УНИВЕРЗИТЕТ У БЕОГРАДУ
МАТЕМАТИЧКИ ФАКУЛТЕТ

Јелена Радојевић

МЕТОДЕ ЗА СМАЊЕЊЕ ДИМЕНЗИЈЕ ПОДАТАКА ЗАСНОВАНЕ НА АНАЛИЗИ ГЛАВНИХ КОМПОНЕНТИ

мастер рад

Београд, 2023.

Ментор:

др Бојана Милошевић, ванредни професор
Универзитет у Београду, Математички факултет

Чланови комисије:

др Марко Обрадовић, доцент
Универзитет у Београду, Математички факултет

др Марија Цупарић, доцент
Универзитет у Београду, Математички факултет

Датум одбране: _____

Садржај

Глава 1

Увод

Алгоритми статистичког и машинског учења су у последњих неколико деценија доживели значајан напредак и њихова примена је свеприсутна. Овај брзи развој је резултат напретка хардвера и све веће доступности уређаја који непрекидно прикупљају огромне количине информација. Статистичко, односно машинско учење има бројне примене у свакодневном животу, као што су персонализовани садржај, препознавање говора и гласа, детекција превара, аутоматско исправљање правописа, побољшање медицинске дијагностике и многе друге. Уз то, у данашње време све чешће се говори о појму "Big Data" који се односи на велике количине података које се генеришу и прикупљају из различитих извора. Алгоритми машинског учења нам помажу да извучемо корисне информације из ових података.

Осим машинског учења и природне науке су током протекле деценије доживеле значајну револуцију због брзог напретка технологије и лабораторијских инструмената. Добар пример јесте биомедицински домен, који је направио изузетан напредак од увођења комплетног секвенцирања генома. Ова постгеномска ера довела је до појаве нових техника високе пропусности које генеришу огромне количине података, што је резултирало експоненцијалним растом биолошких база података. У многим случајевима, ове базе података садрже већи број атрибута или променљивих у поређењу са бројем посматрања. Рецимо, стандардни скупови података се састоје од хиљада променљивих у десетинама узорака. Ова појава није искључиво везана за биомедицинска истраживања, већ се јавља и у другим областима као што су рачунарски вид, где се слике представљају као матрице пиксела, аутоматска анализа текста, анализа временских серија, интернет претраживачи и персонализоване

препоруке и тако даље.

Дакле, све чешће се дешава да базе података које треба обрадити имају велики број променљивих. Што је већи број променљивих теже их је интерпретирати, визуализовати и применити различите методе статистичког и машинског учења. Посебан проблем примене различитих алгоритама на високодимензионим подацима је рачунарска сложеност. Због свега наведеног смањење димензије представља важан корак у припреми података за даљу обраду.

Уобичајени приступ за смањење димензије је коришћење алгоритама које развијају стручњаци у одређеним областима и примењују се само на специфичне проблеме и домене. На пример, постоје алгоритми за обраду текста, алгоритми за препознавање линија или ивица на сликама и слично. Међутим, ови алгоритми претпостављају одређена својства података и решавају само конкретне проблеме, па је пожељно пронаћи алгоритме који су примењиви на ширем спектру проблема. Један такав метод јесте анализа главних компоненти (енг. *Principal Component Analysis, PCA*) која се ослања на принципе линеарне алгебре како би смањила број атрибута и елиминисала корелације у подацима. Кључна идеја је пронаћи нови координатни систем у коме се улазни подаци могу представити са много мање променљивих без значајних губитака.

Смањење димензије података је важан корак у анализи јер омогућава лакше разумевање, интерпретацију и визуализацију комплексних скупова података. Такође може побољшати перформансе модела машинског учења тако што ће смањити прекомерну сложеност и повећати генерализацију.

Глава 2

Основни појмови

У овом поглављу ћемо навести неке основне појмове из теорије вероватноће и линеарне алгебре потребне за разумевање даљег текста.

Дефиниција 2.1. *Нека је $(\Omega, \mathcal{A}, P)$ простор вероватноћа и n произвољан природан број већи од 1. Функција $\mathbf{X} : \Omega \to \mathbf{R}^n$ зове се вишедимензионална случајна величина или случајан вектор ако за сваки Борелов скуп $B \in \mathcal{B}^n$ важи $X^{-1}(B) \in \mathcal{A}$.*

Теорема 2.2. *а) Нека су $X_1, X_2, ..., X_n$ случајне величине. Тада је $\mathbf{X} = (X_1, X_2, ..., X_n)$ n-димензионална случајна величина.*

б) Нека је $\mathbf{X} = (X_1, X_2, ..., X_n)$ n-димензионална случајна величина. Тада су $X_1, X_2, ..., X_n$ случајне величине.

Из претходне теореме следи да је n-димензионална случајни вектор уређена n-торка случајних величина.

Сада ћемо дефинисати два коефицијента који се често узимају за мере зависности случајних величина. То су коваријација и коефицијент корелације.

Теорема 2.3. *Неке су X и Y случајне величине са коначним и строго већим од нуле дисперзијама. Коваријација случајних величина X и Y је број $cov(X, Y)$ задат са*

$$cov(X, Y) = E(X - EX)(Y - EY) = E(XY) - EXEY\,.$$

Коефицијент корелације случајних величина X и Y је број

$$\rho_{X,Y} = \frac{E(XY) - EXEY}{\sqrt{DX}\sqrt{DY}}\,.$$

Случајне величине X и Y су *некорелиране* ако су њихова коваријација, а самим тим и коефицијент корелације једнаки нули.

Теорема 2.4. *Нека су X и Y случајне величине са коначним дисперзијама различитим од нуле. За коефицијент корелације тих случајних величина важи неједнакост $|\rho_{X,Y}| \leq 1$. При томе, једнакост важи ако и само ако је зависност случајних величина X и Y линеарна.*

Дефиниција 2.5. *Нека је $\mathbf{X} = (X_1, X_2, ..., X_n)$ случајан вектор, такав да за све $k \in \{1, 2, ..., n\}$ важи $EX_k^2 < +\infty$. Матрица $B = [b_{kl}]_{n\times n}$, где је $b_{kl} = cov(X_k, X_l)$ зове се коваријациона матрица или матрица коваријансе случајног вектора $\mathbf{X}$.*

Теорема 2.6. *Матрица $B = [b_{kl}]_{n\times n}$ је коваријациона матрица или матрица коваријансе неког случајног вектора $(X_1, X_2, ..., X_n)$ ако и само ако је B симетрична и ненегативно дефинитна[1] матрица.*

У наставку наводимо још неке резултате из линеарне алгебре.

Дефиниција 2.7. *Вектор $v \in \mathbf{R}^n$ је сопствени вектор матрице $\mathbf{A} \in \mathbf{R}^{n\times n}$ са одговарајућом сопственом вредношћу $\lambda \in \mathbf{R}$ ако је*

$$\mathbf{A}v = \lambda v,\ v \neq 0.$$

Дефиниција 2.8. *За матрицу $\mathbf{A} \in \mathbf{R}^{n\times n}$ дефинишемо карактеристични полином од $\mathbf{A}$ као*

$$k_{\mathbf{A}}(z) = det(\mathbf{A} - z\mathbf{I})\,.$$

Теорема 2.9. *$\lambda \in \mathbf{R}^n$ је сопствена вредност матрице $\mathbf{A}$ ако и само ако је $k_{\mathbf{A}}(\lambda) = 0$.*

Теорема 2.10. *Нека је $\mathbf{A} \in \mathbf{R}^{n\times n}$ позитивно семидефинитна матрица. Тада су њене сопствене вредности ненегативне. Сопствени вектори који одговарају различитим сопственим вредностима матрице $\mathbf{A}$ су ортогонални.*

Теорема 2.11. *Свака симетрична матрица $\mathbf{A}$ је ортогонално слична дијагоналној матрици, односно важи $\mathbf{A} = \mathbf{PDP}^T$, где је $\mathbf{P}$ ортогонална, а $\mathbf{D}$ дијагонална матрица.*

[1]За симетричну матрицу $A \in \mathbf{R}^{n\times n}$ кажемо да је ненегативно дефинитна ако за сваки вектор $x \in \mathbf{R}^n$ важи $x^T\mathbf{A}x \geq 0$.

Глава 3

Анализа главних компоненти

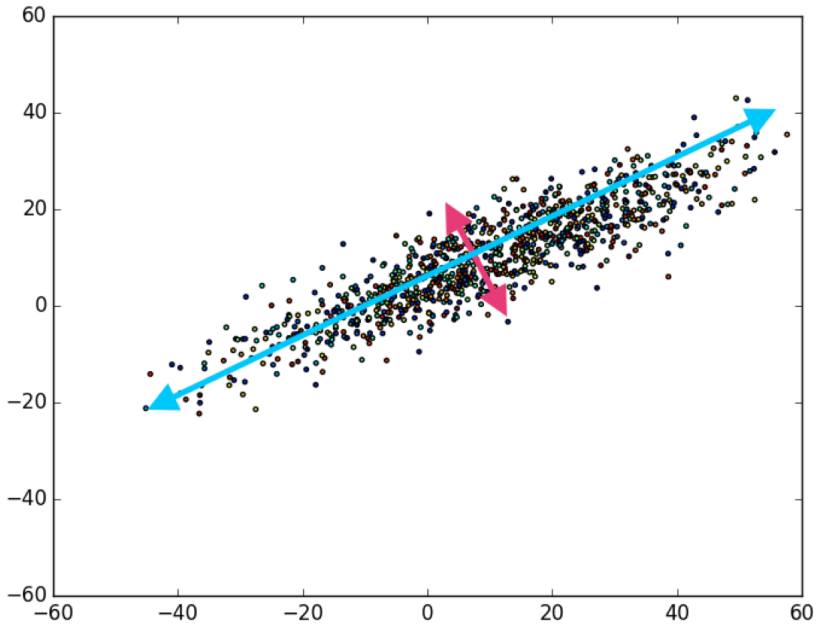
Анализа главних компоненти (енг. *Principal Component Analysis, PCA*) је најчешће коришћен алгоритам за смањење димензије података која се ослања на методе линеарне алгебре. Методу је први пут представио Карл Пирсон 1901. године, који је имао важну улогу у стварању теорије модерне статистике. Иако је вршио израчунавања са само две или три променљиве Пирсон је веровао да се анализа главних компоненти може употребити и за решавање проблема са пуно више променљивих. Неколико деценија касније 1933. године, опис израчунавања као и назив под којим је метод данас познат дао је Харолд Хотелинг, амерички математичар. Међутим, и даље су израчунавања била компликована и заморна када би требало направити анализу са већим бројем атрибута. Широка употреба анализе главних компоненти је уследила тек са појавом рачунара.

3.1 Циљеви анализе главних компоненти

Метод главних компоненти се може користити у различите сврхе, укључујући визуализацију података, избор атрибута и компресиују података. У визуализацији података, метод се може користити за исцртавање високодимензионих података у две или три димензије, што олакшава тумачење. У избору атрибута, метод се може користити за идентификацију најважнијих променљивих у скупу података.

Као што је већ поменуто, основни задатак анализе главних компоненти јесте смањење димензије података уз очување најважнијих информација које ти подаци носе. Да би се то постигло, метод израчунава нове променљиве

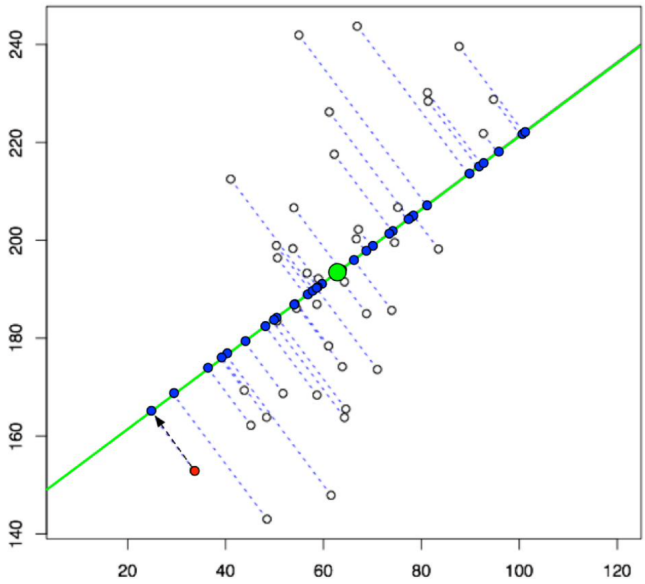
такозване *главне компоненте* које се добијају као линеарне комбинације оригиналних променљивих.



Слика 3.1: Главне компоненте дводимензионог скупа података

Постављањем координатних оса дуж праваца највеће варијабилности, интуитивно се задржава највише информација. Прва главна компонента се простире дуж правца највеће варијабилности скупа, док је друга главна компонента ортогонална на њу и такође је дуж правца највеће варијабилности међу свим ортогоналним правцима. Варијабилност података дуж неког правца се односи на раштрканост пројекција података на тај правац, што је илустровано на слици 3.2.

Метода главних компоненти је веома погодна када између података постоји корелација или приближно линеарна веза, односно када подаци леже у линеарном потпростору, који често има знатно мању димензију у односу на почетни простор. Интуитивно, висока корелисаност међу атрибутима значи да се неки од њих могу приближно изразити у функцији других и да нису сви подједнако важни. Да бисмо ово искористили, потребно је да утврдимо који атрибути су колико корелисани. Стандардни алати којима се описују такве зависности су матрица корелација и коваријансе. У случају да су корелације између великог броја променљивих јаке, нови простор може садржати



Слика 3.2: Варијабилност података дуж неког правца. Приказани правац није оптималан, већ би га требало ротирати супротно смеру кретања казаљке на сату.

велику количину информације, а имати значајно мању димензију. Додатно, ортоганлност новог система и њом условљена некорелисаност нових атрибута, носи рачунске погодности за различите методе.

Све у свему, овај метод је моћан алат за анализу података који нам помаже да сложене скупове поједноставимо и учинимо их лакшим за разумевање и даљи рад.

3.2 Дефиниција главних компоненти

Основни задатак методе главних компоненти јесте одређивање оне линеарне комбинације оригиналних променљивих која ће имати максималну дисперзију. Други, општији задатак ове методе, јесте одређивање неколико линеарних комбинација оригиналних променљивих које ће, поред тога што имају максималну дисперзију, бити међусобно некорелисане, губећи у што мањој мери информацију садржану у скупу оригиналних променљивих. Без умањења општости, претпоставимо да су подаци нулте средње вредности, а потпростор који треба да се уклопи је линеарни. У пракси проседи по колона-

ма углавном нису једнаки нули и због тога се од свих елемената сваке колоне одузима средња вредност те колоне и надаље се ради са тако трансформисаним подацима.

Претпоставимо да је $\mathbf{X} = (X_1, X_2, ..., X_p)$ p-димензиони случајан вектор са матрицом коваријансе $\mathbf{\Sigma}$. Нека је

$$Z_1 = a_{11}X_1 + a_{12}X_2 + ... + a_{1p}X_p = \mathbf{a}_1^T\mathbf{X}$$

линеарна комбинација елемената случајног вектора $\mathbf{X}$, где су $a_{11}, a_{12}, ..., a_{1p}$ коефицијенти линеарне комбинације. Тада је

$$D(Z_1) = D(a_1^T\mathbf{X}) = a_1^T\Sigma a_1\,. \tag{3.1}$$

Наш задатак је да одредимо p-димензиони вектор $\mathbf{a_1}$ који максимизује дисперзију од Z_1, односно квадратну форму $a^T\mathbf{\Sigma}a$. Да би овај проблем имао добро дефинисано решење, потребно је наметнути додатно ограничење, а најчешће ограничење подразумева рад са векторима јединичне норме, односто захтева се

$$a_1^T a_1 = 1\,. \tag{3.2}$$

Класичан приступ решавању овог проблема јесте Лагранжова метода за одређивање условних екстремума. Дакле, потребно је максимизовати функцију

$$\Lambda(a_1; \lambda) = a_1^T\Sigma a_1 - \lambda(a_1^T a_1 - 1)\,, \tag{3.3}$$

где је λ Лагранжов множилац. Диференцирањем Лагранжове функције по коефицијентима a_1 и изједначавањем добијеном израза са нулом, добијамо

$$\frac{\partial\Lambda}{\partial a_1} = 2\mathbf{\Sigma}a_1 - 2\lambda a_1 = 0\,,$$

односно

$$(\mathbf{\Sigma} - a\mathbf{I})a_1 = 0 \iff \mathbf{\Sigma}a = \lambda a \tag{3.4}$$

где је $\mathbf{I}$ јединична матрица формата $p \times p$. Дакле, вектор a_1 мора бити јединични сопствени вектор, а λ одговарајућа сопствена вредност матрице коваријансе $\mathbf{\Sigma}$. Приметимо да за дисперзију случајног вектора $a_1^T\mathbf{X}$ важи:

$$D(a_1^T\mathbf{X}) \overset{(3.1)}{=} a_1^T\mathbf{\Sigma}a_1 \overset{(3.4)}{=} a_1^T\lambda a_1 = \lambda a_1^T a_1 \overset{(3.2)}{=} \lambda \tag{3.5}$$

па да бисмо максимизовали ову дисперзију потребно је да максимизујемо λ. Одавде следи да је λ највећа сопствена вредност матрице коваријансе $\mathbf{\Sigma}$. Прва главна компонента случајног вектора $\mathbf{X}$ једнака је $a_1^T\mathbf{X}$ при чему је a_1

сопствени вектор који одговара највећој сопственој вредности матрице коваријансе $\mathbf{\Sigma}$. Како је $\mathbf{\Sigma}$ позитивно семидефинитна матрица, све њене сопствене вредности су веће или једнаке нули. Претпоставимо да су све сопствене вредности матрице $\mathbf{\Sigma}$ међусобно различите и да су веће од 0, односно да важи $\lambda_1 > \lambda_2 > ... > \lambda_p > 0$. Тада је $D(a_1^T\mathbf{X}) = \lambda_1$.

