одваја регуларне елементе узорка од аутлајера.

[1] енгл- Bag
[2] енгл- Fence

- Петља[3] област у којем су тачке изван вреће, а унутар ограде.



Слика 4.2: Делови врећастог дијаграма

Тачке изван петље сматрамо аутлајерима.

Као и код боксплота, врећасти дијаграм визуелно представља више карактеристика посматраних података, као што су положај (дубинска медијана), распон (величина вреће), корелацију (оријентација вреће), коефицијент асиметрије (облик вреће и петље), репови расподеле (тачке близу ивице петље и аутлајери).

Конструкција врећастог дијаграма

Положајну (полупросторну) функцију дубине $ldepth(\theta, Z)$ неке тачке $\theta \in \mathbb{R}^2$ из скупа података $Z = \{z_1, z_2, ..., z_n\}$ увео је Тјуки[7] 1975. као најмањи број z_i који се налазе у полупростору са границом коју представља права која пролази кроз тачку θ. Ефикаснији алгоритам за $ldepth(\theta, Z)$ су увели Русов и Руц[13] 1996. Дубински регион D_k је скуп свих θ за које је $ldepth(\theta, Z) > k$. То је конвексни полигон за који важи $D_{(k+1)} \subset D_k$. Ови полигони се разликују од конвексног омотача који је претходно објашњен. Дубинска медијана $T^* \in Z$ је дефинисана као θ са максималним

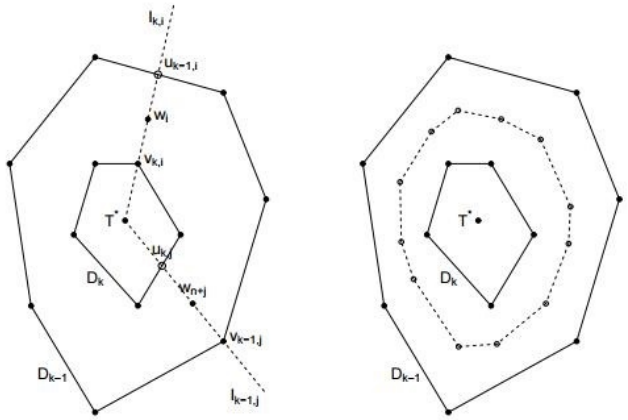
[3]енгл- Loop

$ldepth(\theta, Z)$ ако постоји само једно такво θ. У супротном, медијана се дефинише као центар најдубљег региона. Алгоритам за одређивање дубинске медијане су предложили Русов и Руц[14] 1998.

Сада конструишемо врећу W на следећи начин. Нека је $n_k = \|D_k\|$ кардиналност скупа D_k. Нека су D_k и $D_{(k-1)}$ дубински региони за које важи $n_k \leq \lfloor n/2 \rfloor < n_{(k-1)}$. Одредимо параметар λ који представља релативно растојање вреће од контура дубинских региона D_k и $D_{(k-1)}$

$$\lambda = \frac{\lfloor n/2 \rfloor - n_k}{n_{(k-1)} - n_k}$$

Затим врећу W конструишемо интерполацијом између D_k и $D_{(k-1)}$ (у односу на T^*) користећи λ и темена дубинских региона D_k и $D_{(k-1)}$. Пример конструкције вреће је дат на слици 4.3.



Слика 4.3: Конструкција вреће ($\lambda = 0.5$) приказана у 2 корака

Нека су $V_k = \{v_{(k,1)}, ..., v_{(k,n)}\}$ и $V_{(k-1)} = \{v_{(k-1,1)}, ..., v_{(k,m)}\}$ темена дубинских региона D_k и $D_{(k-1)}$ респективно. За свако $v_{(k,i)} \in V_k$, а $i = 1, ..., n$, нека је $l_{(k,i)}$ права која пролази кроз $v_{(k,i)}$ и T^*. Слично права $l_{(k-1,j)}$ која пролази кроз $v_{(k-1,j)}$ и T^* за свако $v_{(k-1,j)}$, $j = 1, ..., m$. Нека је $u_{(k,i)}$ пресек $l_{(k-1,j)}$ и D_k за сваку праву $l_{(k-1,j)}$. Слично, $u_{(k-1,j)}$ пресек $l_{(k,i)}$ и $D_{(k-1)}$ за сваку праву $l_{(k,i)}$. Темена вреће $W = \{w_1, ...w_{(n+m)}\}$ се дефинишу као

$$W_p = \begin{cases} \lambda v_{(k,i)} + (1-\lambda) u_{(k-1,i)} & \text{, где је } i = 1, ..., n \text{ и } p = 1, ..., n \\ \lambda u_{(k,j)} + (1-\lambda) v_{(k-1,j)} & \text{, где је } j = (n+1), ..., (n+m) \text{ и } p = (n+1), ..., (n+m) \end{cases}$$

Вреће W се добија повезивањем свих темена w_p.

Ограда се добија увећавањам вреће W за коефицијент 3. Ова вредност је добијена симулацијом коју су радили Русов и Руц[15] 1997. Они су за различите вредности n генерисали $m = 10000$ скупова података величине n са нормалном расподелом. Затим су израчунавали W^j, дубинску медијану T^{*j} и растојање d_i^j од W^j до тачке за свако $j = 1, ..., m$. Нека је $\|A\|$ кардиналност скупа A. Русов и Руц су израчунали $\hat{c_n}$ тако да важи