Када желимо да одредимо више од једне линеарне комбинације, поступамо слично као и при одређивању прве главне компоненте, али уз додатни услов да је коваријанса између прве и друге главне компоненте нула. Нека је

$$Z_2 = a_{21}X_1 + a_{22}X_2 + ... + a_{2p}X_p = a_2^T X$$

чије коефицијенте $a_{21}, a_{22}, ..., a_{2p}$ треба одредити уз услов $a_2^T a_2 = 1$, при чему се услов некорелисаности прве и друге главне компоненте своди на $a_2^T a_1 = 0$. Овај услов следи из:

$$\begin{aligned} cov(Z_1, Z_2) &= cov(a_1^T\mathbf{X}, a_2^T\mathbf{X}) \\ &= E[(a_1^T\mathbf{X} - E(a_1^T\mathbf{X}))(a_2^T\mathbf{X} - E(a_2^T\mathbf{X}))^T] \\ &= E[a_1^T(\mathbf{X} - E(\mathbf{X}))(\mathbf{X} - E(\mathbf{X}))^T a_2] \\ &= a_1^T E((\mathbf{X} - E(\mathbf{X}))(\mathbf{X} - E(\mathbf{X}))^T)a_2 \\ &= a_1^T\mathbf{\Sigma}a_2 = a_2^T\mathbf{\Sigma}a_1 \\ &= a_2^T\lambda_1 a_1 = \lambda_1 a_1^T a_2\,. \end{aligned}$$

Даље, формирамо Лагранжову функцију са два множитеља λ и ν

$$\Lambda(a_1, a_2; \lambda, \nu) = a_2^T\mathbf{\Sigma}a_2 - \lambda(a_2^T a_2 - 1) - 2\nu a_2^T a_1\,.$$

Диференцирањем по a_2 и изједначавањем са 0 добијамо

$$\frac{\partial\Lambda}{\partial a_2} = 2\mathbf{\Sigma}a_2 - 2\lambda a_2 - \nu a_1 = 0\,. \tag{3.6}$$

Множењем добијеног израза слева са a_1^T имамо

$$2a_1^T\mathbf{\Sigma}a_2 - 2a_1^T\lambda a_2 - a_1^T\nu a_1 = 0\,.$$

Како су прва два члана у претходном изразу једнака нули следи да је $\nu a_1^T a_1 = 0$, а због $a_1^T a_1 = 1$ следи $\nu = 0$. Уврштавањем добијеног у израз (2.4) добијамо

$$\mathbf{\Sigma}a_2 - \lambda a_2 = 0\,,$$

одакле закључујемо да је λ сопствена вредност матрице $\mathbf{\Sigma}$, а a_2 одговарајући сопствени вектор. Слично као и за дисперзију прве главне компоненте, може се показати да за дисперзију друге главне компоненте важи

$$D(Z_2) = D(a_2^T\mathbf{X}) = a_2^T X a_2 = \lambda\,.$$

Будући да желимо да максимизујемо дисперзију вектора $a_2^T\mathbf{X}$, бирамо за λ што је могуће већу вредност. Како смо претпоставили да су све сопствене вредности матрице $\mathbf{\Sigma}$ различите, следи да је $\lambda \neq \lambda_1$. Ако би важило да је $\lambda = \lambda_1$ то би значило да је $\lambda_2 = \lambda_1$ што је у контрадикцији са условом $a_2^T a_1 = 0$. Дакле, за λ узимамо другу највећу сопствену вредност од $\mathbf{\Sigma}$, то јест $\lambda = \lambda_2$, а a_2 је одговарајући сопствени вектор. Слично се може показати да је k-та главна компонента случајног вектора $\mathbf{X}$ једнака $a_k^T\mathbf{X}$ и $D(a_k^T\mathbf{X}) = \lambda_k$ при чему је λ_k k-та највећа сопствена вредност матрице $\mathbf{\Sigma}$, а a_k одговарајући сопствени вектор за $k = 3, ..., p$.

3.3 Особине главних компоненти

У овом делу ћемо размотрити нека математичка и статистичка својства главних компоненти матрице коваријансе $\mathbf{\Sigma}$ случајног вектора X. Да бисмо извели математичка својства главних компоненти, користимо алгебарски приступ који се ослања на анализу сопствених вредности и сопствених вектора матрице коваријансе $\mathbf{\Sigma}$. Нека је Z p-димензиони вектор чији је k-ти елемент, Z_k, k-та главна компонента вектора $\mathbf{X}$ за $k = 1, ..., p$. Тада је

$$Z = \mathbf{A}^T\mathbf{X}$$

при чему је $\mathbf{A}$ ортогонална матрица чија је k-та колона, a_k, k-ти сопствени вектор матрице $\mathbf{\Sigma}$ придружени одговарајућим сопственим вредностима. Ортогоналност матрице $\mathbf{A}$ следи из особина сопствених вектора ($a_j^T a_j = 1$ и $a_i^T a_j = 0$, $i \neq j$). На овај начин смо главне компоненте дефинисали помоћу ортогоналне линеарне трансформације случајног вектора $\mathbf{X}$. За трансформацију се каже да ортогонална јер се са њом врши ротација координатних оса за известан угао, при чему осе остају управне једна на другу, а угао између ма која два вектора остаје исти након трансформације. Како су колоне матрице $\mathbf{A}$ сопствени вектори матрице $\mathbf{\Sigma}$ важи следеће

$$\mathbf{\Sigma A} = \mathbf{A\Lambda}\,,$$

при чему је $\mathbf{\Lambda}$ дијагонална матрица чији је дијагонални елемент, λ_k, k-та сопствена вредност матрице $\mathbf{\Sigma}$, односно $\lambda_k = D(a_k^T\mathbf{X}) = D(Z_k)$. Како је матрица $\mathbf{A}$ ортогонална важи следеће:

$$\mathbf{A}^T\mathbf{\Sigma}\mathbf{A} = \mathbf{\Lambda},$$

$$\mathbf{\Sigma} = \mathbf{A}\mathbf{\Lambda}\mathbf{A}^T.$$

Теорема 3.1. *Нека је $\mathbf{A} \in \mathbf{R}^{p\times p}$ ортогонална матрица и $\mathbf{X} = (X_1, ..., X_p)$ случајан вектор. Тада ортогонална трансформација $Y = \mathbf{A}\mathbf{X}$ случајног вектора $\mathbf{X}$ и сам случајан вектор $\mathbf{X}$ имају једнаку генерализовану дисперзију као и збир дисперзија компоненти.*

Доказ. Нека је $E(\mathbf{X}) = 0$ и $E(\mathbf{X}\mathbf{X}^T) = \mathbf{\Sigma}$. Тада је $E(Y) = 0$ и $E(YY^T) = E(\mathbf{A}\mathbf{X}(\mathbf{A}\mathbf{X})^T) = \mathbf{A}\mathbf{\Sigma}\mathbf{A}^T$. Генерализована дисперзија вектора Y је

$$det(\mathbf{A}\mathbf{\Sigma}\mathbf{A}^T) = det(\mathbf{A})det(\mathbf{\Sigma})det(\mathbf{A}^T) = det(\mathbf{\Sigma})det(\mathbf{A}\mathbf{A}^T) = det(\mathbf{\Sigma})\,,$$

што је једнако генерализованој дисперзији случајног вектора X. Сума дисперзија компоненти вектора Y је

$$\sum_{i=1}^{p} E(Y_i^2) = tr(\mathbf{A}\mathbf{\Sigma}\mathbf{A}^t) = tr(\mathbf{\Sigma}\mathbf{A}^T\mathbf{A}) = tr(\Sigma I) = \sum_{i=1}^{p} E(X_i^2)\,.$$

□

Последица 3.2. *Генерализована дисперзија вектора главних компоненти једнака је генерализованој дисперзији оригиналног вектора, а збир дисперзија главних компоненти једнак је збиру дисперзија оригиналних случајних промељивих случајног вектора.*

Дакле, ротирањем координатног система нисмо променили укупан варијабилитет система. Како важи

$$\sum_{i=1}^{p} D(Z_i) = \sum_{i=1}^{p} \lambda_i = tr(\Lambda) = \sum_{i=1}^{p} D(X_i)\,,$$

можемо рећи да је удео објашњеног варијабилитета i-том главном компонентом $\frac{\lambda_i}{\sum_{j=1}^{p}\lambda_j}$ за $i = 1, 2, ..., p$. Слично, можемо рећи да првих m главних компоненти објашњава $\frac{\sum_{j=1}^{m}\lambda_j}{\sum_{j=1}^{p}\lambda_j}$ варијабилитета оригиналних података.

Уколико у анализи главних компоненти добијемо релативно висок допринос једне или неколико првих главних компоненти укупној дисперзији, тада је могуће даљу анализу засновати на њима, а не на свим главним компонентама. У овом контексту, занимљиво је указати на разлагање коваријационе матрице $\mathbf{\Sigma}$ на матрице доприноса сваке главне компоненте коваријационој структури оригиналног скупа података.

Теорема 3.3. *(Спектрална декомпозиција матрице коваријансе $\mathbf{\Sigma}$)*
За матрицу коваријансе $\mathbf{\Sigma}$ важи

$$\mathbf{\Sigma} = \lambda_1 a_1 a_1^T + \lambda_2 a_2 a_2^T + ... + \lambda_p a_p a_p^T \,, \tag{3.7}$$

где су $\lambda_1 > \lambda_2 > ... > \lambda_p \geq 0$ сопствени вредности матрице $\mathbf{\Sigma}$, а $a_1, a_2, ..., a_p$ одговарајући сопствени вектори.

Доказ. Из $\mathbf{\Sigma} = \mathbf{A\Lambda A}^T$ расписивањем десне стране једнакости добијамо

$$\mathbf{\Sigma} = \sum_{k=1}^{p} \lambda_k a_k a_k^T \,,$$

чиме је доказ завршен. □

Овај резултат ће се касније показати корисним. Посматрајући дијагоналне елементе, примећујемо да је

$$D(X_j) = \sum_{k=1}^{p} \lambda_k a_{kj}^T \,.$$

Можда главна статистичка импликација овог резултата је та да не само да можемо дисперзије компоненти вектора $\mathbf{X}$ раставити на опадајуће доприносе главних компонент, већ можемо раставити и целу матрицу коваријансе на доприносе $\lambda_k a_k a_k^T$ сваке од главних компоненти.

Задржавајући мањи број главних компоненти од p, коваријациону матрицу $\mathbf{\Sigma}$ апроксимирамо збиром матрица доприноса задржаних главних компоненти. Када допринос укупне дисперзије задржаних компоненти премаши унапред дефинисану вредност, као што је на пример 80%, може се очекивати да ће та апроксимација коваријационе матрице $\mathbf{\Sigma}$ пружити релативно добру репрезентацију коваријационе структуре оригиналног скупа података.

3.4 Главне компоненте са једнаким и/или нултим дисперзијама

На кратко ћемо се осврнути на два проблема која се могу јавити у теорији, али су релативно ретка у пракси. До сада смо претпостављали, имплицитно или експлицитно, да су све сопствене вредности матрице коваријансе међусобно различите и да ниједна од њих није нула.

Једнакост сопствених вредности, и самим тим једнакост дисперзија главних компоненти, може се јавити код одређених матрица. У случају једнакости q сопствених вредности, одговарајући сопствени вектори разапињу q-димензионални простор, али унутар тог простора, осим што су међусобно ортогонални, су произвољни. Геометријски, оно што се дешава за $q = 2$ или $q = 3$ јесте да се главне осе круга или сфере не могу једнозначно дефинисати; сличан проблем се јавља код хиперсфере[1] када је $q > 3$. Дакле, главне компоненте које одговарају сопственим вредностима које се понављају нису јединствено дефинисане.

Други проблем, да је нека од сопствених вредности једнака нули појављује се чешће у пракси, али и тај случај је врло редак. Ако је q сопствених вредности једнако нули, тада је ранг матрице $\mathbf{\Sigma}$ једнак $p - q$ уместо q. Свака главна компонента чија је дисперзија једнака 0 дефинише константну линеарну везу између елемената вектора X. Ако оваква веза постоји, то значи да је једна променљива редундантна за сваки однос, јер се њена вредност може тачно одредити из вредности осталих променљивих које се појављују у тој вези. Због тога бисмо могли број променљивих смањити са p на $p - q$ без губитка информација, те поново израчунати главне компоненте.

3.5 Интерпретација главних компоненти

До сада смо анализу главних компоненти базирали на матрици коваријансе $\mathbf{\Sigma}$. Проблем који се јавља у интерпретацији главних компоненти последица је њихове осетљивости на различите мерне скале оригиналних променљивих. Ако у анализи једна од променљивих има знатно већу дисперзију од осталих,

[1] Хиперсфера представља генерализацију појма сфере у простору произвољне димензије. То је површ која се може дефинисати као скуп свих тачака које се налазе на датом растојању од дате тачке

тада ће та променљива доминирати првом главном компонентом без обзира на корелациону структуру података. Са друге стране, уколико скалирамо променљиве тако да све имају једнаке дисперзије, утицај променљиве на прву главну компоненту ће се променити. Због овога, нема смисла спроводити метод главних компоненти на матрици коваријансе ако променљиве немају приближно сличне вредности дисперзије.

Пример. Нека променљива X_1 представља дужину која може бити мерена у центиметрима или у милиметрима, а променљива X_2 представља тежину у грамима. Нека је $\mathbf{\Sigma}_1$ матрица коваријансе случајног вектора $\mathbf{X} = (X_1, X_2)$ када се дужина мери у центиметрима, а $\mathbf{\Sigma}_2$ матрица коваријансе када се дужина мери у милиметрима:

$$\mathbf{\Sigma}_1 = \begin{bmatrix} 80 & 44 \\ 44 & 80 \end{bmatrix}, \ \mathbf{\Sigma}_2 = \begin{bmatrix} 8000 & 440 \\ 440 & 8000 \end{bmatrix}.$$

Прва главна компонента за $\mathbf{\Sigma}_1$ је $0.707x_1 + 0.707x_2$, а за $\mathbf{\Sigma}_2$ је $0.998x_1 + 0.055x_2$. Видимо да смо променом мерне скале драстично утицали на главне компоненте. Главна компонента матрице $\mathbf{\Sigma}_1$ даје једнаку важност атрибутима X_1 и X_2, док код главна компоненте матрице $\mathbf{\Sigma}_2$ доминира променљива X_1. Дакле, уколико постоје велике разлике у дисперзијама компоненти вектора X, онда ће променљиве са већом дисперзијом доминирати у првој главној компоненту.

Један начин за превазилажење овог проблема јесте да у том случају не користимо директно коефицијенте линеарне комбинације у циљу интерпретације главних компоненти, него да анализу заснивамо на коефицијентима корелације оригиналних променљивих и главних компоненти. Друга могућност је да целу анализу базирамо на корелационој, а не коваријационој матрици оригиналних података.

Сада ћемо одредити коефицијенте корелације између оригиналних променљивих и главних компоненти. Знамо да је $D(\mathbf{Z}) = \mathbf{\Lambda}$ и $D(\mathbf{X}) = \mathbf{\Sigma}$. Коваријанса између $\mathbf{X}$ и $\mathbf{Z}$ је

$$cov(\mathbf{X}, \mathbf{Z}) = cov(\mathbf{X}, \mathbf{AX}) = \mathbf{\Sigma A}^T = (\mathbf{A}^T \mathbf{\Lambda A})\mathbf{A}^T = \mathbf{A}^T \mathbf{\Lambda} = [\mathbf{a_1}\lambda_1, \mathbf{a_2}\lambda_2, ..., \mathbf{a_p}\lambda_p].$$

Коефицијент корелације k-те оригиналне променљиве и j-те главне компоненте дат је следећим изразом

$$\rho_{X_k, Z_j} = \frac{cov(X_k, Z_j)}{\sqrt{D(X_k)}\sqrt{D(Z_j))}} = \frac{\lambda_j a_{jk}}{\sqrt{\sigma_{kk}}\sqrt{\lambda_j}} = a_{jk}\frac{\sqrt{\lambda_j}}{\sqrt{\sigma_{kk}}}, \ j, k = 1, 2, ..., p.$$

У матричном запису, корелациона матрица између вектора оригиналних променљивих и вектора главних компоненти дата је следећим изразом

$$\rho_{\mathbf{XY}} = \mathbf{\Lambda}^{\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}},$$

где је $\mathbf{D}$ дијагонална матрица чији су елементи дисперзије оригиналних променљивих.

Други приступ проблему осетљивости резултата анализе главних компоненти на мерну скалу заснован је на коришћењу корелационе матрице оригиналних променљивих уместо њихове матрице коваријансе. Главне компоненте можемо да дефинишемо на следећи начин

$$\mathbf{Z} = \mathbf{A}^T\mathbf{x}^*$$

где је $\mathbf{A}$ матрица чије колоне представљају сопствени вектори корелационе матрице, а $\mathbf{x}^*$ је стандардизовани вектор, односно i-ти елемент вектора $\mathbf{x}^*$ је $\frac{x_i}{\sqrt{\sigma_{ii}}}$, $i = 1, 2, ..., p$ (подсетимо се да смо раније претпоставили да су променљиве центриране), при чему је x_i i-ти елемент вектора x, а σ_{ii} је дисперзија елемента x_i. Зато корелациону матрицу можемо видети као матрицу коваријансе стандардизованих променљивих. На основу досадашњих резултата, укажимо на основне показатеље анализе у овом специјалном случају. Укупан варијабилитет мерен генерализованом дисперзијом (траг матрице коваријансе) једнак је p, тј. димензији корелационе матрице, а коефицијент корелације између k-те оригиналне променљиве и j-те главне компоненте једнак је $a_{jk}\sqrt{\lambda_j}$. У матричном запису, одговарајућа корелациона матрица дата је изразом $\mathbf{A}\mathbf{\Lambda}^{\frac{1}{2}}$.

Резултати анализе главних компоненти на бази матрице коваријансе, односно корелационе матрице истог скупа података могу се знатно разликовати. Главне компоненте добијене на основу корелационе матрице нису исте нити су директно повезане са главним компонентама рачунатим на основу матрице коваријансе, али поступак добијања компоненти је идентичан у оба случаја. Која од ових матрица ће се користити зависи од тога да ли атрибути варирају у сличним распонима и да ли су сличне природе. Уколико јесу, преферира се матрица коваријансе. Уколико нису, преферира се матрица корелације.

Важно је напоменути колики удео дисперзије оригиналних променљивих је објашњен задржаним скупом главних компоненти. Ова вредност нам показује колико добро главне компоненте апроксимирају дисперзију сваке појединачне оригиналне променљиве. Што је већи удео, то је већи део дисперзије објашњен

сачуваним главним компонентама, што значи да су те компоненте важније у опису варијације података. Ова мера нам помаже да разумемо колико информација је задржано кроз редукцију димензионалности и како се оригинални подаци преносе на главне компоненте. На основу израза за ортогоналну декомпозицију матрице коваријансе имамо да је дисперзија k-те променљиве

$$\sigma_{kk}^2 = \sum_{j=1}^{p} \lambda_j a_{jk}^2 \, , \; k = 1, 2, ..., p \, .$$

Допринос сваке главне компоненте дисперзији k-те променљиве једнак је квадрату коефицијената корелације односне главне компоненте и те оригиналне променљиве. Допринос свих главних компоненти рачунамо на основу корелационе матрице $\mathbf{A}\mathbf{\Lambda}^{\frac{1}{2}}$ тако што саберемо квадрате елемената у њеној k-тој врсти. Уколико смо задржали неколико првих главних компоненти, тад стављањем у однос добијене суме и одговарајуће дисперзије оригиналне променљиве добијамо пропорцију дисперзије те променљиве која је објашњена задржаним главним компонентама. Та пропорција се назива *комуналитет* променљиве и представља проценат објашњења дисперзије оригиналних променљвих задржаним компонентама. Он нам говори колико добро главне компоненте репрезентују варијабилност оригиналне променљиве. Виши комуналитет указује на то да су главне компоненте бољи представници те променљиве и да објашњавају већи део њене варијансе. Комуналитет је користан за разумевање колико информација о дисперзији оригиналне променљиве је сачувано у одабраном скупу главних компоненти. На тај начин можемо проценити колико добро главне компоненте описују и представљају оригиналне променљиве у анализи главних компоненти. Коришћењем корелационе уместо матрице коваријансе оригиналних променљивих одмах добијамо пропорцију дисперзије оригиналне променљиве објашњене задржаним главним компонентама, јер је стандардизацијом променљивих вредност дисперзије једнака јединици.

Дајмо још један начин превазилажења поменутог проблема. Уместо коришћења матрица коваријансе или корелације можемо користити коваријансе од $\frac{x_j}{w_j}$, где су тежине w_j одабране тако да одражавају неку априори идеју о релативном значају променљивих. У специјалном случају када је $w_j = \sqrt{\sigma_{jj}}$ добијамо главне компоненте засноване на корелационој матрици. Различити аутори (видети [11], поглавље 14.2.1) су указивали на то да је избор $w_j = \sqrt{\sigma_{jj}}$ донекле произвољан и да би у неким применама можда биле боље различите

вредности w_j. Међутим, у пракси је ретко да се појави јединствен и адекватан скуп w_j који се природно намеће.

3.6 Узорачке главне компоненте

До сада смо анализу главних компоненти заснивали на популационој коваријансној или корелационој матрици. У овом делу се бавимо анализом главних компоненти добијених из узорачке коваријансне или корелационе матрице. Другим речима, наша анализа се заснива на репрезентативном узорку.

Оцена главних компоненти

Нека је узет узорак од n елемената $\mathbf{x}_1, ..., \mathbf{x}_n$ из p-диомензионе популације са средином μ и матрицом коваријансе $\mathbf{\Sigma}$. Анализу главних компоненти заснивамо на узорачкој матрици коваријансе $\mathbf{S}$ или на узорачкој корелационој матрици $\mathbf{R}$. У оба случаја, метод главних компоненти се користи у дескриптивне сврхе. Међутим, уколико претпоставимо да је популација одакле је узет узорак нормална $\mathcal{N}_p(\mu, \mathbf{\Sigma})$, тада можемо извести бројне резултате који се односе на асимптотска својства главних компоненти.

Оцена максималне веродостојности матрице коваријансе $\mathbf{\Sigma}$, односно $\hat{\mathbf{\Sigma}} = \frac{n-1}{n}\mathbf{S}$, где је $\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T$, представља полазну величину у узорачкој анализи главних компоненти. Уз претпоставку да су сви различити, њене сопствене вредности и сопствени вектори представљају оцене максималне веродостојности одговарајућих популационих сопствених вредности и вектора. Уместо оцене максималне веродостојности матрице коваријнсе $\mathbf{\Sigma}$, можемо користити и непристрасну оцену, тј. узорачку матрицу коваријансе $\mathbf{S}$. Без обзира на то коју процену користимо, главне компоненте су идентичне у оба случаја, као и пропорције објашњене варијансе. Истовремено, $\hat{\mathbf{\Sigma}}$ и $\mathbf{S}$ дају исту узорачку корелациону матрицу $\mathbf{R}$, па избор између оцена постаје небитан.

Нека је $\tilde{\mathbf{z}}_{i1} = \mathbf{a}_1^T\mathbf{x}_i$ за $i = 1, 2, ..., n$ и изаберимо вектор коефицијената $\mathbf{a}_1$, тако да максимизује узорачку дисперзију

$$\frac{1}{n-1}\sum_{i=1}^{n}(\tilde{z}_{i1} - \overline{z}_1)^2, \overline{z}_1 = \frac{1}{n}\sum_{i=1}^{n}\tilde{z}_{i1},$$

уз услов $\mathbf{a}_1^T\mathbf{a_1} = 1$. Затим дефинишемо $\tilde{\mathbf{z}}_{i2} = \mathbf{a}_2^T\mathbf{x}_i$ за $i = 1, 2, ..., n$ и бирамо $\mathbf{a}_2$ тако да максимизује узорачку дисперзију од $\tilde{z}_{i2}$ уз услове да је $\mathbf{a}_2^T\mathbf{a_2} = 1$ и да $\tilde{z}_{i2}$ буде некорелисан са $\tilde{z}_{i1}$ у узорку. Настављајући овај поступак, добијамо узорачку верзију дефиниције главних компоненти дате у одељку 3.2. Дакле, $\mathbf{a}_k^T\mathbf{x}$ се дефинише као k-та узорачка главна компонента, где је $k = 1, 2, ..., p$, а $\tilde{z}_{ik}$ је *скор* за i-то посматрање на k-тој главној компоненти. Скорове главних компоненти користимо у графичком приказу опсервација у дводимензионалном простору генерисаном паровима главних компоненти. У изразу $\mathbf{a}_k^T\mathbf{x}$ смо са $\mathbf{a}_k$ означили оцену вектора коефицијената оригиналних променљивих $\mathbf{x}_1, ..., \mathbf{x}_n$ за k-ту главну компоненту, и ту оцену вектора рачунамо за сваки елеменат узорка.

Пратећи извођење у одељку 3.2, али са узорачким дисперзијама и коваријансама уместо популационих величина, испоставља се да је дисперзија k-те главне компоненте једнаке k-тој највећој сопственој вредности узорачке матрице коваријансе $\mathbf{S}$, а тежине компоненте су одређене одговарајућим сопственим векторима за $k = 1, 2, ..., p$.

Нека је $\tilde{\mathbf{X}}$ матрица димензије $n \times p$ и нека је $[\tilde{\mathbf{X}}]_{ik} = \tilde{x}_{ik}$ при чему је $\tilde{x}_{ik}$ k-ти елемент од $\mathbf{x}_i$. Нека је $\tilde{\mathbf{Z}}$ такође матрица димензије $n \times p$ и $[\tilde{\mathbf{Z}}]_{ik} = \tilde{z}_{ik}$. Тада важи

$$\tilde{\mathbf{Z}} = \tilde{\mathbf{X}}\mathbf{A}$$

при чему је $\mathbf{A}$ ортогонална матрица димензије $p \times p$ чија је k-та колона једнака $\mathbf{a}_k$. За узорачку матрицу коваријансе $\mathbf{S}$ важи:

$$[\mathbf{S}]_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(\tilde{x}_{ij} - \overline{x}_j)(\tilde{x}_{ik} - \overline{x}_k)\,,$$

где је

$$\overline{x}_j = \frac{1}{n}\sum_{i=1}^{n}\tilde{x}_{ij}\,, j = 1, 2, ..., p\,.$$

Стога, матрицу $\mathbf{S}$ можемо записати као

$$\mathbf{S} = \frac{1}{n-1}\mathbf{X}^T\mathbf{X}\,, \tag{3.8}$$

при чему је $\mathbf{X}$ матрица формата $n \times p$ и $[\mathbf{X}]_{ij} = \tilde{x}_{ij} - \overline{x}_j$. Затим, дефинишемо матрицу скорова као

$$\mathbf{Z} = \mathbf{X}\mathbf{A}\,.$$

Важно је напоменути да скорови главних компоненти ове матрице имају једнаке дисперзије и коваријансе као и матрице $\tilde{\mathbf{Z}}$, али су аритметичке средине колона те матрице нула, а не $\overline{z}_k$, за $k = 1, 2, ..., p$.

Наводимо још спектралну декомпозицију узорачке матрице коваријансе

$$\mathbf{S} = l_1 a_1 a_1^T + l_2 a_2 a_2^T + ... + l_p a_p a_p^T \,. \tag{3.9}$$

Према досадашњем излагању следи закључак, да у основи анализе главних компоненти леже управо сопствене вредности и вектори матрице коваријансе или корелационе матрице. Коришћење узорачких оцена одговарајућих популационих величина имплицитно подразумева да ће се оне, због случајних варијација, разликовати од својих популационих пандана. У циљу одређивања интервала поверења и тестирања хипотеза о значајности сопствених вредности наводимо њихова асимптотска понашања, под претпоставком да је случајан узорак од n елемената узет из вишедимензионе нормалне расподеле.

Нека је за матрицу коваријансе $\mathbf{\Sigma}$, са $\mathbf{\Lambda}$ означена дијагонална матрица сопствених вредности $\lambda_1, \lambda_2, ..., \lambda_p$. Имамо асимптотску расподелу узорачких сопствених вредности (видети [11])

$$\sqrt{n}(\hat{\lambda} - \lambda) \sim \mathcal{N}_p(\mathbf{0}, \mathbf{2\Lambda}^2) \,,$$

где смо са $\hat{\lambda}^T = [\hat{\lambda}_1, \hat{\lambda}_2, ..., \hat{\lambda}_p]$ означили оцену вектора сопствених вредности λ засновану на узорачкој матрици коваријансе. Сопствене вредности у великим узорцима су међусобно независни. Ако формирамо матрицу

$$\mathbf{\Omega}_j = \lambda_j \sum_{k=1, k\neq j}^{p} \frac{\lambda_k}{(\lambda_k - \lambda_j)^2} \mathbf{a}_j \mathbf{a}_j^T \,,$$

тада имамо и асимптотску расподелу сопствених вектора

$$\sqrt{n}(\hat{\mathbf{a}}_j - \mathbf{a}_j) \sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Omega}_j) \,,$$

при чему је $\hat{\lambda}_j$ независно распоређено од елемената придруженог сопственог вектора $\hat{\mathbf{a}}_j$. Према наведеним резултатима можемо направити интервал поверења за сваку сопствену вредност с обзиром на њихову расподелу. Важи, $\hat{\lambda}_j \sim \mathcal{N}_p(\lambda_j, \frac{2\lambda_j^2}{n})$ што имплицира асимптотски $100(1-\alpha)\%$ инетрвал поверења за оцену j-те сопствене вредности

$$\frac{\hat{\lambda_j}}{1 + z_{[\frac{\alpha}{2}]}\sqrt{\frac{2}{n}}} \leq \lambda_j \leq \frac{\hat{\lambda_j}}{1 - z_{[\frac{\alpha}{2}]}\sqrt{\frac{2}{n}}} \,,$$

где је $z_{[\frac{\alpha}{2}]}$ горњи $100\frac{\alpha}{2}$ перцентил стандардизоване нормалне расподеле.

Добијени интервал поверења може бити веома широк, чак и за велике вредности n. Из тог разлога, поступак одређивања броја главних компоненти које треба задржати у даљој анализи не можемо заснивати на коришћењу добијене интервалне оцене сопствене вредности.

Декомпозиција матрице на сингуларне вредности

У овом делу ћемо описати познат резултат из линеарне алгебре који је битан и у контексту анализе главних компоненти - декомпозиција матрице на сингуларне вредности.

Нека је $\mathbf{X}$ произвољна матрица димензије $n \times p$. Матрицу $\mathbf{X}$ записати на следећи начин

$$\mathbf{X} = \mathbf{U}\mathbf{L}\mathbf{A}^T, \tag{3.10}$$

где

- $\mathbf{U}$ и $\mathbf{A}$ су матрице формата $n \times r$ и $p \times r$, редом, које имају ортонормиране колоне, па важи $\mathbf{U}^T\mathbf{U} = \mathbf{I}_r$ и $\mathbf{A}^T\mathbf{A} = \mathbf{I}_r$,
- $\mathbf{L}$ је дијагонална матрица формата $r \times r$,
- r је ранг матрице $\mathbf{X}$.

Претпоставимо да имамо реализације $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$ случајног вектора $((X_1^1, X_2^1, ..., X_p^1), ..., (X_1^n, X_2^n, ..., X_p^n))$. Ове податке записујемо у матрицу $\tilde{\mathbf{X}}$ димензије $n \times p$ тако да је $[\tilde{\mathbf{X}}]_{ik} = \tilde{x}_{ik}$, при чему је $\tilde{x}_{ik}$ k-ти елемент од $\mathbf{x}_i$. Доказ постојања спектралне декомпозиције за произвољну матрицу $\mathbf{X}$ димензије $n \times p$ ћемо спровести на матрици чији су елементи $[\mathbf{X}]_{ij} = \tilde{x}_{ij} - \overline{x}_j$.

Из (3.8) и (3.9) знамо да је спектрална декомпозиција матрице $\mathbf{X}^T\mathbf{X}$

$$(n-1)\mathbf{S} = \mathbf{X}^T\mathbf{X} = l_1^\star a_1 a_1^T + l_2^\star a_2 a_2^T + ... + l_p^\star a_p a_p^T,$$

при чему је $l_1^\star = (n-1)l_i$, за $i = 1, 2, ..., p$. Из претпоставке да је ранг матрице $\mathbf{X}$ једнак r следи да је и ранг матрице $\mathbf{X}^T\mathbf{X}$ такође једнак r. Одатле је последњих $p - r$ сопствених вредности тих матрица једнако нули. Према томе важи

$$(n-1)\mathbf{S} = \mathbf{X}^T\mathbf{X} = l_1^\star a_1 a_1^T + l_2^\star a_2 a_2^T + ... + l_r^\star a_r a_r^T.$$

Матрицу $\mathbf{A}$ дефинишемо као матрицу формата $p \times r$ чија је k-та колона

$$\mathbf{a}_k^\star = (n-1)a_k \, , k = 1, 2, ..., r$$

то јест $\mathbf{a}_k^\star$ је сопствени вектор који одговара сопственој вредности $l_k^\star$. Матрицу $\mathbf{U}$ дефинишемо као матрицу формата $n \times r$ чија је k-та колона

$$\mathbf{u}_k = l_k^{\star -\frac{1}{2}} \mathbf{X}\mathbf{a}_k^\star \, , k = 1, 2, ..., r \, .$$

Матрицу $\mathbf{L}$ дефинишемо као дијагоналну матрицу формата $r \times r$ чији је k-ти дијагонални елемент $l_k^{\star -\frac{1}{2}}$. Овако изабране матрице задовољавају прва два услова, па сада можемо показати да важи $\mathbf{X} = \mathbf{ULA}^T$. Дакле, имамо:

$$\begin{aligned} \mathbf{ULA}^T &= \mathbf{U} \begin{bmatrix} l_1^{\star \frac{1}{2}} \mathbf{a}_1^{\star T} \\ l_2^{\star \frac{1}{2}} \mathbf{a}_2^{\star T} \\ \vdots \\ l_r^{\star \frac{1}{2}} \mathbf{a}_r^{* T} \end{bmatrix} = \sum_{k=1}^{r} l_k^{\star -\frac{1}{2}} \mathbf{X}\mathbf{a}_k^{\star \frac{1}{2}} l_k^{\star \frac{1}{2}} \mathbf{a}_k^{* T} \\ &= \sum_{k=1}^{r} \mathbf{X}\mathbf{a}_k^* \mathbf{a}_k^{* T} = \sum_{k=1}^{p} \mathbf{X}\mathbf{a}_k^* \mathbf{a}_k^{* T} \, . \end{aligned}$$

Последња једнакост важи јер је $\mathbf{a}_k^*$ за $k = r+1, r+2, ..., p$ сопствени вектор матрице $\mathbf{X}^T\mathbf{X}$ који одговара сопственој вредности која је једнака нули. Вектор $\mathbf{X}\mathbf{a}_k^*$ је вектор скорова k-те главне компоненте, док дисперзија једнака нули последњих $p - r$ главних компоненти имплицира да је $\mathbf{X}\mathbf{a}_k^* = 0$ за $k = r + 1, r + 2, ..., p$. Према томе, важи

$$\mathbf{ULA}^T = \mathbf{X} \sum_{k=1}^{p} \mathbf{a}_k^* \mathbf{a}_k^{* T} = \mathbf{X} \, .$$

Декомпозиција матрице на сопствене вредности даје ефикасан начин за израчунавање главних компоненти. Јасно је да проналаском матрица $\mathbf{U}, \mathbf{L}$ и $\mathbf{A}$ које задовољавају једнакост (3.6) долазимо до сопствених вектора (из матрице $\mathbf{A}$) и корена сопствених вредности (из матрице $\mathbf{L}$) матрице $\mathbf{X}^T\mathbf{X}$, а тиме и до коефицијената и стандардних девијација главних компоненти за узорачку матрицу коваријансе $\mathbf{S}$. Матрица $\mathbf{U}$ садржи скалиране скорове. Ако једнакост (3.6) помножимо са десне стране матрицом $\mathbf{A}$, добијамо

$$\mathbf{XA} = \mathbf{ULA}^T\mathbf{A} = \mathbf{UL} \, ,$$

јер је $\mathbf{A}^T\mathbf{A} = \mathbf{I}_r$. Матрица $\mathbf{XA}$ је формата $n \times r$ и њена k-та колона садржи скорове k-те главне компоненте. Скорови су дати формулом

$$z_{ik} = u_{ik} l_k^{\frac{1}{2}} \text{ за } i = 1, ..., n,\ k = 1, ..., r$$

или у матричном запису $\mathbf{Z} = \mathbf{UL}$.

Поред наведеног, декомпозиција матрице нам даје додатни увид у то како главне компоненте раде. Даје нам алгебарски и графички начин за представљање и тумачење главних компоненти.

3.7 Тестирање значајности главних компоненти

Анализа главних компоненти представља метод за смањење димензије података и као таква није заснована на теоријском моделу. У поступку смањења димензије није нам унапред наглашено са колико главних компонената је потребно извршити анализу да бисмо обухватили значајан део укупне дисперзије. Из досадашњег излагања можемо закључити да нас у анализи интересују главне компоненте које имају највеће сопствене вредности. Међутим, то не значи да су аналитички мање интересантне главне компоненте са мањим сопственим вредностима. Оне су корисне у процесу утврђивања одступања од претпоставке нормалности. Због свега тога, можемо говорити и о тестирању значајности главних компоненти у виду тестова сопствених вредности. Ово је још један од начина одабира главних компоненти које су битне за анализу, поред одабира према величини њихових сопствених вредности.

Најпознатији тест за сопствене вредности матрице коваријансе приписује се Бартлету и користи се за тестирање хипотезе да су последње сопствене вредносте међусобно једнаке. Дакле, тестирамо

$$H_{0k} : \lambda_{k+1} = \lambda_{k+2} = ... = \lambda_{k+p}$$

против алтернативне хипотезе да су бар две од последњих $p - k$ сопствених вредности међусобом различите. У свом оригиналном раду, Хотелинг је проучавао проблем тестирања једнакости две узастопне сопствене вредности, а тестови за H_{0k} су касније разматрани од стране неколико аутора, укључујући и Бартлета. Одлука о томе колико главних компоненти треба задржати

може се донети тестирањем нулте хипотезе за различите вредности k. Ако прихватимо нулту хипотезу, тада у анализи користимо само првих k главних компоненти јер за њих претпостављамо да обухватају значајан део укупног варијабилитета, а да последњих $p-k$ главних компоненти једнаког варијабилита мере само "шум" у подацима.

Тест статистика за тестирање наведене хипотезе конструисана је уз претпоставку о нормалности, а заснована је на коришћењу принципа количника највеће варијабилности (енг. *Likelihood Ratio test*)

$$LR = \left(\frac{\prod_{j=k+1}^{p} \hat{\lambda}_j}{\left[\frac{1}{p-k} \sum_{j=k+1}^{p} \hat{\lambda}_j \right]^{p-k}} \right)^{\frac{n}{2}},$$

где су $\hat{\lambda}_j$ сопствене вредности узорачке матрице коваријансе. Из израза за тест статистику можемо закључити да је тест заснован на поређењу геометријске и аритметичке средине последњих $p-k$ сопствених вредности. Ако је нулта хипотеза тачна, тада LR статистика има вредност једнаку нули. У супротном, удаљавајући се од нулте хипотезе разлика између аритметичке и геометријске средине ће се повећавати, што резултира мањом вредношћу LR статистике. У том случају одбацујемо нулту хипотезу о једнакости последњих $p-k$ сопствених вредности. Тачна расподела тест статистике је компликована, али можемо користити познате резултате из статистичког заккатеључивања који се односе на LR тестове. Наиме, $-2\ln LR$ има приближно χ^2 рсподелу код које је број степени слободе једнак разлици између броја независно променљивих параметара под $H_{0k} \cup H_{1k}$ и под H_{0k}. Израчунавање броја степени слободе није једноставно, али испоставља се да је $\nu = \frac{1}{2}(p-k+2)(p-k-1)$. Тада, под нултом хипотезом, важи:

$$\left((p-k) \ln \overline{\lambda}_{p-k} - \sum_{j=k+1}^{p} \ln \hat{\lambda}_j \right) \sim \chi^2_\nu ,$$

где је $\overline{\lambda}_{p-k}$ аритметичка средина последњих $p-k$ сопствених вредности узорачке матрице коваријансе. Апроксимација се може побољшати ако n заменимо са $n' = n - \frac{1}{6}(2p+11)$, тако да се H_{0k} одбацује са нивоом значајности α ако

$$\left(n - \frac{1}{6}(2p+11) \right) \left((p-k) \ln \overline{\lambda}_{p-k} - \sum_{j=k+1}^{p} \ln \hat{\lambda}_j \right) \geq \chi^2_{\nu,\alpha} .$$

Практичан поступак примене овог теста се може описати на следећи начин. Прво се тестира хипотеза да су све сопствене вредности међусобно једнаке, $k = 1$. Уколико се одбаци ова хипотеза, поставља се нова хипотеза да су све сопствене вредности, осим прве, међусобно једнаке, $k = 2$. Ако и ова хипотеза буде одбачена, поступак тестирања се наставља, али сада тестирамо нулту хипотезу да су све сопствене вредности, осим прве две, међусобно једнаке, $k = 3$. Поступак се наставља све док се не прихвати хипотеза о једнакости последњих $p - k$ сопствених вредности.