$$\|\{d_i^j > \hat{c_n}\}\| = 0,005m,$$

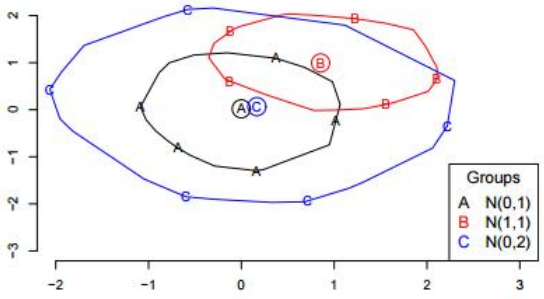
где је $i = 1, ..., n$,$j = 1, ..., m$. Значи да је 5% података дозвољено да буду аутлајери. Симулацијом је израчунато да за $n \geq 15$, вредност броја $\hat{c_n}$ је око 3. За мање n вредност $\hat{c_n}$ може бити доста велика. Русов и Руц су затим увели другу симулацију комбиновањем неколико узорака величине n таквих да имају 1000 растојања у тоталу. У таквом сету су израчунали граничну вредност такву да постоји 5 аутлајера, узимајући средњу вредност пете и шесте највеће вредности од d_i. Овај процес се понавља 1000 пута и резултат је да за $n < 10$ гранична вредност доста варира. Зато за мале сетове података је предложено да се дефинише само дубинска медијана T^* и праве између T^* и осталих тачака у скупу.

Након увећавања вреће фактором $c = 3$, тачке изван или на граници ограде се сматрају аутлајерима. Ограда се дефинише као најмањи могући конвексни слој који обухвата тачке изван вреће, а унутар увећане вреће фактором 3. Област између вреће и ограде се назива петља.

Метода вређастог дијаграма се такође може користити за поређење неколико група података, представљајући одговарајуће дијаграме на истом графику. Очигледно је да на том приказу сви дијаграми морају бити уочљиви, а то се најбоље постиже цртањем само вреће и петље заједно са медијаном за сваку групу у различитим бојама или симболима.

У следећем примеру су представљена 3 сета података $A = (X_1, Y_1)$, $B = (X_2, Y_2)$ и $C = (X_3, Y_3)$ где су X_1, Y_1 случајни узорци из нормалне расподеле $N(0, 1)$, X_2, Y_2 случајни узорци из нормалне расподеле $N(1, 1)$ и X_3, Y_3 случајни узорци из нормалне расподеле $N(0, 2)$. Сваки од узорака има по 200 елемената. Вреће сваке од група су визуелно представљене на слици 4.2 заједно са дубинском медијаном користећи различите боје и бројеве за сваку од група.

Када се елементи узорка $z_i = (x_i, y_i)$ трансформишу неком од трансформација као што су транслација или било која несингуларна линеарна трансформација, сам дијаграм се трансформише према тој трансформацији. То је зато што је полупросторна функција дубине инваријантна на таква пресликавања, као и конвексни полигони. Тако да ће



Слика 4.4: Врећасти дијаграми за 3 нормалне расподеле

након трансформације све тачке које су биле унутар вреће остати унутар вреће, као и аутлајери који су били изван ограде, остаће изван ограде.

Поглавље 5

Откривање аутлајера на подацима

У претходном делу овог рада смо детаљно објаснили појам статистичке функције дубине, методе за детекцију аутлајера базиране на Махаланобисовом растојању, као и неке од графичких метода које користе полупросторну функцију дубине као што је вређасти дијаграм. Било је и речи о робусним оценама и њиховој примени у откривању аутлајера. У наставку ћемо навести примере аутлајер детекције базиране на различитим претходно описаним методама. Тестираћемо методе за симулиране податке, као и за реалне податке.

5.1 Симулирани подаци

На самом почетку ћемо приказати резултате Монте Карло методе са 10 000 понављања, у којем желимо да добијемо просечан број аутлајера за сваку од метода. Овај поступак примењујемо на подацима из нормалне расподеле са освртима на различите средње вредности контаминираних података.

За основну расподелу изабрали смо $N(0, E)$ која има удео од 95% од узорка чији је обим 200. Интересује нас да видимо просечан број аутлајера за све 3 методе применом робусних оцена и без њихове примене. Резултати су приказани у следећој табели.

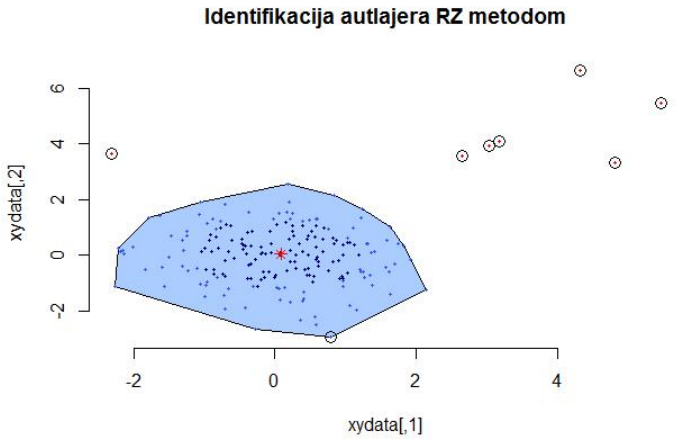
Метод/Расподела	$N((2,2),E)$	$N((2,3),E)$	$N((3,3),E)$	$N((4,4),E)$	$N((4,5),E)$
FGR	1.55	3.71	5.68	8.14	8.82
BG	11.24	12.12	12.66	13.2	13.24
RZ	6.56	7.89	8.89	10.34	10.75
FGR robust	4.39	8.01	9.95	11.47	11.57
BG robust	13.03	15.01	15.95	16.41	16.43
RZ robust	8.39	10.89	12.27	13.09	13.11

Из претходне табеле видимо да је метод FGR детектовао најмањи број аутлајера за сваку од расподела. За њега се може рећи да је либералан. Супротно FGR методи, метод BG детектује највећи број аутлајера. Могло би се закључити да метод BG детектује и поједине "благе" аутлајере. Користећи неробусне оцене за сваку од метода можемо видети да број детектованих ауталајера расте са повећавањем средње вредности контаминираних података. То се може објаснити тиме што за мале промене средње вредности много су веће шансе да се подаци из две расподеле преклапају. То је једна од отежавајућих околности за детекцију таквих аутлајера. Из тог разлога је најбоље упоредити резултате добијене коришћењем више различитих метода, како неробусних тако и робусних. Детектовање и уклањање великог броја аутлајера, посебно у случајевима малих сетова података, може знатно променити анализу резултата модела. Зато је веома важно дубље погледати шта ти детектовани аутлајери представљају за дати сет података, пре него што се елиминишу.

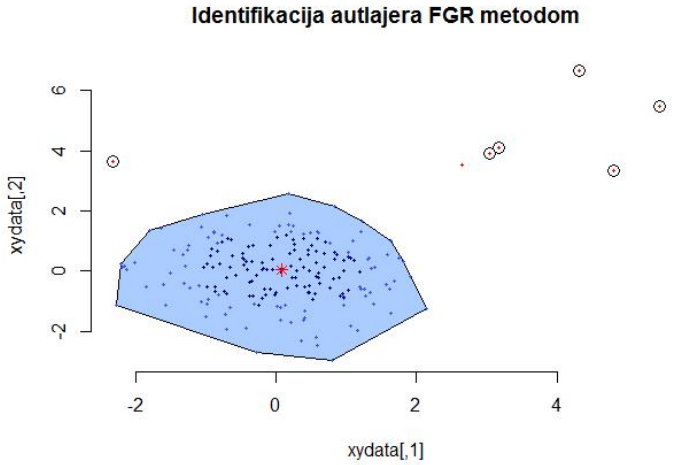
5.1.1 Нормална расподела

Промена средње вредности

Као основну расподелу изабрали смо $N(0,E)$ која има удео од 95% од целог узорка. Покушајмо да у овом случају померимо вектор средње вредности тако што ћемо на такав узорак додати контаминираних 5% елемената узорка из $N((4,4),E)$. Применом методе врећастог дијаграма која се заснива на полупросторној дубини и RZ, FGR, BG метода које се ослањају на Махаланобисово растојање добијамо следеће резултате.

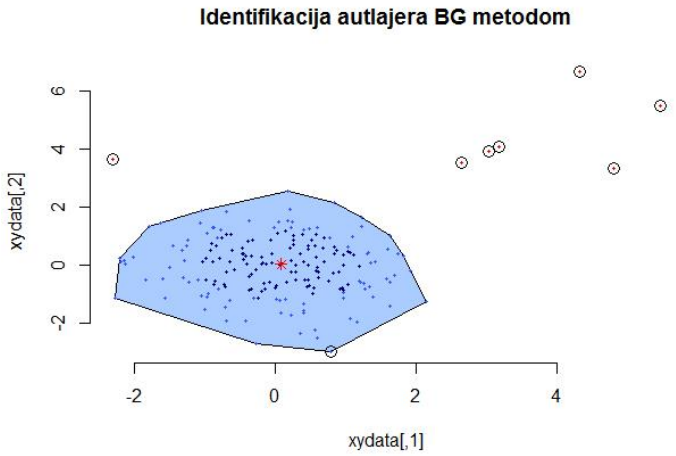


Слика 5.1: RZ метод - $N((4,4),E)$ расподела



Слика 5.2: FGR метод - $N((4,4),E)$ расподела

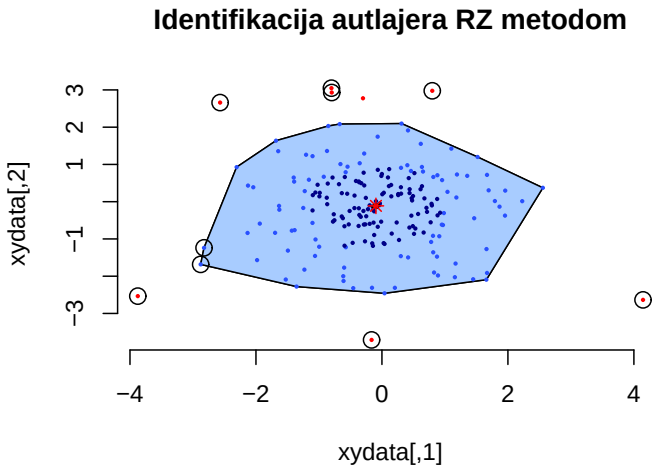
На претходном примеру можемо видети да се методе RZ и BG поклапају и детектују 8 аутлајера. Док метода FGR детектује 6 аутлајера који се поклапају са остале две методе. За разлику од њих врећасти дијаграм је детектовао 7 аутлајера.



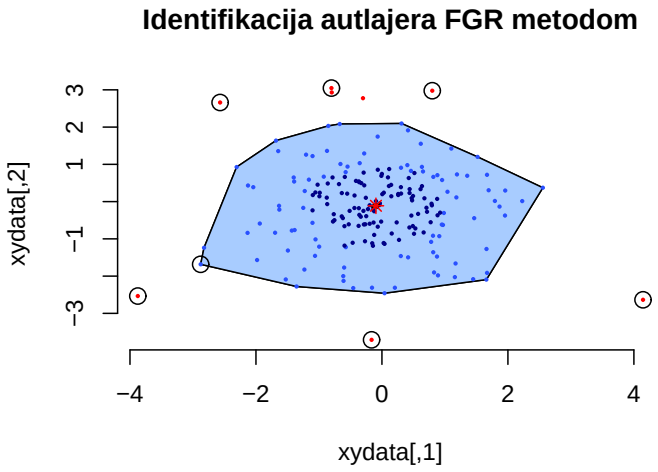
Слика 5.3: BG метод - $N((4,4),E)$ расподела