Тестирање значајности последњих сопствених вредности нам може дати одговор на питање о броју главних компоненти које ћемо задржати у анализи. Мањкавост овог приступа огледа се у задржавању сувише великог броја главних компоненти.

3.8 Избор броја главних компоненти

Као што смо већ рекли, основни циљ методе главних компоненти јесте смањење димензије података, односно смањење броја променљивих. Оно што желимо јесте да након примене методе имамо мање компоненти него што имамо оригиналних променљивих, наравно уз услов да ове компоненте објасне што више варијације у подацима. Природно се намеће питање, како одлучити колико главних компоненти треба задржати? При одабиру броја главних компоненти могу се користити различити критеријуми, а овде ћемо приказати неколико уобичајених.

Први приступ се ослања на фиксирање кумулативне пропорције укупне дисперзије коју објашњава одабрани скуп главних компоненти. У зависности од посматраног проблема, одабере се кумулативна пропорција, на пример 80% или 90% укупне дисперзије, и затим се повећава број задржаних главних компоненти све док се не постигне претходно постављена граница. Важно је напоменути да ова метода има субјективну природу јер се број главних компоненти одређује на основу произвољно фиксиране вредности кумулативне пропорције објашњене дисперзије. То значи да се граница одређује према процени или приоритетима истраживача, а не на основу објективних критеријума.

Други приступ сугерише задржавање оних главних компоненти чија је дисперзија, λ_k, већа од просечне вредности $\overline{\lambda} = \frac{1}{p}\sum_{k=1}^{p} \lambda_k$. Ако уместо матрице

коваријансе користимо матрицу корелеције, тада је просечна вредност дисперзије једнака 1. У том случају овај критеријум нам каже: задржи оне главне компоненте код којих је дисперзија већа од јединице. Овај критеријум се често користи у факторској анализи, где је познат под називом "критеријум јединичног корена" или "Кајзерово правило" (енг. *Kaiser's rule*).

Следећи критеријум избора користи геометријску средину сопствених вредности, односно дисперзија главних компоненти. Генерализована дисперзија једнака је производу сопствених вредности, то јест $\prod_{j=1}^{p} \lambda_j$. Ако добијену вредност дигнемо на степен $\frac{1}{p}$ добићемо геометријску средину сопствених вредности. Дакле, просечна генерализована дисперзија дата је геометријском средином сопствених вредности, па према овом критеријуму задржавамо оне главне компоненте чија је дисперзија већа од геометријске средине свих сопствених вредности.

Још ћемо размотрити критеријум који се ослања на дијаграм осипања (енг. *Scree graph*), односно на графички приказ сопствених вредности у опадајућем поретку, а предложио га је Кател 1966. године у раду [7]. Формира се тако што се сопствене вредности прикажу на графикону у односу на њихове одговарајуће бројеве компоненти и пружа визуелну процену колико дисперзије објашњава свака главна компонента.

Графикон приказује криву која опада, при чему првих неколико главних компоненти има значајно веће сопствене вредности од осталих. Тачка где крива почиње да се равна или постаје релативно равна означава тачку на којој укључивање додатних компоненти неће значајно допринети објашњеној дисперзији. Ова тачка се често користи као праг или граница за одабир броја задржаних главних компоненти. Правило лакта није од помоћи уколико на графикону нема очигледног прелома или уколико их има више од једног.

Важно је напоменути да избор броја главних компоненти зависи од специфичних потреба анализе, жељене димензије података и баланса између смањења димензије и губитка информација. У општем случају показало се да када је $p \geq 20$, Кајзеров критеријум је рестриктиван у смислу да укључује сувише мали број главних компоненти. Насупрот њему, правило лакта задржава велики број главних компоненти у даљој анализи. Пожељно је експериментисати са различитим праговима и критеријумима како би се пронашао најбољи избор за конкретан случај.

3.9 Ротација главних компоненти

Ротација задржаних главних компоненти је поступак којим се врши промена оријентације или међусобног положаја главних компоненти ради лакше интерпретације. Постоје два основна типа ротације: ортогонална (енг. *orthogonal*), када нове осе остају међусобно ортогоналне, и коса (енг. *oblique*), када нове осе нису обавезно ортогоналне. Циљ ротације је постићи „једноставнију" структуру задржаних компоненти која је лакша за интерпретацију. Ротација се обично изводи у простору задржаних компоненти и може утицати на расподелу објашњене дисперзије између компоненти. Избор потпростора за ротацију има велики утицај на резултате, па је препоручљиво испробати више варијанти потпростора како би се проверила робустност интерпретације.

Ортогонална ротација главних компоненти

Ортогонална ротација се дефинише матрицом ротације $\mathbf{W}$, где редови означавају оригиналне факторе, а колоне нове (ротиране) факторе. На пресеку реда m и колоне n имамо косинус угла између првобитне и нове осе: $w_{m,n} = \cos\theta_{m,n}$. Матрица ротације је ортогонална и стога важи $\mathbf{W}^T\mathbf{W} = \mathbf{I}$.

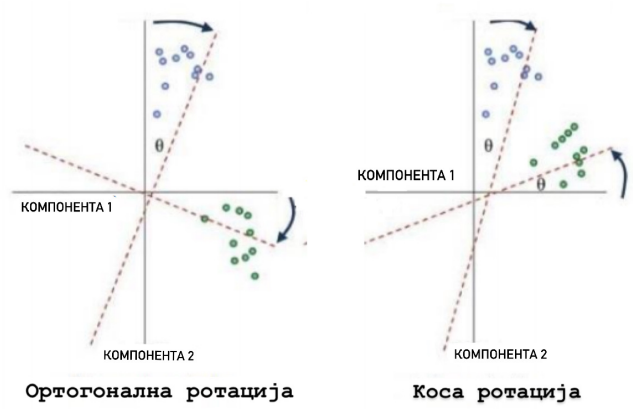
Varimax ротација, коју је развио Кајзер 1958. године (погледати [14]), је најпопуларнија метода ротације. Циљ *varimax* ротације је постизање једноставног обрасца оптерећења компоненти, где свака компонента има висока оптерећења само за одређен број променљивих, док остале променљиве имају нула или мала оптерећења. Ова ротација нам помаже да идентификујемо јасне везе између променљивих и компоненти, олакшавајући интерпретацију резултата. Одређивање ортогоналне матрице се врши максимизовањем суме дисперзије квадрата коефицијената. У већини статистичких софтвера *varimax* је подешен као подразумевани критеријум.

Коса ротација главних компоненти

За разлику од метода ортогоналне ротације као што је *varimax*, *promax* омогућава косо (неортогонално) ротирање, што значи да ротиране компоненте могу бити у корелацији једна са другом.

Promax ротација укључује процену матрице која дефинише трансформацију оригиналних главних компоненти. Ова матрица трансформације укључу-

је углове ротације и коефицијенте корелације између ротираних компоненти. Углови ротације одређују оријентацију компоненти, док коефицијенти корелације обухватају међусобне односе између компоненти. Ротација се обично врши итеративно, прилагођавајући вредности како би се побољшало слагање између ротираних компоненти и посматраних података. Омогућавајући корелацију између ротираних компоненти, *promax* ротација може боље ухватити основну сложеност и међузависности у подацима. Посебно је корисна када се очекује да су димензије које се анализирају међусобно корелисане, јер пружа реалистичнију репрезентацију односа између варијабли.



Слика 3.3: Ортогонална и коса ротација главних компоненти

3.10 Примена методе главних компоненти

Све методе које су наведене и описане кроз рад су илустроване кроз примере помоћу програмског језика *Python*. Ови примери демонстрирају различите концепте и технике и могу се пронаћи на следећем *GitHub* репозиторијуму [2].

База података *Iris*

У овом примеру, упознаћемо се са методом главних компоненти на популарној бази података *Iris*. База података *Iris* је често коришћен скуп података у области машинског учења и статистике. Она садржи информације о различитим врстама цветова. База се састоји од 150 инстанци, при чему свака инстанца има 4 нумеричка атрибута, то су: дужина и ширина латица (енг.

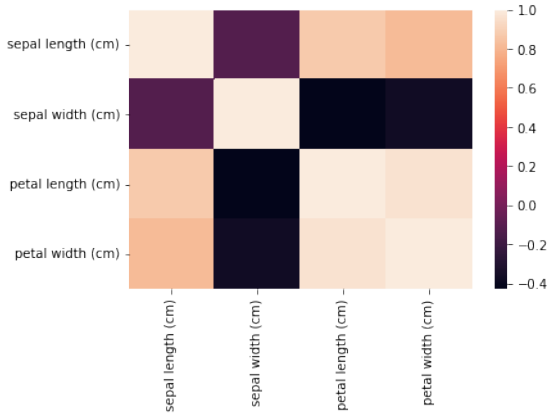
petal) и дужина и ширина чашице цвета (енг. *sepal*). Цветови су подељени у три класе *Setosa, Versicolor* и *Virginica*, при чему свака класа има 50 инстанци.

Након увоза свих потребних библиотека, прво учитавамо базу података *Iris* која већ постоји у *sklearn* библиотеци. Затим конвертујемо базу података у *pandas data frame* ради лакше манипулације. Пре примене анализе глав-

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
0	5.1	3.5	1.4	0.2
1	4.9	3.0	1.4	0.2
2	4.7	3.2	1.3	0.2
3	4.6	3.1	1.5	0.2
4	5.0	3.6	1.4	0.2

Слика 3.4: Првих неколико редова из базе података *Iris*

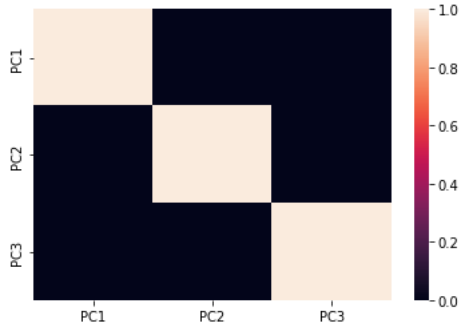
них компоненти или других техника статистичког или машинског учења, добра пракса је стандардизација података. Користимо *StandardScaler* из *sklearn* библиотеке како бисмо стандардизовали скуп атрибута. Затим приказујемо корелацију између скалираних података помоћу топлотне мапе. Корелација између различитих атрибута је приказана путем боја, где тамнија нијанса означава мању корелацију, а светлија нијанса означава већу корелацију. Ова анализа нам помаже да визуелно сагледамо међусобне односе атрибута пре примене даљих техника анализе података.



Слика 3.5: Топлотна мапа корелације скупа података *Iris*

Са топлотне мапе приказане на слици 3.5 видимо да постоје високе корелације између дужине чашице и дужине латица, као и између дужине

латица и ширине латица. Дакле, има смисла смањити димензију података. Метод главних компоненти примењујемо на скалирани скуп података. За то *Python* нуди још једну уграђену класу названу *PCA* која се налази у модулу *sklearn.decomposition*. Ради илустрације, број компоненти у нашем финалном скупу смо поставили на 3, што уствари значи да ће наш крајњи скуп атрибута имати 3 колоне. Сада, када смо применили метод и добили смањени скуп атрибута, проверићемо корелацију између различитих главних компоненти, поново користећи топлотну мапу.



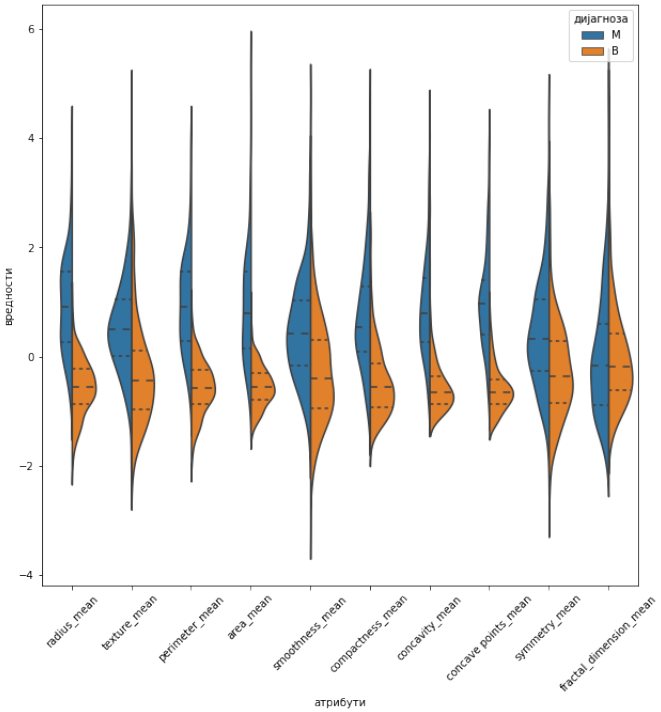
Слика 3.6: Топлотна мапа након примене анализе главних компоненти

Са топлотне мапе приказане на слици 3.6 јасно видимо да не постоји корелација између различитих добијених главних компоненти (*PC1, PC2* и *PC3*). Дакле, прешли смо са простора више димензије на простор мање димензија, при чему смо осигурали да је корелација између добијених главних компоненти минимална. На тај начин смо остварили циљеве методе главних компоненти.

База података *Breast Cancer*

Ова база података је део библиотеке *scikit-learn* и често се користи у едукативне сврхе и за тестирање алгоритама класификације. Матрица атрибута садржи податке о туморима дојке. Сваки ред у матрици представља један узорак (пацијента), а сваки атрибут представља различите карактеристике тумора (нпр. радијус, текстура, обим, глаткоћа итд.). Укупно има 30 атрибута. Такође имамо и низ циљних вредности који означава дијагнозу тумора. Циљна вредност ”0” представља бенигни тумор, док циљна вредност ”1” представља малигни тумор.

Визуализација података. Да бисмо визуализовали податке, користићемо *seaborn* графиконе. Пре употребе графикона виолине (*violin plot*), потребно је нормализовати или стандардизовати податке. Разлике између вредности атрибута су велике да би се могле уочити на графикону. Графиконе групишемо у 3 групе, при чему свака група садржи 10 атрибута како бисмо их лакше посматрали. Не заборавимо, не бирамо атрибуте, већ се упознајемо са подацима, као када гледамо карту пића на вратима паба.

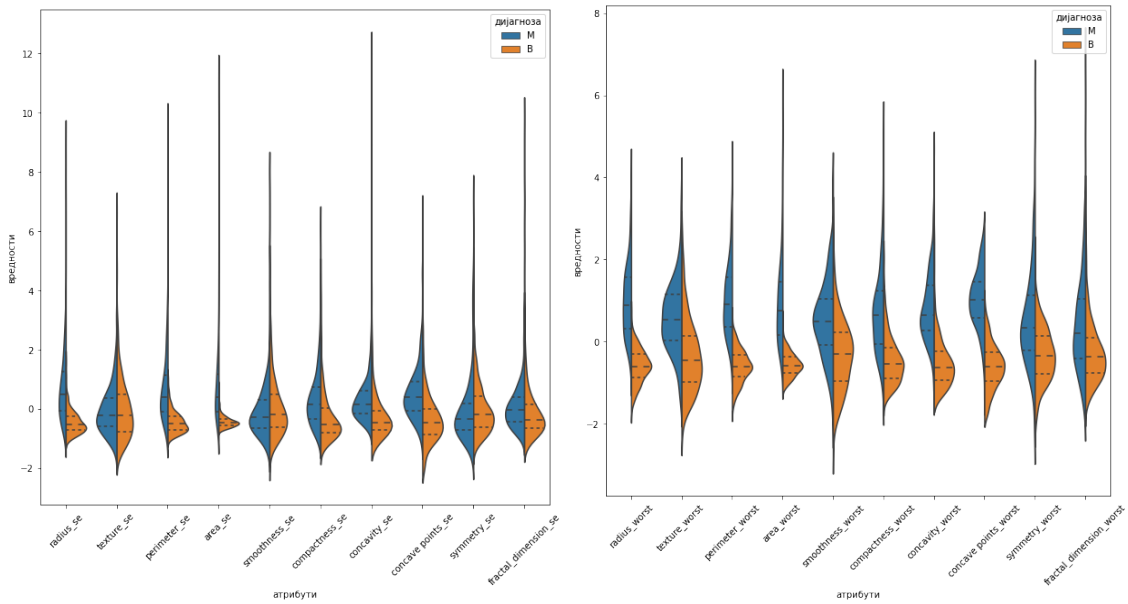


Слика 3.7: Графикон виолине првих 10 атрибута

Протумачимо графикон приказан на слици 3.7. На примеру атрибута *texture_ mean*, медијана малигних и бенигних тумора изгледа као да су раздвојени, што може бити добро за класификацију. Међутим, на примеру атрибута *fractal_ dimension_ mean* медијана малигних и бенигних тумора не изгледа раздвојено, што не пружа довољно информација за класификацију. Ово тумачење значи да атрибут *texture_ mean* може бити користан у процесу класификације јер раздваја медијану између малигних и бенигних тумора, док атрибут *fractal_ dimension_ mean* не пружа довољно информација за разликовање између ове две класе.

Са слике 3.8, примећујемо још да атрибути *concavity_ worst* и *concave*

point_ worst изгледају слично, али како можемо утврдити да ли су они корелирани или не?

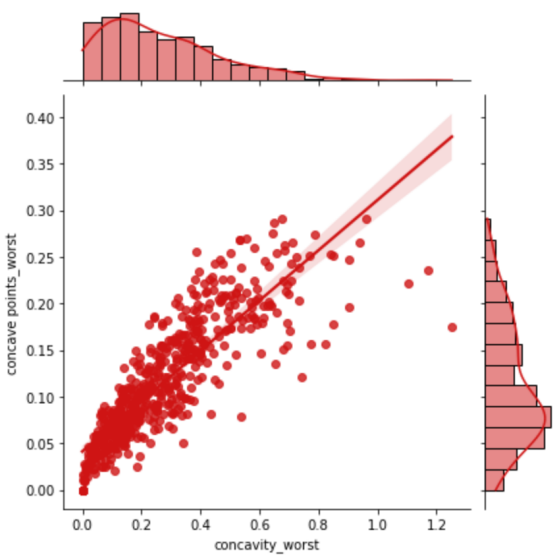


Слика 3.8: Графикон виолине

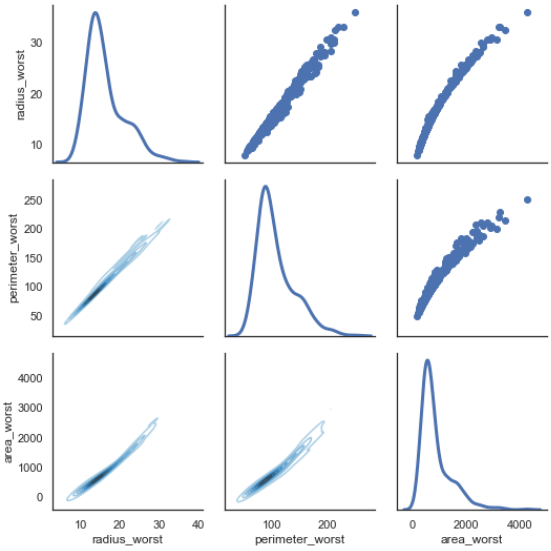
Да бисмо детаљније упоредили ова два атрибута, користићемо заједнички графикон (енг. *joint plot*). Ако погледамо слику 3.9, видимо да су они заиста корелирани. Вредност *Pearsonr* је мера корелације, при чему вредност 1 представља највишу корелацију. Стога, вредност од 0.86 је довољна да кажемо да су ови атрибути корелирани. Још увек не бирамо атрибуте, само покушавамо да стекнемо слику о њима.