Промена дисперзије

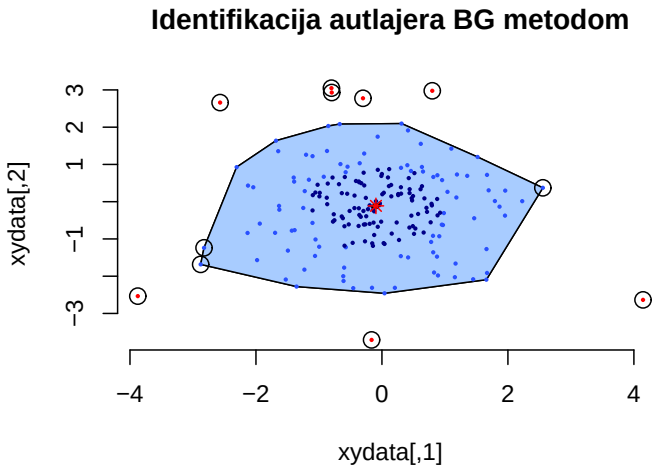
Као основну расподелу изабрали смо $N(0,E)$ која има удео од 95% од целог узорка. Покушајмо да у овом случају померимо вредност за σ_1, σ_2 тако што ћемо на такав узорак додати контаминираних 5% елемената узорка из $N(0,W)$, $W = \begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix}$. Као и у претходном примеру, упоредимо резултате добијене RZ, FGR, BG методама са методом вређастог дијаграма.



Слика 5.4: RZ метод - $N(0, W)$ расподела



Слика 5.5: FGR метод - $N(0, W)$ расподела

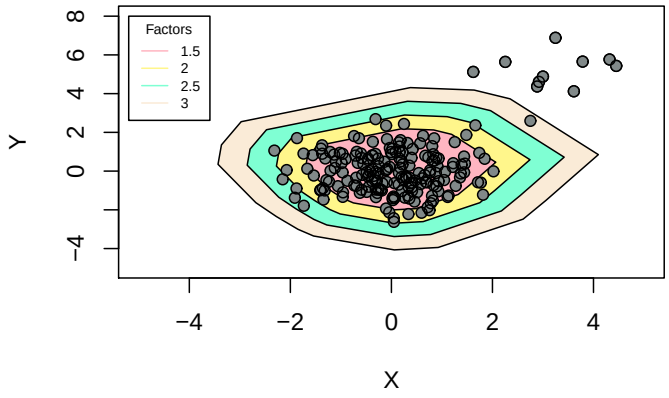


Слика 5.6: BG метод - $N(0, W)$ расподела

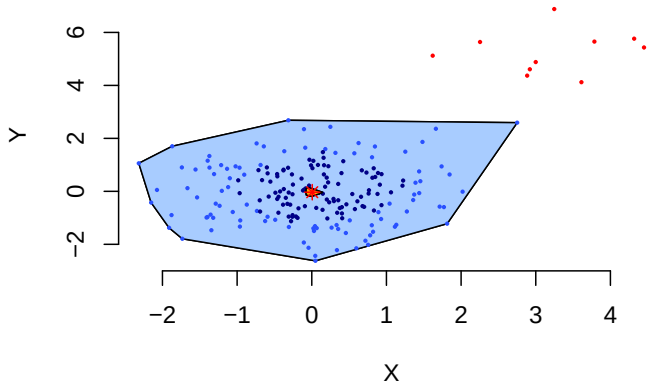
Као што можемо видети са претходних графика, када само променимо коваријациону матрицу контаминираних података добијамо мањи број детектованих аутлајера методом вређастог дијаграма. Посматрајући RZ, FGR, BG методе и у овом случају закључујемо да BG метод детектује већи број аутлајера из разлога што он узима у обзир и благе аутлајере. Ако уместо узорачке средње вредности и узорачке коваријационе матрице, искористимо робусне MCD оцене добијамо веома сличне резултате. Методе RZ, FGR и BG коришћењем робусних оцена детектују 10, 7 и 14 аутлајера респективно. Можемо извести закључак да коришћење робусних оцена код података са нормалном расподелом нема велику корист, јер је број дектованих аутлајера веома сличан оном броју добијеном неробусним оценама.

Балон дијаграм

Ако бисмо за исти скуп симулираних података обима 200 са 5% контаминираних елемената узорка са очекивањем (3,3) применили метод балон дијаграма за различите вредности фактора увећања, добили бисмо следећи график. Овај график се може упоредити са стандардним вређастим дијаграмом, код којег је фактор увећања 3.