Шта је са поређењем три или више карактеристика? У ту сврху можемо користити мрежу парова (енг. *pair grid plot*). Са слике 3.10 откривамо још једну ствар, *radius_ worst*, *perimeter_ worst* и *area_ worst* су у корелацији. Ако желимо да посматрамо корелације између свих атрибута, користимо топлотну мапу која је стара, али моћна метода цртања.

Након анализе коју смо спровели, приметили смо да постоји корелација између одређених атрибута у скупу података. То сугерише да ови атрибути пружају сличне информације и да су повезани. Када имамо високо корелиране атрибуте, може се размотрити смањење димензије скупа података. Смањење димензије може бити корисно јер нам омогућава да радимо са мањим бројем атрибута који задржавају већину информација садржаних у оригиналном скупу података. У ту сврхе користимо анализу главних компоненти, како би-



Слика 3.9: Заједнички графикон атрибута *concavity_worst* и *concave points_worst*



Слика 3.10: Мрежа парова

смо добили компактнији скуп атрибута који одржава већину варијабилности у подацима.

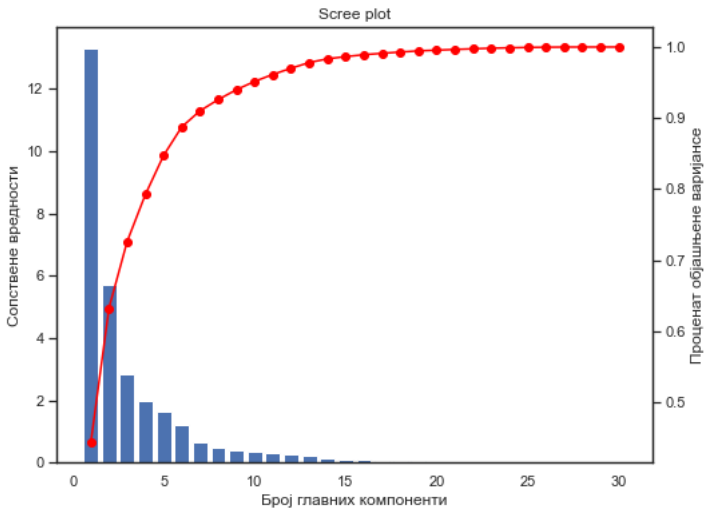
Одређивање броја главних компоненти. Главне компоненте можемо изабрати уз помоћ различитих критеријума о којима је већ било речи. Према Кајзеровом правилу, треба задржати само оне компоненте чија је сопствена вредност већа од 1. Како је то случај код првих шест главних компоненти, према овом критеријуму њих бисмо задржали у даљој анализи.

Кумулативна пропорција објашњене дисперзије износи редом: за прву главну компоненту 44.27%, за прве две главне компоненте 63.24%, за прве три 72.63%, за прве четири 79.23% , за првих пет 84.73% и тако даље. Ако унапред фиксирамо износ објашњене дисперзије на 80%, онда према овом критеријуму треба задржати пет главних компоненти.

Аритметичка средина сопствених вредност износи $\overline{\lambda} = 1.00176$, па према другом критеријуму задржавамо оне главне компоненте код којих је $\hat{\lambda} > \overline{\lambda}$. Како је то случај код првих шест главних компоненти, према овом критеријуму толико ћемо их и задржати у даљој анализи.

Критеријум заснован на геометријској средини сугерише задржавање четрнаест главних компоненти, пошто је геометријска средина једнака 0.095, а првих четрнаест сопствених вредности су веће од ове вредности. По мом

мишљењу, нема потребе задржати толики број компоненти, а и остали критеријуми сугеришу задржавање мањег броја компоненти.



Слика 3.11: Дијаграм осипања

На крају, коришћење правила лакта заснивамо на слици 3.11. Са слике читамо да за потребе даље анализи треба задржати шест главних компоненти.

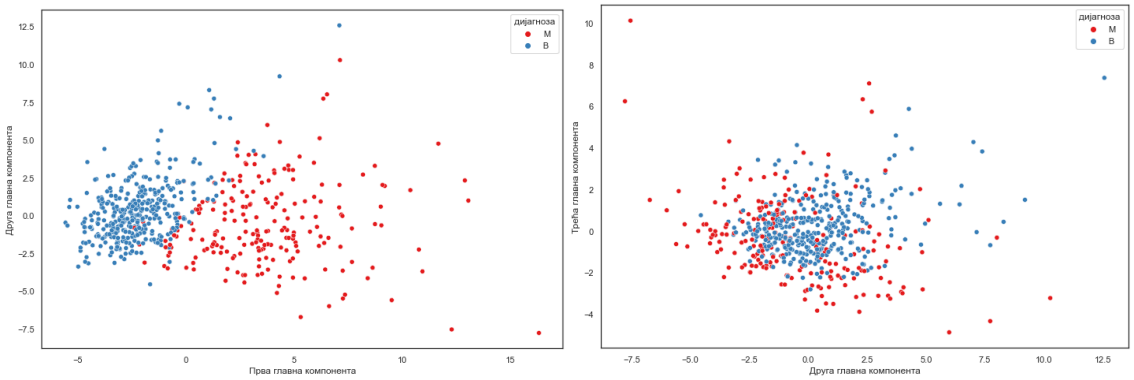
Коришћењем различитих критеријума избора броја главних компоненти у даљој анализи бисмо задржали шест главних компоненти, чиме смо значајно смањили димензију подтака.

Анализа компоненти. Изабрали смо две компоненте са највећим сопственим вредностима и приказали дводимензиони графикон који приказује разлику између малигног и бенигног тумора на основу ове две компоненте. Такође, можемо експериментисати са различитим скуповима компоненти како бисмо видели који пар даје бољу визуализацију ове две групе пацијената.

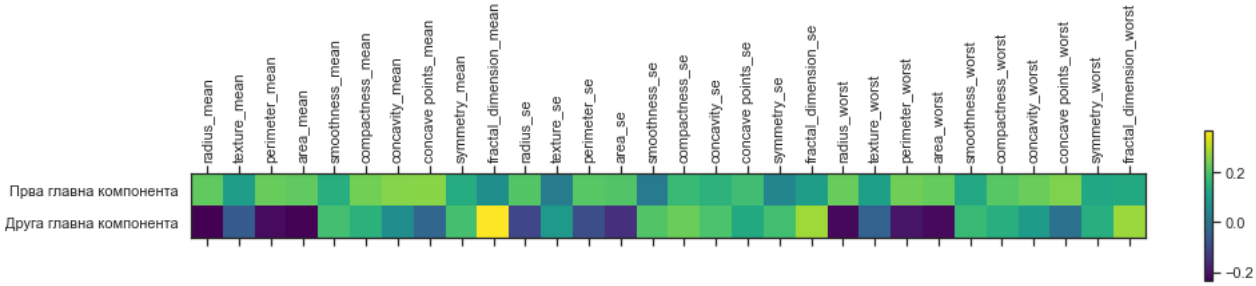
Са слике 3.12 се јасно види да прва пројекција боље раздваја класе, што је донекле и очекивано јер прва компонента садржи највећи део информација.

Матрица *pca_components_* садржи информације о међусобном односу сваког од полазних атрибута и добијених главних компоненти. Ово можемо искористити за приказивање зависности у форми топлотне мапе (слика 3.13). Први ред се односи на прву, а други на другу главну компоненту. Сваки од атрибута је описан појединачном колоном.

Ротација главних компоненти. Циљ ротације јесте добијање једноставне структуре у којој главне компоненте треба да буду што независније. То зна-



Слика 3.12: Разлика између малигног и бенигног тумора на основу две главне компоненте



Слика 3.13: Приказ доприноса сваког атрибута првој и другој главној компоненти

чи да свака главна компонента треба да буде одређена специфичним скупом променљивих, а да при томе буде што мање променљивих које се заједнички јављају у већем броју главних компоненти.

Приликом ротације се мењају оптерећења и требало би да та оптерећења дају јаснију интерпретацију. Вредности у табели, слика 3.15, су оптерећења која представљају корелације између главних компоненти и оригиналних променљивих. Из табеле посматрамо само она оптерећења која су по апсолутној вредности већа од 0.8. Као што можемо приметити, прва главна компонента је у корелацији са *mean_ radius, mean_ perimeter, mean_ area, mean_ concave_ points, area_ error, worst_ radius, worst_ perimeter* и *worst_ area*. Такође, можемо рећи да прва главна компонента у великој мери објашњава и *worst_ concav_ points*. Друга главна компонента је у корелацији са *mean_ fractal_ dimension* и *worst_ fractal_ dimension*, као и да у великој мери овјашњава *compactness_ error*. Ако погледамо табелу са слике 3.14, видимо

да највећа оптерећења (по апсолутној вредности) поново стоје уз променљиве *mean_fractal_dimension* и *worst_fractal_dimension*.

	PC1	PC2
mean radius	0.218902	-0.233857
mean texture	0.103725	-0.059706
mean perimeter	0.227537	-0.215181
mean area	0.220995	-0.231077
mean smoothness	0.142590	0.186113
mean compactness	0.239285	0.151892
mean concavity	0.258400	0.060165
mean concave points	0.260854	-0.034768
mean symmetry	0.138167	0.190349
mean fractal dimension	0.064363	0.366575
radius error	0.205979	-0.105552
texture error	0.017428	0.089980
perimeter error	0.211326	-0.089457
area error	0.202870	-0.152293
smoothness error	0.014531	0.204430
compactness error	0.170393	0.232716
concavity error	0.153590	0.197207
concave points error	0.183417	0.130322
symmetry error	0.042498	0.183848
fractal dimension error	0.102568	0.280092
worst radius	0.227997	-0.219866
worst texture	0.104469	-0.045467
worst perimeter	0.236640	-0.199878
worst area	0.224871	-0.219352
worst smoothness	0.127953	0.172304
worst compactness	0.210096	0.143593
worst concavity	0.228768	0.097964
worst concave points	0.250886	-0.008257
worst symmetry	0.122905	0.141883
worst fractal dimension	0.131784	0.275339

Слика 3.14: Оптерећења главних компоненти

	PC1_rotated	PC2_rotated
mean radius	0.970997	-0.069656
mean texture	0.397554	0.071648
mean perimeter	0.975154	-0.015268
mean area	0.974140	-0.060050
mean smoothness	0.218300	0.647705
mean compactness	0.562677	0.758401
mean concavity	0.734731	0.606297
mean concave points	0.858586	0.416476
mean symmetry	0.199281	0.648111
mean fractal dimension	-0.247270	0.871016
radius error	0.773546	0.168930
texture error	-0.055598	0.216846
perimeter error	0.770577	0.211887
area error	0.821021	0.067401
smoothness error	-0.204726	0.445807
compactness error	0.248246	0.795120
concavity error	0.239135	0.690989
concave points error	0.414301	0.609782
symmetry error	-0.092047	0.455943
fractal dimension error	-0.021913	0.765337
worst radius	0.982325	-0.024003
worst texture	0.382458	0.102199
worst perimeter	0.984901	0.033087
worst area	0.971915	-0.028794
worst smoothness	0.189410	0.592063
worst compactness	0.481518	0.686836
worst concavity	0.595771	0.628303
worst concave points	0.794960	0.452130
worst symmetry	0.210847	0.520328
worst fractal dimension	0.075300	0.810224

Слика 3.15: Оптерећења главних компоненти након ротације

Глава 4

Анализа главних компоненти са кернел трансформацијама

Стандардна метода главних компоненти омогућава само линеарно смањење димензије. Међутим, уколико подаци имају сложеније структуре које се не могу добро представити у линеарном потпростору, стандардна метода главних компоненти неће бити од велике помоћи. Анализа главних компоненти са кернел трансформацијама нам омогућава генерализацију стандардне методе главних компоненти за нелинеарно смањење димензије. Предности методе са кернел трансформацијама су нарочито видљиве код анализе података који се не могу адекватно моделовати линеарним моделима, као што су слике, аудио записи и друге комплексне врсте података. Ова техника омогућава боље очување информација о структури података, што је кључно у многим применама, укључујући препознавање облика, класификацију и друге анализе података. Међутим, ова метода има и нека ограничења, као што је одабир одговарајућег кернела (функције језгра) и одговарајућих параметара, што може бити тешко и дуготрајно. Такође може бити рачунски захтевна за велике скупове података, јер захтева израчунавање кернел матрице за све парове података.

Напомена: Иако се у различитим математичким дисциплинама термин кернел преводи као језгро, ми ћемо у даљем раду, у духу терминологије машинског учења, користити енглеску реч. У различитим контекситма, праве се врло различите претпоставке везане за кернеле, па се и сами концепти који се под тим изразом подразумевају врло разликују. Овај термин означава функцију која се користи за рачунање сличности између података у вишедимензионом простору.

4.1 Кернели (функције језгра)

У контексту статистичког учења, кернел се дефинише као позитивно семидефинитна, симетрична функција $\kappa\colon \mathbf{X} \times \mathbf{X} \to \mathbf{R}$ која задовољава Мерцеров услов.

Мерцеров услов каже да за било који скуп тачака $x_1, x_2, ..., x_n$ у простору улаза $\mathbf{X}$, матрица $\mathbf{K}$ са елементима $\mathbf{K}_{ij} = \kappa(x_i, x_j)$ мора бити позитивно семидефинитна. Другим речима, за било који коначан скуп података, матрица кернела треба да има све сопствене вредности веће или једнаке нули.

Својство позитивне семидеифинитности обезбеђује да кернел представља валидан скаларни производ (енг. *inner product*) у потенцијалном високодимензионом простору атрибута. То нам омогућава да имплицитно изводимо прорачуне у том вишедимензионом простору, без експлицитног мапирања тачка података у њега, што је познато као кернел трик (енг. *„kernel trick"*)[1]. Овај трик омогућава ефикасно рачунање у алгоритмима као што су анализа главних компоненти са кернел трансформацијама, метод потпорних вектора (енг. *support vector machine*) и другим методама заснованим на кернелима.

Најчешће коришћени кернели су:

- константни кернел (енг. *constant kernel*)

$$\kappa(\mathbf{x}, \mathbf{y}) = c\,,$$

- линеарни кернел (енг. *linear kernel*)

$$\kappa(\mathbf{x}, \mathbf{y}) = \mathbf{x} \cdot \mathbf{y}\,,$$

- полиномни кернел (енг. *polinomyal kernel*)

$$\kappa(\mathbf{x}, \mathbf{y}) = (\mathbf{x} \cdot \mathbf{y} + c)^d\,,$$

  где је d параметар који одређује степен полинома, а параметар c је слободан члан који омогућава померање полинома дуж вертикалне осе

[1]То значи да приликом примене технике кернел трик и трансформације података у вишедимензиони простор, не морамо дословно израчунати све вредности тих трансформисаних тачака и стварно их преместити у тај простор. У традиционалном приступу, ако бисмо желели да радимо са подацима у вишедимензионом простору, прво бисмо их трансформисали користећи одређену функцију пресликавања и тиме добили нове координате за сваку тачку у новом простору. Међутим, то може бити рачунски захтевно, посебно када радимо са великим скупом података. Кернел трик нам омогућава да радимо са подацима као да су већ трансформисани у вишедимензионом простору.

- RBF (*Radial Basis Function*) кернел, такође познат као Гаусов кернел

$$\kappa(\mathbf{x}, \mathbf{y}) = e^{-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}},$$

  где $\|\cdot\|$ представља Еуклидско растојање између вектора $\mathbf{x}$ и $\mathbf{y}$, а σ је слободни параметар који контролише ширину функције

- сигмоидни кернел (енг. *sigmoid kernel*)

$$\kappa(\mathbf{x}, \mathbf{y}) = \tanh(\alpha \mathbf{x} \cdot \mathbf{y} + c),$$

  где параметар α контролише нагиб сигмоидне фунцкије, а параметар c је слободан члан који омогућава померање дуж вертикалне осе

где су $\mathbf{x}$ и $\mathbf{y}$ вектори, а $\cdot$ је скаларни производ.

Ови кернели су само неки од најчешће коришћених. Постоје и други који се користе у зависности од специфичних проблема и захтева алгоритама. Избор одговарајућег кернела зависи од природе података и циља проблема који желимо да решимо.

4.2 Конструисање кернел матрице

Претпоставимо да имамо нелинеарну трансформацију $\phi(\mathbf{x})$ из оригиналног D-димензионалног простора атрибута у M-димензионални простор атрибута F, при чему је обично $M > D$. Тада се свака тачка података $\mathbf{x}_i$ пројектује у тачку $\phi(\mathbf{x}_i)$. Можемо применити стандардан метод главних компоненти у новом простору промењљивих, али то може бити рачунски захтевно. Срећом, можемо да користимо методе кернела да поједноставимо прорачун.

Ради једноставности, претпостављамо да пројекција нових промењљивих има средину нула:

$$\frac{1}{N} \sum_{i=1}^{N} \phi(\mathbf{x}_i) = 0\,.$$

То је лако постићи у улазном простору атрибута, али теже у простору F, јер не можемо експлицитно израчунати средњу вредност од $\phi(x_i)$ у том простору. Међутим, постоји начин да то урадимо, што доводи до благо модификованих једначина за анализу главних компоненти базираних на кернелима (видети [19]).

Матрица коваријансе пројектованих атрибута је формата $M \times M$ и израчунава се на следећи начин:

$$\mathbf{C} = \frac{1}{N} \sum_{i=1}^{N} \phi(\mathbf{x}_i)\phi(\mathbf{x}_i)^T \,. \tag{4.1}$$

Њене сопствене вредности и сопствени вектори су дати са

$$\mathbf{C}\mathbf{v}_k = \lambda_k \mathbf{v}_k \,, \tag{4.2}$$

где је $k = 1, 2, ..., M$. Из једнакости (4.1) и (4.2) имамо

$$\frac{1}{N} \sum_{i=1}^{N} \phi(\mathbf{x}_i)\phi(\mathbf{x}_i)^T \mathbf{v}_k = \lambda_k \mathbf{v}_k \,, \tag{4.3}$$

што се може преформулисати као

$$\mathbf{v}_k = \sum_{i=1}^{N} a_{ki}\phi(\mathbf{x}_i) \,. \tag{4.4}$$

Сада, заменом $\mathbf{v}_k$ из једнакости (4.4) у једнакост (4.3), добијамо

$$\frac{1}{N} \sum_{i=1}^{N} \phi(\mathbf{x}_i)\phi(\mathbf{x}_i)^T \sum_{j=1}^{N} a_{kj}\phi(\mathbf{x}_j) = \lambda_k \sum_{i=1}^{N} a_{ki}\phi(\mathbf{x}_i) \,. \tag{4.5}$$

Ако дефинишемо функцију кернела као:

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j) \,,$$

и помножимо обе стране једнакости (4.5) са $\phi(\mathbf{x}_l)^T$, добијамо

$$\frac{1}{N} \sum_{i=1}^{N} \kappa(\mathbf{x}_l, \mathbf{x}_i) \sum_{j=1}^{N} a_{kj}\kappa(\mathbf{x}_i, \mathbf{x}_j) = \lambda_k \sum_{i=1}^{N} a_{ki}\kappa(\mathbf{x}_l, \mathbf{x}_i) \,.$$

Можемо користити матричну нотацију

$$\mathbf{K}^2\mathbf{a}_k = \lambda_k N \mathbf{K}\mathbf{a}_k \,,$$

где је

$$\mathbf{K}_{i,j} = \kappa(\mathbf{x}_i, \mathbf{x}_j) \,, \tag{4.6}$$

а $\mathbf{a}_k$ је N-димензионални вектор колоне који садржи вредности a_{ki}:

$$\mathbf{a}_k = [a_{k1}\, a_{k2} \ldots a_{kN}]^T \,.$$

Вектор $\mathbf{a}_k$ добијамо из једначине

$$\mathbf{K}\mathbf{a}_k = \lambda_k N \mathbf{a}_k \,, \tag{4.7}$$

а резултујући кернел главних компоненти се може израчунати користећи

$$y_k(\mathbf{x}) = \phi(\mathbf{x})^T \mathbf{v}_k = \sum_{i=1}^{N} a_{ki} \kappa(\mathbf{x}, \mathbf{x}_i) \,. \tag{4.8}$$

Ако новодобијени скуп података $\{\phi(\mathbf{x}_i)\}$ нема средњу вредност једнаку нули, можемо користити Грамову матрицу $\widetilde{\mathbf{K}}$ као замену за кернел матрицу $\mathbf{K}$. Грамова матрица се дефинише на следећи начин

$$\widetilde{\mathbf{K}} = \mathbf{K} - \mathbf{1}_N \cdot \mathbf{K} - \mathbf{K} \cdot \mathbf{1}_N + \mathbf{1}_N \cdot \mathbf{K} \cdot \mathbf{1}_N \,, \tag{4.9}$$

где је $\mathbf{1}_N$ матрица димензије $N \times N$ са свим елементима једнаким $\frac{1}{N}$.