Слика 5.7: Балон дијаграм

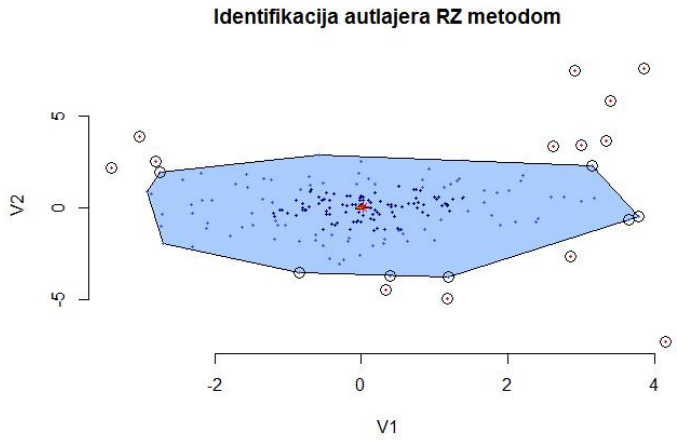


Слика 5.8: Врећасти дијаграм

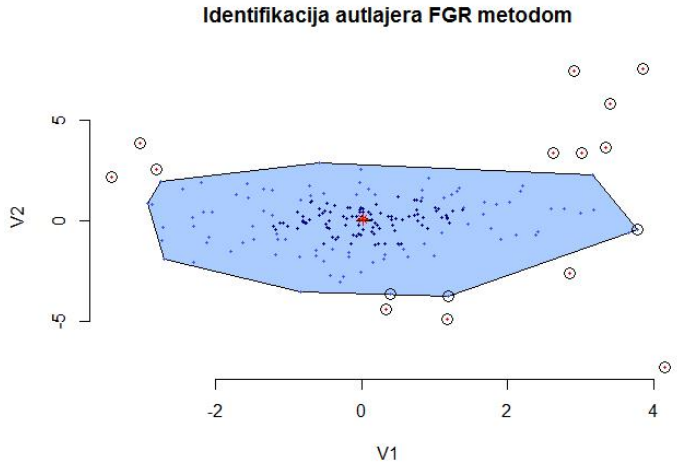
Балон дијаграм се користи за детекцију благих аутлајера, па је приликом увећања за фактор 1.5 детектовано много више аутлајера него у случају врећастог дијаграма. Уколико би фактор увећања променили у 3, добили би исте аутлајере као и у случају врећастог дијаграма, што не мора да буде увек случај.

5.1.2 Студентова расподела

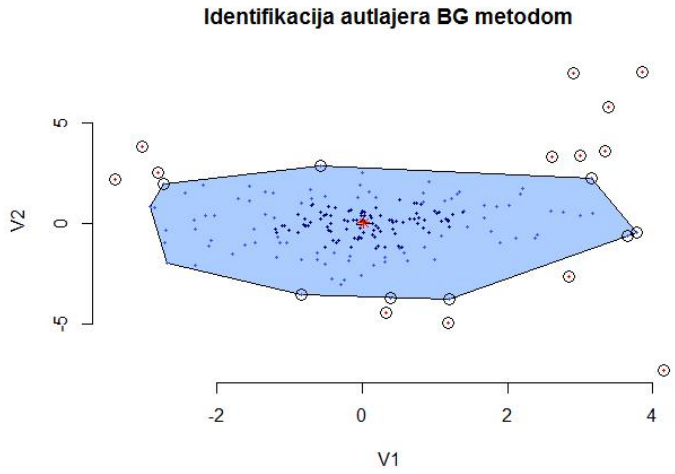
У наредном примеру ћемо користити робусне RZ, FGR, BG методе за узорак из Студентове расподеле. За основну расподелу узећемо Студентову расподелу t_3 са очекивањем (0,0). Посматрајући све три методе на узорку са Студентовом расподелом t_3 и 5% контаминираних података сА очекивањем (3,3) добијамо следеће графике.



Слика 5.9: RZ метод - t_3 расподела са очекивањем (3,3)

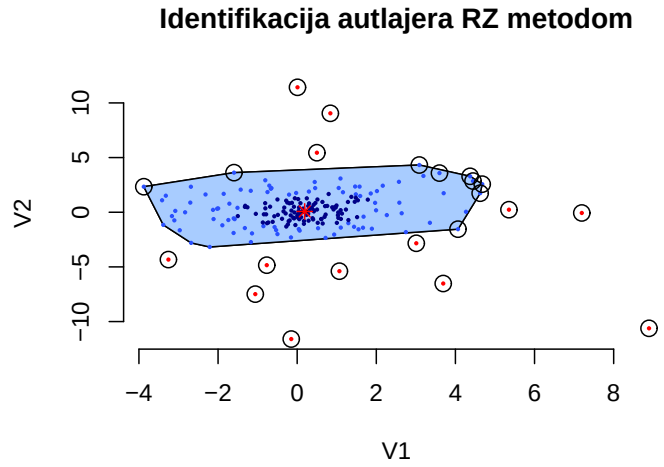


Слика 5.10: FGR метод - t_3 расподела са очекивањем (3,3)

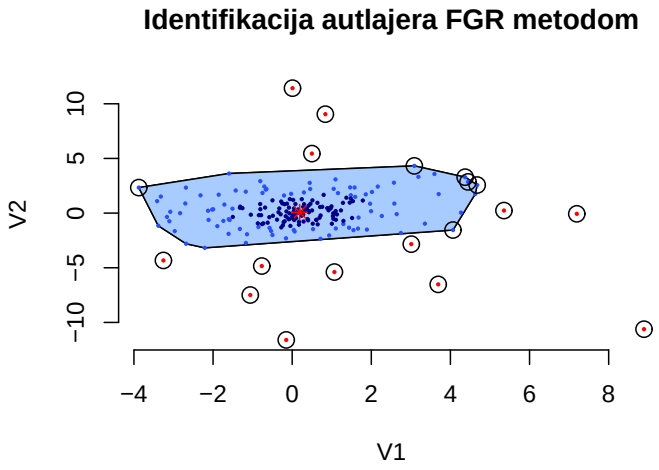


Слика 5.11: BG метод - t_3 расподела са очекивањем (3,3)

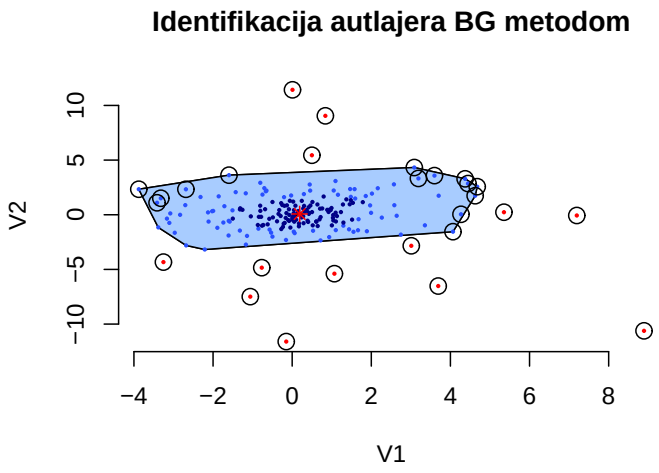
Посматрајмо сада случај Студентове расподеле са 2 степена слободе и очекивањем (2,2).



Слика 5.12: RZ метод - t_2 расподела са очекивањем (2,2)



Слика 5.13: FGR метод - t_2 расподела са очекивањем (2,2)



Слика 5.14: BG метод - t_2 расподела са очекивањем (2,2)

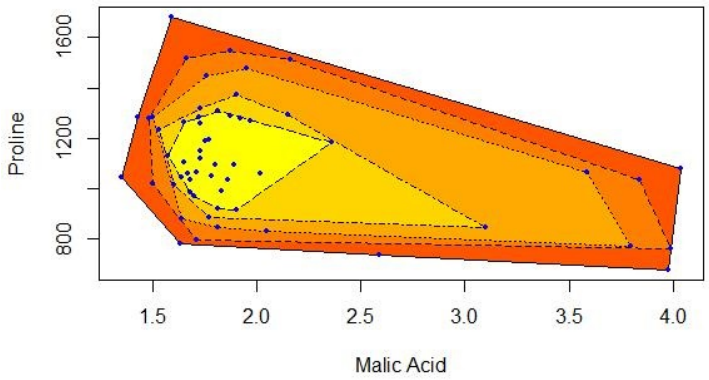
Са претходна 2 примера за Студентову расподелу можемо закључити да је број детектованих аутлајера за сваку од 3 методе знатно већи (преко 20 аутлајера) него у случају нормалне расподеле. У оба примера су коришћене робусне MCD оцене за параметре. За случај Студентове расподеле вређасти дијаграм се показао као најбоље решење, јер је детектовао око 13 аутлајера у оба случаја.

5.2 Реални подаци

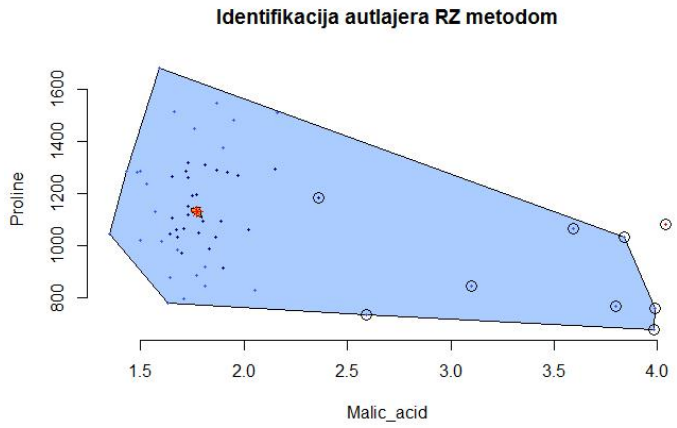
5.2.1 Подаци о вину

У поглављу о робусним оценама смо имали пример винских података које ћемо користити за детекцију аутлајера. Скуп података о вину се састоји од 13 променљивих (ниво алкохола, јабучне киселине, магнезијума, фенола, флаваноида, пепела, алкалност пепела, боја вина, интезитет боје, разблаженост вина, ниво пролина) које представљају мере квалитета 3 различите сорте италијанског вина. Посматраћемо прву групу која садржи 59 вина узимајући у обзир само 2 променљиве пролин и јабучну киселину.

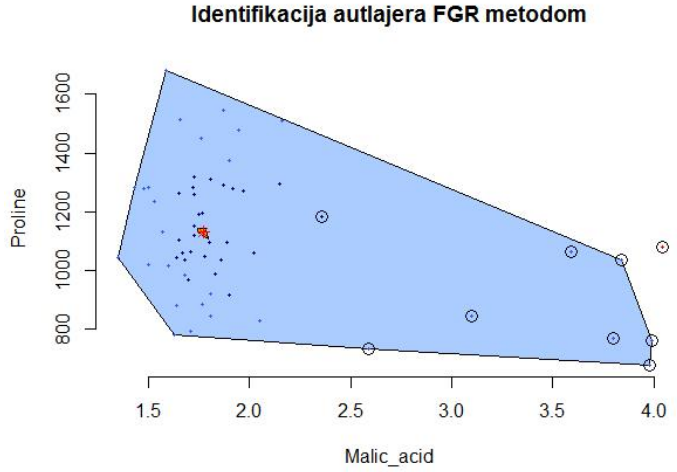
Прво ћемо формирати конвексне полигоне помоћу функције дубине преко конвексних слојева. Са самог приказа се може видети да постоје аутлајери који померају расподелу података у десно. Затим ћемо применити исте методе за детекцију као и у претходним примерима симулираних података. Једина разлика је што ћемо користити робусне MCD оцене за FGR, RZ, BG методе. Разлог зашто користимо робусне оцене је описан у одељку о MCD оценама.