Моћ кернел метода је у томе што не морамо експлицитно да израчунамо $\phi(\mathbf{x}_i)$, већ можемо директно да конструишемо матрицу кернела из скупа података за обуку $\{\mathbf{x}_i\}$.

Два најчешће коришћена кернела за моделовање нелинеарних односа и решавање проблема који нису линеарно сепарабилни у оригиналном простору су полиномни кернел (енг. *polynomial kernel*)

$$\kappa(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^T \mathbf{y})^d \tag{4.10}$$

или

$$\kappa(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^T \mathbf{y} + c)^d \,,$$

где је $c > 0$ константа, и Гаусов кернел (енг. *Gaussian kernel*)

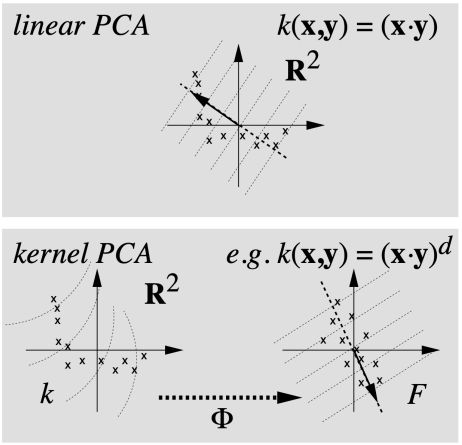
$$\kappa(\mathbf{x}, \mathbf{y}) = e^{-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}} \,, \tag{4.11}$$

са параметром σ.

На слици 4.1, која је преузета из [19], приказана је основна идеја методе главних компоненти која користи кернеле.

Да резимирамо, стандардни кораци за смањење димензије помоћу методе главних компоненти са кернел трансформацијама су следећи:

- конструкција кернел матрице $\mathbf{K}$ из скупа података за обуку $\{\mathbf{x}_i\}$ користећи једнакост (4.6),



Слика 4.1: Основна идеја анализе главних компоненти са кернел трансформацијама: У неком високодимензионом простору атрибута F (доле десно), изводимо класични PCA, баш као PCA у улазном простору (горе). Пошто је F нелинеарно повезан са улазним простором (преко Φ), контурне линије константних пројекција на главни сопствени вектор (нацртан као стрелица) постају нелинеарне у улазном простору. Важно је напоменути да не можемо нацртати предслику сопственог вектора у улазном простору, јер он можда чак и не постоји. Кључно за метод са кернел трансформацијама је чињеница да није потребно да се изврши пресликавање у F простор. Сва потребна израчунавања се изводе коришћењем функције кернела k у улазном простору (овде: $\mathbf{R}^2$).

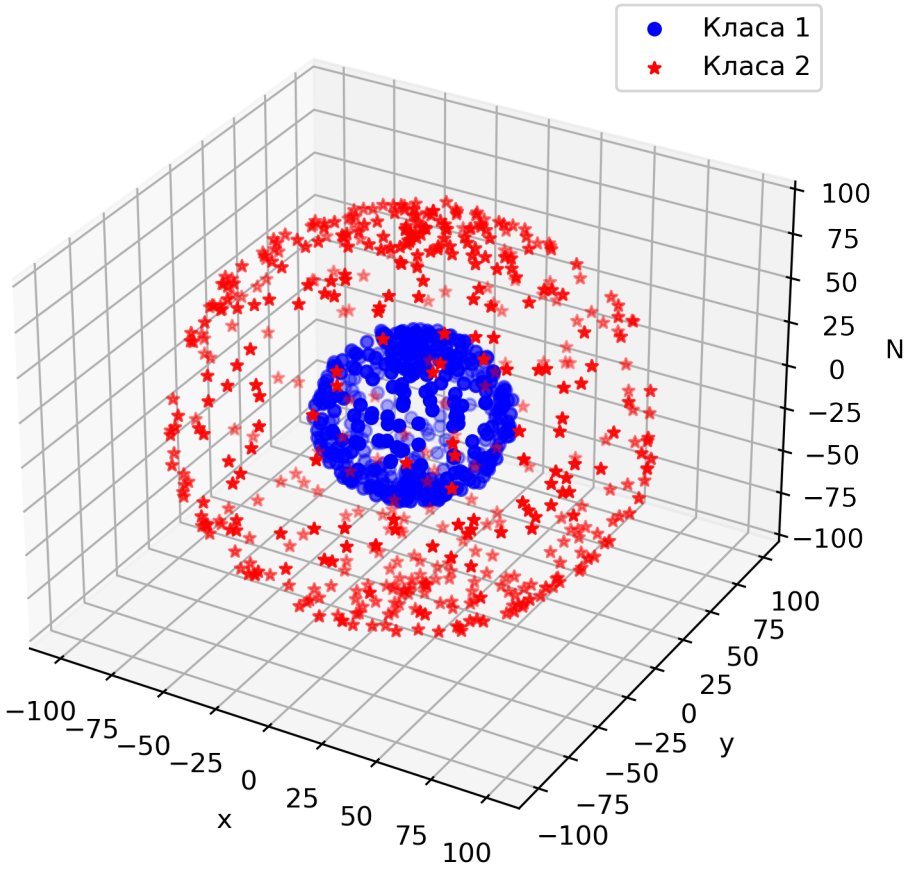
- израчунати Грамову матрицу $\widetilde{\mathbf{K}}$ помоћу једнакости (4.9),
- помоћу једнакости (4.7), при чему се $\mathbf{K}$ замени са $\widetilde{\mathbf{K}}$, одредити векторе $\mathbf{a}_i$,
- израчунати главне компоненте $y_k(\mathbf{x})$ користећи једнакост (4.8).

4.3 Примери

Вештачки скуп података

Пре него што алгоритам тестирамо на стварним подацима, генерисаћемо један вештачки скуп података и на њему тестирати наш алгоритам. Користићемо податке са два концентрична сферна скупа.

Претпоставимо да имамо једнак број тачака података распоређене на две концентричне сферне површине. Ако је N укупан број свих тачака података,

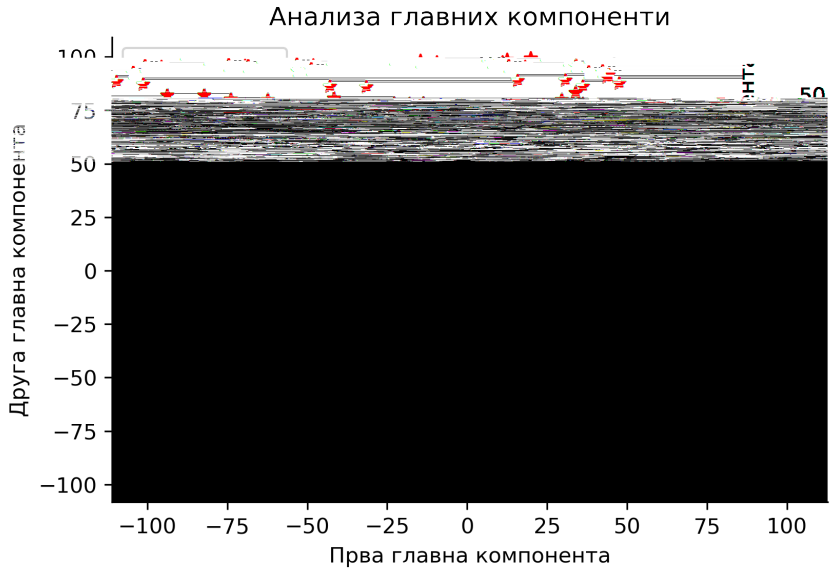


Слика 4.2: 3D приказ вештачких података са две концетричне сфере

онда имамо $\frac{N}{2}$ тачака класе 1 на сфери полупречника r_1 и $\frac{N}{2}$ тачака класе 2 на сфери полупречника r_2. У сферном координатном систему, нагиб (поларни угао) θ узима вредности из интервала $[0, \pi]$, а азимут ϕ узима вредности из интервала $[0, 2\pi)$, за обе класе. Наша посматрања тачака података су координате (x, y, z) у Декартовом координатном систему, и све три координате су узнемирене Гаусовим шумом стандардне девијације σ_{noise}. Постављамо $N = 1000$, $r_1 = 40$, $r_2 = 100$, $\sigma_{noise} = 1$ и приказујемо $3D$ график података на слици 4.2.

Резултати стандардне методе и методе која користи кернеле. Да бисмо визуализовали резултате, пројектујемо оригиналне тродимензионе податке у дводимензиони простор помоћу стандардне методе главних компоненти и методе која користи кернеле. За метод са кернел трансформацијама користимо полиномни кернел са параметром $d = 5$ и Гаусов кернел са пара-

метром $\sigma = 27$. Резултати све три методе дати су на слици 4.3, слици 4.4 и слици 4.5, редом.

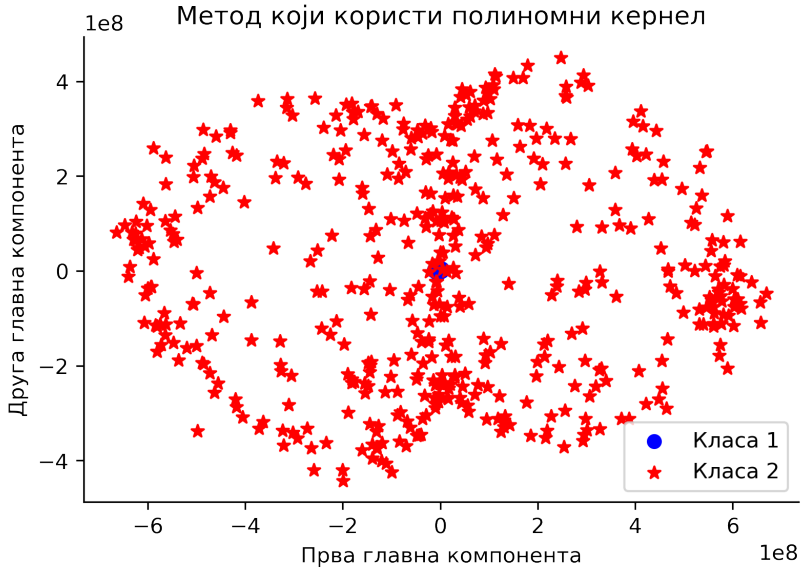


Слика 4.3: Резултати стандардне методе за вештачке податаке са две концентричне сфере

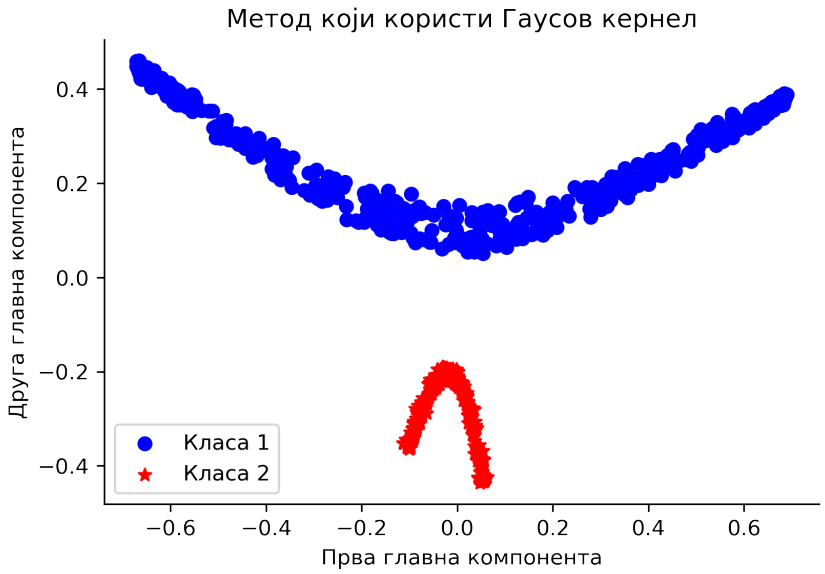
Треба напоменути да иако обележавамо тачке из различитих класа различитим бојама, заправо вршимо ненадгледано учење. Ни стандардни метод ни метод који користи кернеле не узимају ознаке класа као улазне податке.

На добијеним сликама можемо видети да стандардни метод не открива никакве структурне информације оригиналних података, односно резултујуће главне компоненте не дају потпростор где су подаци линеарно одвојиви. За метод који користи полиномни кернел, у новом простору атрибута, тачке података класе 2 су груписане, док су тачке података класе 1 распршене. Међутим, оне још увек нису линеарно одвојиве. За метод који користи Гаусов кернел, две класе су потпуно линеарно одвојиве.

Дискусија о примеру. Избор параметара за метод који користи кернеле директно одређује перформансе модела, па је то прво питање које се намеће. Како изабрати параметре? За Гаусов кернел који је дат са (4.11) најважнији параметар је σ. Гаусов кернел је функција удаљености $\|\mathbf{x}-\mathbf{y}\|$ између вектора $\mathbf{x}$ и $\mathbf{y}$. Ако желимо да раздвојимо различите класе у новом простору атрибута, тада би параметар σ требало да буде мањи од међукласних растојања, и већи од растојања унутар класа. Међутим, ми не знамо колико има класа у



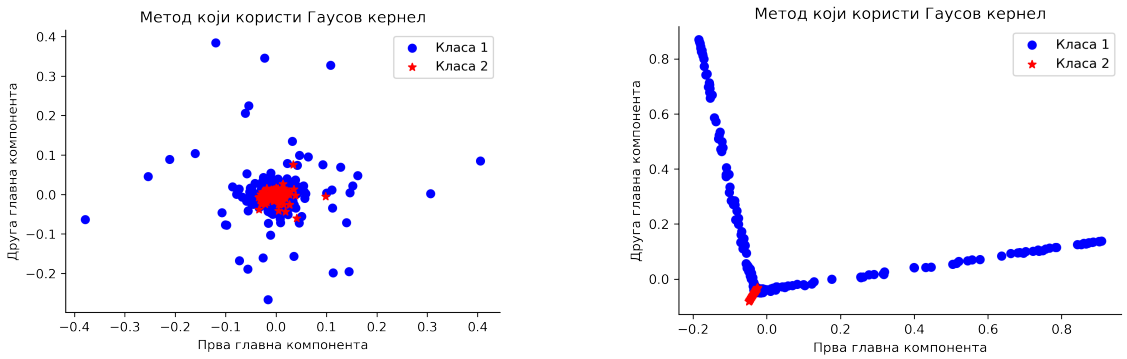
Слика 4.4: Резултати методе која користи полиномни кернел за податаке са две концетричне сфере



Слика 4.5: Резултати методе која користи Гаусов кернел за податаке са две концетричне сфере

подацима, па није лако проценити међукласна или унутаркласна растојања. Алтернативно, можемо поставити σ на малу вредност како бисмо ухватили само информације о суседима за сваку тачку података. Избор σ зависи од скупа података и може се добити помоћу техника подешавања хиперпараметара

као што је претрага мреже (енг. *Grid Search*). Подешавање хиперпараметара је само по себи широка тема и овде смо користили σ-вредност за коју смо открили да даје „добре“ резултате. На слици 4.6 можемо видети како се резултати методе мењају са избором параметра σ.



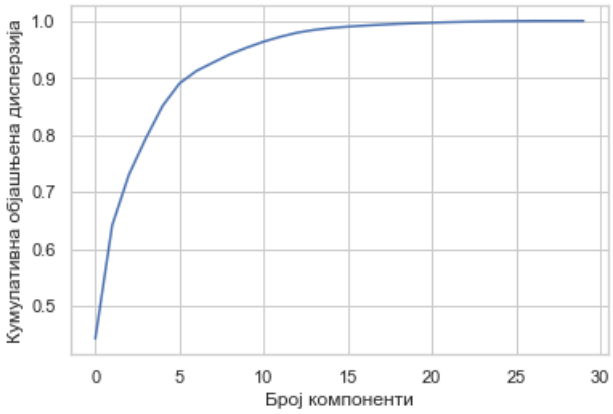
Слика 4.6: Резултати методе која користи Гаусов кернел са параметром $\sigma = 0.03$ и $\sigma = 5$ редом

Интуитивно објашњење Гаусовог кернела. Као што смо могли да видимо, у експерименту класификације вештачких података, Гаусов кернел са правилно одабраним параметром σ може лепо да раздвоји две класе, што класична метода главних компоненти није успела.

Интуитивно, метод главних компоненти са Гаусовим кернелом користи удаљености између различитих тачака скупа за обучавање, слично као метод k најближих суседа или методе кластеровања. Са добро изабраним параметром σ, метод главних компоненти који користи Гаусов кернел ће имати одговарајући распон хватања, што ће побољшати везу између тачака података које су блиске једна другој у оригиналном простору атрибута. Затим, применом анализе сопствених вектора, сопствени вектори ће описати правце у простору високих димензија у коме су различити кластери података најраспршенији.

База података *Breast Cancer*

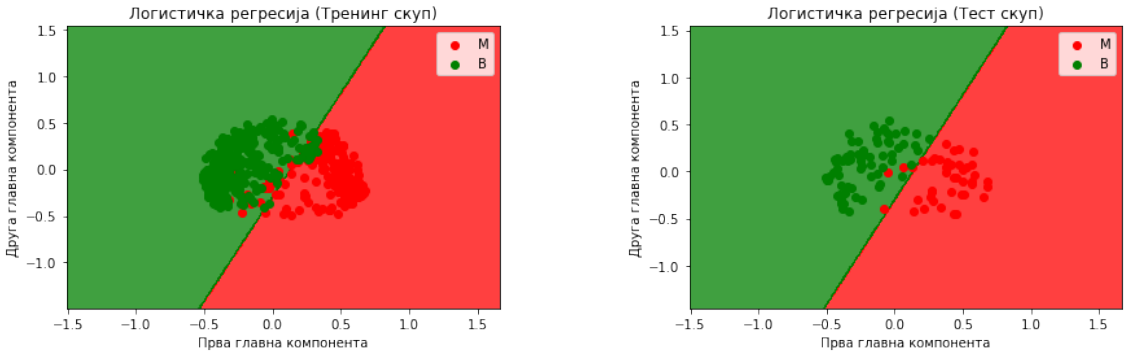
У овом делу ћемо се фокусирати на примену методе главних компоненти која користи кернеле на скупу података о раку дојке. Главни циљ нам је да смањимо димензију скупа података и да побољшамо перформансе модела класификације. Осим тога, желимо да упоредимо перформансе методе са кернел трансформацијама са стандардном методом како бисмо утврдили који од ових приступа боље одговара нашем скупу података.



Слика 4.7: Однос броја компоненти и објашњене дисперзије

Применићемо радијалну базну функцију (*RBF*), односно Гаусов кернел за смањење димензије наших података. Оно што можемо видети са слике 4.7 и из анализе добијених компоненти јесте да првих пет компоненти објашњава преко 80% дисперзије, тачније 85.23%. Дакле, даљу анализу бисмо могли засновати на тих пет компоненти, међутим ми ћемо у даљем раду задржати само прве две компоненте, ради лакше визуализације.