Слика 5.15: Конвексни полигони



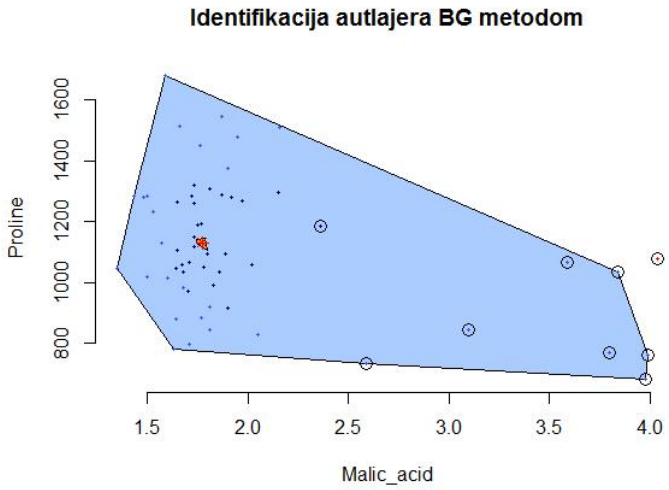
Слика 5.16: RZ метод



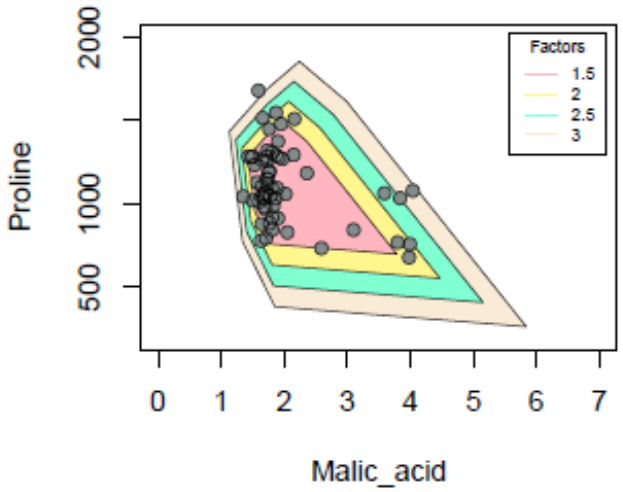
Слика 5.17: FGR метод

Из претходног можемо закључити да је врећасти дијаграм идентификовао само један аутлајер, док су остале 3 методе идентификовале истих 9 аутлајера. Ово је један од примера ефекта маскирања. Применом робусних оцена смо успели да откријемо замаскиране аутлајере, што није пошло за руком методи врећастог дијаграма.

Упоредимо претходне резултате са балон дијаграмом.



Слика 5.18: BG метод



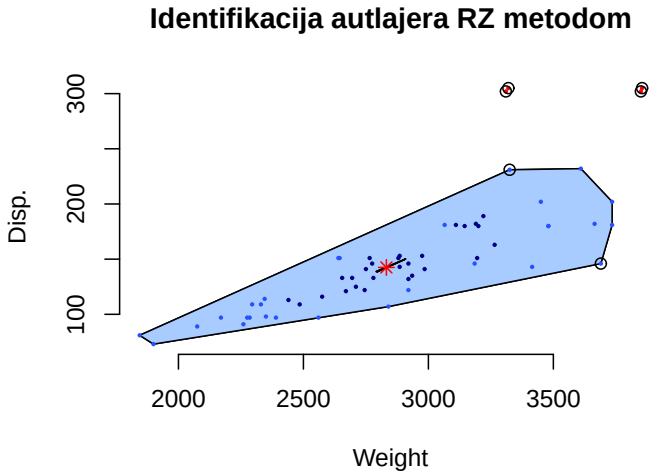
Слика 5.19: Балон дијаграм

Ако упоредимо резултате добијене методом балон дијаграма са фактором увећања 3 и врећастим дијаграмом можемо видети да су обе методе детектовале по један аутлајер, али овог пута то је различит елемент узорка.

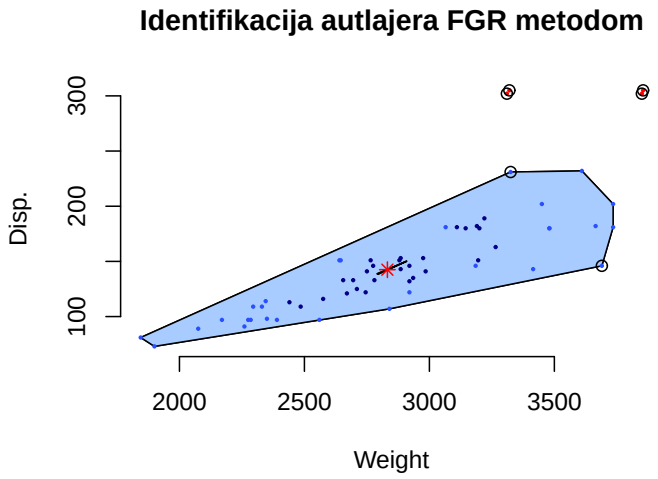
5.2.2 Подаци о аутомобилима

У часопису " *American Statistician*" Русов је објавио рад[10] о конструкцији врећастих дијаграма. Као пример за графичко представљање овог дијаграма је коришћен скуп података о аутомобилима који садржи информације о тежини и снази мотора за 60 аутомобила. Упоредимо врећасти дијаграм са аутлајерима добијеним применом RZ, FGR и BG метода. Из следећих графика можемо закључити да су методе RZ, FGR детектовале исте аутлајере, док је последњи метод BG пронашао један више аутлајер. Што се тиче врећастог дијаграма, он је детектовао најмањи број аутлајера, 4.

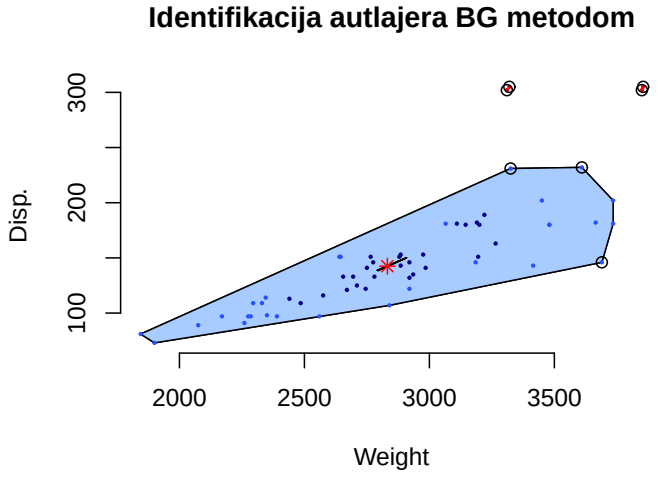
Ако погледамо колике су средње вредности за тежину и снагу аутомобила из овог скупа података, добијамо да је просечна тежина аутомобила 2900.83 кг док је просечна снага 152.05 коњских снага. Ако из скупа података уклонимо 4 аутлајера који су детектовани методом врећастог дијаграма, добијамо да је просечна тежина аутомобила 2852.05 кг док је просечна снага 141.23 коњске снаге. Из претходног можемо закључити да су аутлајери имали већи утицај на просечну коњску снагу мотора. Ако мало боље погледамо податке, видећемо да од 4 детектована аутлајера, 2 аутомобила (Chevrolet Camaro V8, Ford Mustang V8) су спортског типа, док су друга 2 (Chevrolet Caprice V8, Ford LTD Crown Victoria V8) стари модели аутомобила, такозвани "олдтајмери" који су веома дугачки. Сва 4 аутомобила имају велики број коњских снага, преко 300, као и тежину која је преко 3 тоне.