Користећи ове компоненте, обучићемо модел логистичке регресије и проверити његову исправност користећи матрицу конфузије. Подаци су подељени тако да тренинг скуп садржи 455 опсервација, а тест скуп 114 опсервација, при чему је подела случајна. Смањивање димензије скупа података са тридесет атрибута на две главне компоненте довело је до модела са добрим перформансама. Овај модел има тачност од 96.49% на тест скупу.

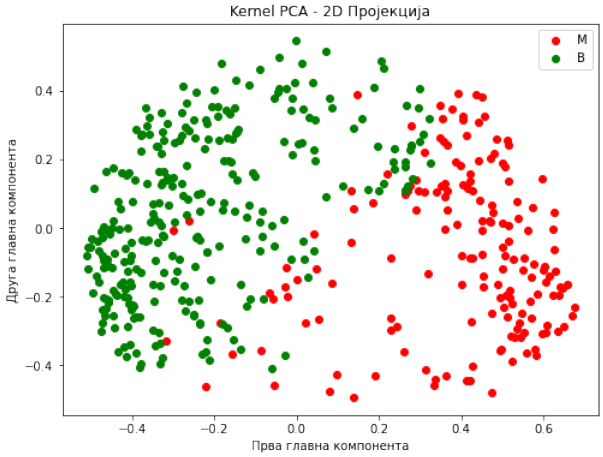


Слика 4.8: Класификација тумора логистичком регресијом на тренинг и тест скупу помоћу прве две главне компоненнте

На слици 4.8 можемо да видимо како су подаци класификовани за оба ску-

па, тренинг и тест. Такође, можемо видети да смо добили атрибуте који су линеарно одвојиви.

На истом скупу података смо обучили модел логистичке регресије уз претходну примену стандардне методе главних компоненти, при чему смо поново задржали само прве две компоненте. Овај модел се показао бољим и има тачност од 99.12% на тест скупу. На крају, када упоредимо резултате оба метода можемо закључити да су подаци скоро линеарно раздвојиви и да стандардна метода даје боље резултате.



Слика 4.9: 2D пројекција прве две компоненте код анализе главних компоненти са кернел трансформацијама

Ако погледамо слику 4.9, могли бисмо закључити да главне компоненте добијене применом стандардне методе боље раздвајају класе од оних које смо добили применом кернела. Свакако, у оба случаја смо смањили димензију наших података, али и задржали значајне информације за класификацију.

Глава 5

Анализа главних компоненти проређених података

Анализа главних компоненти има очигледнуману, а то је да су све главне компоненте линеарне комбинације свих p променљивих и коефицијенти (енг. *loadings*) су обично различити од нуле. То често отежава тумачење добијених компоненти. Један начин да се ово превазиђе јесу технике ротације које помажу при тумачењу добијених главних компоненти. Други приступ је да се вештачки поставе тежине са малим апсолутним вредностима на нулу, чиме се ефикасно смањује број експлицитно коришћених променљивих. Међутим, овакав приступ постављању прага може бити обмањујући.

Мек Кејб је 1984. у свом раду [15] предложио алтернативу методе главних компоненти која идентификује подскуп главних променљивих. Џолиф, Трендафилов и Удин су 2003. године у раду [13] увели SCoTLASS (*Simplified Component Technique-LASSO*), метод за добијање модификованих главних компоненти које могу имати нулте тежине. Ови методи имају за циљ постизање смањења димензије уз истовремено смањење броја експлицитно коришћених променљивих.

На основу увида да се метод главних компоненти може формулисати као оптимизациони проблем сличан регресији са квадратном казненом функцијом, односно са квадратним ограничењима на коефицијенте модела, уводимо нови приступ (енг. *Sparse Principal Component Analysis*) за процену главних компоненти са ретким оптерећењем. Овај приступ директно интегрише ласо (енг. *lasso - least absolute shrinkage and selection operator*) казну (путем еластичне мреже) у критеријум регресије, што доводи до модификованог метода

главних компоненти са ретким оптерећењима.

5.1 Ласо и еластична мрежа

Посматрамо модел линеарне регресије са n опсервација и p предиктора. Нека $Y = (y_1, ..., y_n)^T$ представља вектор одзива, а $X = [X_1, ..., X_p]$ предикторе, где је $X_j = (x_{1j}, ..., x_{nj})^T$ j-ти предиктор за $j = 1, ..., p$. Претпоставимо да су сви X_j и Y центрирани, односно да им је средња вредност нула.

Ласо је статистичка метода пенализованих најмањих квадрата која намеће ограничења на l_1 норму регресионих коефицијената. Дакле, ласо процене коефицијената $\hat{\beta}_{lasso}$ се добијају минимизовањем следећег критеријума

$$\hat{\beta}_{lasso} = \arg\min_{\beta} \left(\|Y - \sum_{j=1}^{p} X_j \beta_j\|^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right) ,$$

где је λ ненегативно.

Због природе l_1 казне, неки коефицијенти ће бити смањени тачно на нулу. Ово својство чини ласо изузетно корисним за селекцију најважнијих променљивих између много предиктора. Дакле, кључни циљ овог метода је пронаћи баланс између тачности предвиђања и једноставности модела, што се постиже смањењем небитних коефицијената на нулу. Овај процес омогућава стварање ретког модела који користи само најважније атрибуте за предвиђање, што олакшава тумачење резултата и смањује ризик од преприлагођавања на малим узорцима.

Еластична мрежа генерализује ласо како би превазишла њене недостатке, истовремено задржавајући повољна својства. За било које ненегативне вредности λ_1 и λ_2, оцена коефицијената $\hat{\beta}_{en}$ дата је са

$$\hat{\beta}_{en} = (1 + \lambda_2) \left(\arg\min_{\beta} \left(\|Y - \sum_{j=1}^{p} X_j \beta_j\|^2 + \lambda_2 \sum_{j=1}^{p} |\beta_j|^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| \right) \right) .$$

Еластична мрежа је конвексна комбинација гребене (енг. *ridge*) и ласо регресије. Јасно је да је ласо специјалан случај еластичне мреже када је $\lambda_2 = 0$.

5.2 Мотивација и објашњење ретке анализе главних компоненти

Сама чињеница да су коефицијенти ретки како у ласо тако и у еластичној мрежи је директна последица l_1 казне и не зависи од функције квадратне грешке. Интересантан поступак који добија ретке коефицијенте директним наметањем l_1 ограничења на анализу главних компоненти јесте SCoTLASS (*Simplified Component Technique-LASSO*). Он сукцесивно максимизује варијансу

$$a_k^T(\mathbf{X}^T\mathbf{X})a_k\,,$$

уз услов

$$a_k^T a_k = 1 \text{ и (за } k \geq 2)\ a_h^T a_k = 0,\ h < k\,,$$

уз додатно ограничење

$$\sum_{j=1}^{p} |a_{kj}| \leq t\,,$$

за неки параметар подешавања t. Иако довољно мала вредност t доводи до коефицијената једнаких нули, нема много смерница у SCoTLASS-у за одабир одговарајуће вредности за t. Могли бисмо пробати са неколико вредности t, али висока рачунарска сложеност SCoTLASS-а чини то неприхватљивим решењем. Ова висока рачунарска сложеност вероватно је последица чињенице да SCoTLASS није конвексан оптимизациони проблем. Такође, показало се да коефицијенти добијени путем SCoTLASS-а нису довољно ретки када је потребан висок проценат објашњеног варијабилитета.

Овде ћемо видети другачији приступ модификовању анализе главних компоненти. Показаћемо како се стандардан метод може преформулисати као проблем регресије (гребене регресије). Затим уводимо ласо казну тако што гребену регресију претварамо у регресију еластичне мреже.

Директне ретке апроксимације

Без умањења општости, претпоставимо да све колоне матрице $\mathbf{X}$ имају средњу вредност једнаку нули. Нека је спектрална декомпозиција матрице $\mathbf{X}$ дата са

$$\mathbf{X} = \mathbf{UDV}^T\,.$$

Нека $\mathbf{Z} = \mathbf{UD}$ представља главне компоненте, а колоне матрице $\mathbf{V}$ представљају одговарајућа оптерећења главних компоненти.

Прво ћемо размотрити једноставан регресиони приступ анализи главних компоненти. Свака главна компонента је линеарна комбинација p променљивих, па се њени коефицијенти могу добити регресијом главне компоненте на p променљивих.

Теорема 5.1. *За свако i, означимо са $Z_i = U_i D_{ii}$ i-ту главну компоненту. Нека је λ позитивно и нека је $\hat{\beta}_{ridge}$ оцена коефицијената која се добија помоћу гребене методе*

$$\hat{\beta}_{ridge} = \arg\min_{\beta} \left(\|Z_i - \mathbf{X}\beta\|^2 + \lambda \|\beta\|^2 \right) . \tag{5.1}$$

Нека је $\hat{v} = \frac{\beta_{ridge}}{\|\beta_{ridge}\|}$, тада је $\hat{v} = V_i$.

Доказ теореме, као и њена формулација се може наћи у раду [24]. Ова теорема нам даје везу између анализе главних компоненти и методе регресије. Када је $n > p$ и $\mathbf{X}$ пуног ранга, теорема не захтева позитивно λ. Када је $p > n$ и $\lambda = 0$, обична вишеструка регресија нема јединствено решење које је баш V_i. Исто се дешава када је $n > p$ и матрица $\mathbf{X}$ није пуног ранга. Међутим, метод главних компоненти увек даје јединствено решење у свим ситуацијама. Као што је приказано у теореми, ова неодређеност се елиминише позитивном гребеном казном ($\lambda\|\beta\|^2$). Након нормализације, коефицијенти су независни од λ, стога се гребена казна не користи за пенализацију регресионих коефицијената, већ да би се осигурала реконструкција главних компоненти.

Даље, додајемо l_1 казну у (5.1) и разматрамо следећи проблем оптимизације

$$\hat{\beta} = \arg\min_{\beta} \left(\|Z_i - \mathbf{X}\beta\|^2 + \lambda \|\beta\|^2 + \lambda_1 \|\beta\|_1 \right) , \tag{5.2}$$

где је $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_1|$ l_1 норма од β. $\hat{V}_i = \frac{\hat{\beta}}{\|\hat{\beta}\|}$ је апроксимација за V_i и $\mathbf{X}\hat{V}_i$ је апроксимација i-те главне компоненте. Ова техника се назива и наивна еластична мрежа (енг. *naive elastic net*), која се разликује од праве еластичне мреже само по фактору скалирања $(1 + \lambda)$. Пошто користимо нормализоване коефицијенте, фактор скалирања не утиче на V_i. Довољно велико λ_1 даје ретке $\hat{\beta}$, самим тим ретке и $\hat{V}_i$. Са фиксним λ, (5.2) се ефикасно решава за све λ_1 коришћењем ларсен (енг. LARS-EN - least angle regression - elastic net) алгоритма. Дакле, можемо флексибилно изабрати ретку апроксимацију за i-ту главну компоненту.

Ретке главне компоненте на основу SPCA критеријума

Теорема 5.1 зависи од резултата анализе главних компоненти, па није права алтернатива. Међутим, може се користити у двостепеној истраживачкој анализи тако што прво извршимо анализу главних компоненти, а затим користимо (5.2) да бисмо пронашли одговарајуће ретке апроксимације.

Сада ћемо представити критеријум заснован на регресији за директно добијање главних компоненти, без потребе за претходним израчунавањем. Овај приступ нам омогућава да директно процењујемо главне компоненте као линеарну комбинацију променљивих и пронађемо најважније векторе који објашњавају варијабилност у подацима. Нека $\mathbf{x}_i$ означава вектор i-тог реда матрице $\mathbf{X}$. Прво ћемо разматрати прву главну компоненту.

Теорема 5.2. *Нека за свако $\lambda > 0$ важи*

$$\begin{gathered}(\hat{\alpha}, \hat{\beta}) = \arg\min_{\alpha,\beta} \sum_{i=1}^{n} \|\mathbf{x}_i - \alpha\beta^T \mathbf{x}_i\|^2 + \lambda\|\beta\|^2 \\ \textit{уз услов } \|\alpha\|^2 = 1\,.\end{gathered} \tag{5.3}$$

Тада је $\hat{\beta}$ пропорционално првој главној компоненти V_1.

Следећа теорема проширује Теорему 5.2 како би се извео цео низ главних компоненти.

Теорема 5.3. *Претпоставимо да разматрамо првих k главних компоненти. Нека је $\mathbf{A}_{p\times k} = [\alpha_1, ..., \alpha_k]$ и $\mathbf{B}_{p\times k} = [\beta_1, ..., \beta_k]$. Нека за свако $\lambda > 0$ важи*

$$\begin{gathered}(\hat{\mathbf{A}}, \hat{\mathbf{B}}) = \arg\min_{\mathbf{A},\mathbf{B}} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{A}\mathbf{B}^T \mathbf{x}_i\|^2 + \lambda \sum_{j=1}^{k} \|\beta_j\|^2 \\ \textit{уз услов } \mathbf{A}^T\mathbf{A} = I_{k\times'}\,.\end{gathered} \tag{5.4}$$

Тада је $\hat{\beta}_j$ пропорционално V_j за $j = 1, 2, ..., k$.

Претходне две теореме ефикасно трансформишу анализу главних компоненти у регресиони проблем. Критични елемент је циљна функција $\sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{A}\mathbf{B}^T\mathbf{x}_i\|^2$. Ако је $\mathbf{B} = \mathbf{A}$ онда

$$\sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{A}\mathbf{B}^T\mathbf{x}_i\|^2 = \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{A}\mathbf{A}^T\mathbf{x}_i\|^2\,,$$

чији минимум под ортонормалним ограничењем на $\mathbf{A}$ је тачно првих k вектора оптерећења стандардне методе главних компоненти. Теорема 5.3 показује да можемо добити тачну алализу главних компоненти чак и уз попуштање ограничења $\mathbf{B} = \mathbf{A}$ и додавање казне за гребену регресију. Како ћемо видети касније, ове генерализације нам омогућавају да флексибилно модификујемо традиционални метод.

Нећемо спроводити доказ ових теорема, већ ћемо дати интуитивно објашњење. Приметимо да је

$$\sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{A}\mathbf{B}^T\mathbf{x}_i\|^2 = \|\mathbf{X} - \mathbf{X}\mathbf{B}\mathbf{A}^T\|^2 .$$

С обзиром да је матрица $\mathbf{A}$ ортогонална, нека је $\mathbf{A}_\perp$ било која ортогонална матрица таква да је $[\mathbf{A}; \mathbf{A}_\perp]$ ортогонална $p \times p$ матрица. Тада имамо

$$\begin{aligned} \|\mathbf{X} - \mathbf{X}\mathbf{B}\mathbf{A}^T\|^2 &= \|\mathbf{X}\mathbf{A}_\perp\|^2 + \|\mathbf{X}\mathbf{A} - \mathbf{X}\mathbf{B}\|^2 \\ &= \|\mathbf{X}\mathbf{A}_\perp\|^2 + \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{A}\mathbf{B}^T\mathbf{x}_i\|^2 . \end{aligned} \tag{5.5}$$

Претпоставимо да је $\mathbf{A}$ дато, онда би оптимално $\mathbf{B}$ које минимизује (5.2) требало да минимизује

$$\arg\min_{\mathbf{B}} \sum_{i=1}^{n} \{\|\mathbf{X}\alpha_j - \mathbf{X}\beta_j\|^2 + \lambda\|\beta_j\|^2\}$$

што је еквивалентно са k независних проблема гребене регресије. Конкретно, ако $\mathbf{A}$ одговара обичним главним компонентама, то јест $\mathbf{A} = \mathbf{V}$, онда по Теореми 5.1 знамо да би $\mathbf{B}$ требало бити пропорционално са $\mathbf{V}$.

Настављамо са повезивањем методе главних компоненти и регресије, при чему користимо ласо приступ како бисмо добили ретке коефицијенте у регресији. У ту сврху, критеријуму (3.7) додајемо ласо казну и разматрамо следећи оптимизациони проблем

$$\begin{aligned} (\hat{\mathbf{A}}, \hat{\mathbf{B}}) = \arg\min_{\mathbf{A},\mathbf{B}} &\sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{A}\mathbf{B}^T\mathbf{x}_i\|^2 + \lambda \sum_{j=1}^{k} \|\beta_j\|^2 + \sum_{j=1}^{k} \lambda_{1,j}\|\beta_j\|_1 \\ &\text{уз услов } \mathbf{A}^T\mathbf{A} = I_{k\times k} . \end{aligned} \tag{5.6}$$

Док се исто λ користи за свих k компоненти, дозвољено је да се различити $\lambda_{1,j}$ користе за кажњавање оптерећења различитих главних компоненти. Поново, ако је $p > n$, потребно је користити позитивно λ како бисмо добили тачну анализу главних компоненти када ограничење реткости (казна ласо) нестане ($\lambda_{1,j} = 0$).

Општи алгоритам.

- Иницијализујемо матрицу $\mathbf{A}$ са $\mathbf{V}[,1:k]$, која представља коефицијенте (енг. *loadings*) првих k обичних главних компоненти.
- За фиксирано $\mathbf{A} = [\alpha_1, ..., \alpha_k]$, решавамо следеће за $j = 1, 2, ..., k$:

$$\beta_j = \arg\min_{\beta} (\alpha_j - \beta)^T \mathbf{X}^T \mathbf{X} (\alpha_j - \beta) + \lambda \|\beta\|^2 + \lambda_{1,j} \|\beta\|_1 \,,$$

  где је β вектор коефицијената који репрезентује j-ту модификовану главну компоненту.

- За фиксирано $\mathbf{B} = [\beta_1, ..., \beta_k]$, рачунамо сингуларну декомпозизију матрице $\mathbf{X}^T\mathbf{X}\mathbf{B} = \mathbf{U}\mathbf{D}\mathbf{V}^T$, а затим ажурирамо матрицу $\mathbf{A} = \mathbf{U}\mathbf{V}^T$.
- Понављамо кораке $2 - 3$ до конвергенције.
- Нормализујемо $\hat{V}_j = \frac{\beta_j}{\|\beta_j\|}$, $j = 1, ..., k$.

Напомене:

- Емпиријски докази сугеришу да се излаз горњег алгоритма не мења много при промени λ. За податке са $n > p$ (где је n број узорака, а p број обележја), подразумевани избор λ може бити нула. У пракси, λ се обично бира као мала позитивна вредност како би се превазишли потенцијални проблеми са колинеаритетом у матрици $\mathbf{X}$.
- У принципу, можемо испробати неколико комбинација $\{\lambda_{1,j}\}$ да бисмо пронашли добар избор параметара подешавања, пошто горњи алгоритам конвергира прилично брзо. Постоји пречица коју пружа директна ретка апроксимација (5.2). Алгоритам ларсен, описан у раду [23], ефикасно даје целу секвенцу прорећених апроксимација за сваку главну компоненту и одговарајуће вредности $\lambda_{1,j}$. Тако можемо изабрати $\lambda_{1,j}$ који даје добар компромис између дисперзије и прорећености. Када правимо компромис између дисперзије и прорећености, ипак већи приоритет дајемо дисперзији.

Прилагођена укупна дисперзија

Обичне главне компоненте су некорелисане и њихова оптерећења су ортогонална. Нека је $\hat{\mathbf{\Sigma}} = \mathbf{X}^T\mathbf{X}$, тада је $\mathbf{V}^T\mathbf{V} = \mathbf{I}_k$ и $\mathbf{V}^T\hat{\mathbf{\Sigma}}\mathbf{V}$ је дијагонала матрица.

Лако је проверити да је ово могуће само за стандардне главне компоненте, где оптерећења могу задовољити оба услова. Ретка анализа главних компоненти не намеће експлицитно услов некорелисаности компоненти.