Слика 5.20: RZ метод



Слика 5.21: FGR метод



Слика 5.22: BG метод

Закључак

Статистичке функције дубине су од великог значаја за анализу података. Концепт дубина података омогућава генерализацију концепта поретка у вишедимензионом случају. Још једна велика предност овог концепта је то да је већина метода базираних на статистичким функцијама дубине робусна, што их чини посебно погодним за анализу стварних података, где су аутлајери често присутни.

Морамо напоменути да је употреба графичких метода за детекцију аутлајера од великог значаја, јер се на први поглед може видети који елементи узорка су идентификоване као аутлајери.

Идентификација аутлајера коришћењем FGR, RZ и BG метода је детаљно објашњена и приказана је њихова практична примена. Све поменуте методе су базиране на Махаланобисовом растојању. У примерима са симулираним подацима се показало да је проценат детекованих аутлајера највећи за BG метод, док FGR метод детектује знатно мање аутлајера. У претходним примерима се показало да нема великог бенефита од коришћења робусних оцена приликом откривања аутлајера за симулиране податке из нормалне расподеле. Сличан закључак се не може извести за Студентову расподелу и реалне податке. Коришћење робусне MCD оцене се показало као најбоље решење за реалне скупове података, посебно у случају ефекта маскирања.

Литература

[1] Ali S. Hadi, A Modification of a Method for the Detection of Outliers in Multivariate Samples, Journal of the Royal Statistical Society, Vol.56, No.2, 393-396, 1994

[2] Claudia Becker, Ursula Gather, The masking breakdown point of multivariate outlier identification rules, Journal of the American Statistical Association, Vol.94, No.447, 947–955, 1999

[3] Claudia Becker, Roland Fried, Sonja Kuhnt, Understanding biplots, Springer, Chapther 2, Depth Statistics 17–34, 2013

[4] Claudia Becker, Roland Fried, Sonja Kuhnt, Robustness and Complex Data Structures, Springer, 2013

[5] Frank R. Hampel, A general qualitative definition of robustness, The Annals of Mathematical Statistics, Vol.42, No.6, 1887–1896, 1971

[6] H. Caussinus, A. Ruiz, Interesting Projections of Multidimensional Data by Means of Generalized Principal Component Analyses, Compstat, 121–126, 1990

[7] John W. Tukey, Mathematics and the Picturing of Data, Proceeding of International Congress of Mathematics, Vancouver, Vol.2, 523–532, 1974

[8] Peter Filzmoser, Identification of Multivariate Outliers: S Performance Study, Austian Journal of Statistics, Vol.34, No.2, 127–138, 2005

[9] Peter Filzmoser, Robert G. Garrett, Clemens Reimann, Multivariate outlier detection in exploration geochemistry, Computers and Geosciences, Vol.31, No.5, 579-587, 2005

[10] Peter J. Rousseeuw, Ida Ruts, The Bagplot: A Bivariate Boxplot , The American Statistician, Vol.53, No.4, 382–387, 1999

[11] Peter J. Rousseeuw, B.C. Van Zomeren, Unmasking multivariate outliers and leverage points, Journal of the American Statistical Association, Vol.85, No.411, 633–651, 1990

[12] Peter J. Rousseeuw, Katrien Van Driessen, A Fast Algorithm for the Minimum Covariance Determinant Estimator, Technometrics, Vol.41, 212–223, 1999

[13] Peter J. Rousseeuw, Ida Ruts, Bivariate Location Depth, Applied Statistics (JRSS-C), Vol.45, No.4, 516–526, 1996

[14] Peter J. Rousseeuw, Ida Ruts, Constructing the Bivariate Tukey Median, Statistica Sinica, Vol.8, 827–839, 1998

[15] Peter J. Rousseeuw, Ida Ruts, The Bagplot-Bivariate Box-and-Whiskers Plot, Technical Report, Universitaire Instelling Antwerpen, Belgium, 1997

[16] Yijun Zuo, Robert Serfling, General Notions of Statistical Depth Functions, The Annals of Statistics, Vol.28, No.2, 461–482, 2000

[17] Dokument o MCD ocenama, preuzet sa `https://wis.kuleuven.be/stat/robust/papers/2010/wire-mcd.pdf`

[18] Prezentacija o dubinama preko konveksnih slojeva, preuzeto sa `https://hea-www.harvard.edu/AstroStat/Stat310MM_VI_VII/hsl_20060907.pdf`

[19] Materijali o balon dijagramu sa konferencije "Statistički izazovi u modernoj astronomiji", preuzeto sa `http://sites.stat.psu.edu/~hlee/PRESENTATION/SAMSI06.pdf`

[20] Baza podataka o parametrima tri sorte italijanskih vina, preuzeto sa `https://archive.ics.uci.edu/ml/datasets/Wine`

Биографија

Катарина Јеремић је рођена у Чачку 25. фебруара 1992. године. Завршила је Гимназију у Чачку 2011. године. Школовање је наставила на Математичком факултету Универзитета у Београду на којем је дипломирала 2015. године на смеру Статистика, финансијска и актуарска математика. Након завршеног факултета је почела да ради у области визуелизације података.

Области интересовања су математичка статистика са применом на визуелно представљање података.