Нека је $\hat{\mathbf{Z}}$ модификована главна компинента. Обично се укупна дисперзија објашњена помоћу $\hat{\mathbf{Z}}$ рачуна као траг матрице $\hat{\mathbf{Z}}^T\hat{\mathbf{Z}}$. Ово има смисла када су $\hat{\mathbf{Z}}$ некорелисане. Међутим, ако су оне корелисане, траг од $\hat{\mathbf{Z}}^T\hat{\mathbf{Z}}$ је превише оптимистичан за представљање укупне дисперзије. Претпоставимо да је $(\hat{Z}_i,\, i = 1, 2, ..., k)$ првих k модификованих главних компоненти, а $\hat{Z}_{k+1}$ је $(k+1)$-ва модификована главна компонента добијена кроз процес рачунања додатних главних компоненти изван првих k компоненти. Желимо да израчунамо укупну дисперзију објашњену од стране првих $k+1$ модификованих главних компоненти, која би требало да буде збир дисперзије објашњене преко првих k модификованих главних компоненти и додатне дисперзије од $\hat{Z}_{k+1}$. Ако је $\hat{Z}_{k+1}$ у корелацији са $(\hat{Z}_i,\, i = 1, 2, ..., k)$, тада њена дисперзија садржи доприносе од $(\hat{Z}_i,\, i = 1, 2, ..., k)$, које не треба укључивати у укупну дисперзију с обзиром на присуство $(\hat{Z}_i,\, i = 1, 2, ..., k)$.

Сада уводимо формулу за израчунавање укупне дисперзије објашњене од стране $\hat{\mathbf{Z}}$, која узима у обзир корелације између $\hat{\mathbf{Z}}$. Користимо пројекцију регресије да уклонимо линеарну зависност између корелираних компоненти. Означимо са $\hat{Z}_{j\cdot 1,\ldots,j-1}$ остатке (резидуалне вредности) компоненте $\hat{Z}_j$ након што је та компонента прилагођена или коригована на основу свих претходних компоненти $\hat{Z}_1, \hat{Z}_2, ..., \hat{Z}_{j-1}$. То можемо записти као

$$\hat{Z}_{j\cdot 1,\ldots,j-1} = \hat{Z}_j - \mathbf{H}_{1,\ldots,j-1}\hat{Z}_j$$

где $\mathbf{H}_{1,\ldots,j-1}$ представља линеарну пројекцију или регресију компоненте $\hat{Z}_j$ на основу $\hat{Z}_1, \hat{Z}_2, ..., \hat{Z}_{j-1}$. Затим прилагођена дисперзија од $\hat{Z}_j$ је $\|\hat{Z}_{j\cdot 1,\ldots,j-1}\|^2$, а укупна објашњена дисперзија је дефинисана као сума прилагођених дисперзија за све компоненте, тј. $\sum_{j=1}^{k}\|\hat{Z}_{j\cdot 1,\ldots,j-1}\|^2$. Када су модификоване главне компоненте $\hat{\mathbf{Z}}$ некорелисане, нова формула се слаже са $\mathrm{tr}(\hat{\mathbf{Z}}^T\hat{\mathbf{Z}})$.

Треба имати на уму да претходна израчунавања зависе од реда $\hat{Z}_i$. Међутим, пошто постоји природан редослед у анализи главних компоненти, редослед овде није проблем. Коришћењем QR декомпозиције, лако можемо израчунати прилагођену дисперзију. Претпоставимо $\hat{\mathbf{Z}} = \mathbf{QR}$, где је $\mathbf{Q}$ ортогонална матрица, а $\mathbf{R}$ је горње троугаона матрица. Онда је лако видети

да

$$\|\hat{Z}_{j\cdot 1,\ldots,j-1}\|^2 = \mathbf{R}^2_{jj}\,.$$

Отуда је укупна објашњена дисперзија једнака $\sum_{j=1}^{k} \mathbf{R}^2_{jj}$.

5.3 Примери

Вештачки формирани подаци

У овом примеру, имамо три скривена фактора V_1, V_2 и V_3, који су генерисани на следећи начин:

$$V_1 \sim \mathcal{N}(0, 150)$$

$$V_2 \sim \mathcal{N}(0, 200)$$

$$V_3 = -0.325V_1 + 0.9V_2 + \epsilon,\ \epsilon \sim \mathcal{N}(0, 1)$$

где су V_1, V_2 и ϵ независни. Затим је конструисано 10 променљивих X_i на следећи начин:

- $X_i = V_1 + \epsilon_i^1$ где $\epsilon_i^1 \sim \mathcal{N}(0, 1)$, $i = 1, 2, 3, 4$,
- $X_i = V_2 + \epsilon_i^2$ где $\epsilon_i^2 \sim \mathcal{N}(0, 1)$, $i = 5, 6, 7, 8$,
- $X_i = V_3 + \epsilon_i^3$ где $\epsilon_i^3 \sim \mathcal{N}(0, 1)$, $i = 9, 10$,
- $\{\epsilon_i^j\}$ су независне, $j = 1, 2, 3$, $i = 1, ..., 10$.

Користимо тачну матрицу коваријансе од $(X_1, X_2, ..., X_{10})$ за примене стандардне и методе ретких главних компоненти у популацијском окружењу. То значи да матрица коваријансе која се користи за ове методе садржи тачне популацијске параметре уместо процењених вредности на основу узорка.

Обично, када примењујемо стандардан метод, метод ретких главних компоненти или било коју другу технику смањења димензије, радимо са узоркованим подацима, а матрица коваријансе се процењује на основу посматраних података. Међутим, у овом специфичном примеру, вештачки подаци се генеришу са познатим основним факторима и дисперзијама за опажене променљиве $(X_1, X_2, ..., X_{10})$. Стога је могуће користити тачну матрицу коваријансе добијену из познатих популацијских параметара.

	Метод главних компоненти			Метод главних компоненти прорeђених података	
	PC1	PC2	PC3	PC1	PC2
X_1	0.15236616	0.46807832	-0.10628757	0	-0.49998365
X_2	0.1545458	0.46695739	-0.13845437	0	-0.49997505
X_3	0.15143775	0.46865299	-0.00698521	0	-0.50005688
X_4	0.15181933	0.46835821	-0.09629752	0	-0.49998441
X_5	-0.38046607	0.17555079	0.21451387	0.49999383	0
X_6	-0.38010746	0.1766794	0.27842797	0.50003378	0
X_7	-0.38095722	0.17364795	0.26574665	0.49996166	0
X_8	-0.38057934	0.17487002	0.32391996	0.50001073	0
X_9	-0.40483727	0.02291948	-0.55648469	0	0
X_{10}	-0.40474594	0.02379739	-0.59283419	0	0
Објашњена дисперзија	60.46%	38.99%	0.12%	49.94%	49.82%

Слика 5.1: Резултати примера симулације: оптерећења и дисперзија

Дисперзије три основна фактора су, редом, 150, 200 и 178.8. Број променљивих повезаних с тим факторима је 4, 4 и 2. Из тога произлази да су V_2 и V_1 скоро подједнако важни, а много важнији од V_3. Прве две главне компоненте заједно објашњавају 99.45% укупне дисперзије. Ови подаци сугеришу да нам је потребно размотрити само две изведене варијабле са „исправним" штурим репрезентацијама. Идеално, прва изведена компонента би требало да реконструише фактор V_2 користећи само (X_5, X_6, X_7, X_8), док би друга изведена варијабла требало да реконструише фактор V_1 користећи само (X_1, X_2, X_3, X_4).

У ствари, ако поступно максимизујемо дисперзију прве две изведених компоненте уз услов ортонормалности, и притом ограничимо број ненултих оптерећења на четири, прва изведена компонента ће једнако додељивати ненулта оптерећења на (X_5, X_6, X_7, X_8), а друга изведена варијабла ће их једнако додељивати на (X_1, X_2, X_3, X_4).

На слици 5.1 су дати резултати поређења. Јасно је да метод ретких главних компоненти правилно идентификује скупове важних променљивих и пружа ретке репрезентације првих двеју главних компоненти. То се може објаснити чињеницом да метода експлицитно користе ласо као казнену функцију.

База *Pitprops*

Скуп података *Pitprops*, који је представио Џеферс 1967. године у раду [1], постао је класичан пример који показује тешкоће у тумачењу главних компоненати. У овом скупу података забележене су величине и особине 180 потпорних дрвених стубова (дрвених греда) који се користе за подршку крововима тунела у рудницима угља. Доступно је 13 атрибута и то су: горњи пречник стуба (*topidam*), дужина стуба (*length*), садржај влаге у дрвету (*moist*), спе-

цифична тежина дрвета у време теста (*testsg*), специфична тежина дрвета када је потпуно осушено у рерни (*ovensg*), број годишњих прстенова на врху стуба (*ringtop*), број годишњих прстенова при дну стуба (*ringbut*), максимална савијеност (*bowmax*), удаљеност тачке максималне савијености од врха стуба (*bowdist*), број чворова у спирали (*whorls*), дужина чистог дела стуба од врха (*clear*), просечан број чворова по спирали (*knots*) и просечни пречник чворова (*diaknot*).

Атрибути	PC1	PC2	PC3	PC4	PC5	PC6
topdiam	-0.404	-0.217	0.207	-0.091	0.083	0.12
length	-0.406	-0.186	0.235	-0.103	0.113	0.163
moist	-0.124	-0.54	-0.141	0.078	-0.35	-0.256
testsg	-0.173	-0.456	-0.352	0.055	-0.356	-0.054
ovensg	-0.057	0.17	-0.481	0.049	-0.176	0.626
ringtop	-0.284	0.014	-0.475	-0.063	0.316	0.052
ringbut	-0.4	0.19	-0.253	-0.065	0.215	0.003
bowmax	-0.294	0.189	0.243	0.286	-0.185	-0.055
bowdist	-0.357	-0.017	0.208	0.097	0.106	0.034
whorls	-0.379	0.248	0.119	-0.205	-0.156	-0.173
clear	0.011	-0.205	0.07	0.804	0.343	0.175
knots	0.115	-0.343	-0.092	-0.301	0.6	-0.17
diaknot	0.113	-0.309	0.326	-0.303	-0.08	0.626
Објашњена дисперзија	32.45%	18.29%	14.44%	8.53%	7.00%	6.27%

Слика 5.2: Оптерећења првих шест главних компоненти *Pitprops* скупа података

Као демонстрацију, размотрили смо првих шест главних компоненти. Слика 5.2 приказује резултате стандардне методе главних компоненти, док слика 5.3 приказује резултате анализе главних компоненти прорећених података. Будући да је ово уобичајен скуп података са $n \gg p$, односно скуп где је број опсервација много већи од броја атрибута, поставили смо параметар $\lambda = 0$. Вредности за параметар $\lambda_1 = (0.06, 0.16, 0.1, 0.5, 0.5, 0.5)$ су изабране тако да свака ретка апроксимација објасни готово исту количину дисперзије као и обична главна компонента. Табела са слике 5.3 приказује добијене главне компоненте на прорећеним подацима и одговарајућу прилагођену дисперзију.

Можемо видети велику разлику између ове две методе. Основни алгоритам је прихватио више променљивих, што значи да све променљиве се равноправно узимају у обзир при формирању компоненти. Са друге стране, метода главних компоненти прорећених податка одбацила је већи део атрибута, фокусирајући се само на најзначајније елементе. Овакав приступ омогућава смањење димензије података и истовремено задржава само кључне информације олакшавајући интерпретацију резултата. У поређењу са стандардним главним компонентама које описују 87% укупне варијабилности података, главне компоненте добијене помоћу методе прорећених података објашњавају мању

Атрибути	PC1	PC2	PC3	PC4	PC5	PC6
topdiam	-0.477	0	0	0	0	0
length	-0.476	0	0	0	0	0
moist	0	0.785	0	0	0	0
testsg	0	0.619	0	0	0	0
ovensg	0.177	0	0.641	0	0	0
ringtop	0	0	0.589	0	0	0
ringbut	0.25	0	0.492	0	0	0
bowmax	-0.344	-0.021	0	0	0	0
bowdist	-0.416	0	0	0	0	0
whorls	-0.4	0	0	0	0	0
clear	0	0	0	-1	0	0
knots	0	0.013	0	0	-1	0
diaknot	0	0	-0.016	0	0	1
Прилагођена дисперзија	28.00%	13.97%	13.30%	7.40%	6.80%	6.20%

Слика 5.3: Оптерећења првих шест главних компоненти на прорећеним подацима скупа *Pitprops*

количину дисперзије (75.67% наспрам 86.98%) са знатно ређом структуром оптерећења. Важне променљиве повезане са шест главних компонената се скоро не преклапају, што додатно олакшава и чини тумачења јаснијим. Прва главна компонента додељује оптерећења променљивим *topdiam*, *length*, *ringbut*, *bowmax* и *wholrs*, пружајући општу меру величине; друга главна компонента додељује слична оптерећења атрибутима *moist* и *testsg* и мери степен сушења; трећа главна компонента, која додељује приближна оптерећења променљивим *ovensg*, *ringtop* и *ringbut*, узима у обзир брзину раста и чврстоћу дрвета; наредне три главне компоненте представљају *clear*, *knots* и *diaknot*, редом.

Ови резултати наглашавају предности и ограничења обе технике. Стандардна метода задржава све информације садржане у оригиналним атрибутима и не губи податке. Међутим, када оригинални атрибути имају високу димензионалност, резултујуће главне компоненте такође могу бити високе димензије. То може довести до веће димензије простора главних компоненти, што може отежати анализу и захтевати више рачунарских ресурса за њено извођење. Са друге стране, метода главних компоненти прорећених података омогућава компресију података и истицање битних карактеристика, али може изгубити мање значајне детаље. Избор између ових метода зависиће од специфичних потреба анализе и природе података. У ситуацијама где је важно истаћи суштинске карактеристике, метода главних компоненти прорећених података може пружити корисније резултате, док стандардна метода остаје важна када је очување свих информација приоритет.

Глава 6

Робусна анализа главних компоненти

Традиционална метода је осетљива на присуство аутлајера (енг. *outliers*) или аномалија у подацима, што може довести до искривљених и непоузданих резултата. Управо из тог разлога, уводи се концепт робусне анализе главних компоненти који се фокусира на решавање проблема аутлајера приликом примене анализе главних компоненти. Робусна анализа главних компоненти је модификација стандардне методе која је отпорнија на присуство аутлајера и има способност издвајања доминатне структуре чак и у присуству екстремних вредности. Ова техника је посебно корисна у ситуацијама када је важно задржати поуздану процену главних компоненти упркос присуству аутлајера.

Овај приступ користи методе као што су минимизација медијане апсолутног одступања (енг. *Median Absolute Deviation*) или минимизација l_1-норме за процену главних компоненти. Уместо максимизовања варијансе као што је случај са традиционалном методом, ова метода се фокусира на идентификацију атрибута који се мање мењају и пружају поузданију процену главних компоненти.

Даље ћемо представити концепт робусне методе главних компоненти и његову примену у смањењу димензије података. Приказаћемо кораке примене методе и анализирати њене предности и ограничења. Такође ћемо пружити увид у различите ситуације у којима је ова метода посебно корисна и размотрити њену ефикасност у поређењу са стандардном методом. Доказ главне теореме која ће бити касније наведена, као и примере примене можете погледати у раду [5].

Аутлајери

У контексту анализе података, изузеци се односе на запажања која се значајно разликују од типичног обрасца или дистрибуције скупа података. Утицај уклањања изузетка на резултате анализе може варирати у зависности од конкретне анализе која се спроводи. У неким случајевима, уклањање изузетка може имати значајан утицај на резултате, док у другим случајевима може бити безначајно. Запажања која, када се уклоне, имају велики утицај на резултате анализе називају се утицајна запажања. Док су већина утицајних запажања изузеци у неком погледу, изузеци не морају уопште бити утицајни.

Концепт утицаја је специфичан за сваку анализу. Запажање које је утицајно за једну врсту анализе или интересни параметар можда неће имати исти ниво утицаја у другој анализи или на други параметар. На пример, у анализи главних компоненти, запажања која су утицајна за одређивање коефицијената (оптерећења) главне компоненте можда неће нужно бити утицајна на дисперзију те главне компоненте, и обрнуто.

Мотивација

Претпоставимо да нам је дата велика матрица података M и знамо да се она може декомпоновати као

$$M = L_0 + S_0\,,$$

где L_0 има низак ранг, а S_0 је ретка; при томе, обе компоненте могу бити произвољне величине. Не знамо нискодимензиони простор колона и редова матрице L_0, чак ни њихову димензију. Слично томе, не знамо локације ненултих елемената матрице S_0, чак ни колико их има. Поставља се питање, можемо ли се надати тачном (можда чак и потпуном) и ефикасном опоравку и компоненте ниског ранга и ретке компоненте?

Једно решење које је доказано тачно и скалабилно за горенаведени проблем вероватно би имало утицај на данашња открића у научном истраживању које се ослања на велике количине података. Недавна експлозија огромних количина високодимензионих података у научним, инжењерским и друштвеним областима представља изазов, али и прилику за многе области као што су обрада слика, видео записа, мултимедијална обрада, анализа података о релевантности веба, претрага, биомедицинско снимање и биоинформатика.

У таквим доменима примене, подаци се рутински налазе у хиљадама или чак милијардама димензија, при чему број узорака понекад има сличан ред величине. Ова експанзија високодимензионих података захтева развој скалабилних и ефикасних метода за анализу и обраду.

Да бисмо ублажили проклетство димензионалности и скалирања, ослањамо се на најједноставнију и најкориснију претпоставку, а то је да се подаци налазе близу неког нискодимензионог потпростора. То значи да ако све тачке података сложимо као колоне матрице M, та матрица треба да има (приближно) низак ранг: математички,

$$M = L_0 + N_0 \,,$$

где L_0 има низак ранг, а N_0 је матрица малих шумова. Класична анализа главних компоненти тражи најбољу процену ранга k за L_0 (у смислу L_2 норме) решавањем

$$\begin{aligned} &\text{минимизовати } \|M - L\| \\ &\text{под условом да је } rang(L) \leq k \,. \end{aligned}$$

Овај проблем се може ефикасно решити помоћу декомпозиције сингуларних вредности и има неколико оптималних својстава када је шум N_0 мали и има расподелу независних једнако расподељених Гаусових вредности.

Робусна анализа главних компоненти

Метод главних компоненти је можда најчешће коришћени статистички алат за анализу података и редукцију димензионалности данас. Међутим, његова крхкост у погледу грубо оштећених посматрања често доводи у питање његову валидност - један грубо оштећен унос у матрици M може учинити процењени $\hat{L}$ произвољно удаљеним од стварне $\hat{L_0}$. Нажалост, грубе грешке су сада свеприсутне у модерним апликацијама попут обраде слика, анализе веб података и биоинформатике, где нека мерења могу бити произвољно оштећена (због преклапања, злонамерних манипулација или кварова сензора) или једноставно нису релевантна за нискодимензиону структуру коју покушавамо да идентификујемо. У литератури су током неколико деценија истраживања истражене и предложене бројне природне методе за побољшање робусности метода главних компоненти. Репрезентативни приступи укључују технике

функције утицаја, мултиваријантно одсецање, алтернативно минимизовање и техника случајног узорковања. Нажалост, ниједан од ових приступа не пружа полиномијално време извршавања са јаким гаранцијама перформанси под широким условима. Случајни приступи узорковања гарантују скоро оптималне процене, али имају експоненцијалну комплексност у рангу матрице L_0, док алгоритми за скраћивање имају релативно нижу рачунску сложеност, али гарантују само локално оптимална решења.

Нови проблем који овде разматрамо може се сматрати идеализованом верзијом робусне методе главних компоненти, у којој желимо да повратимо матрицу ниског ранга L_0 из високо оштећених мерења $M = L_0 + S_0$. За разлику од појма малог шума N_0 у класичном приступу, уноси у S_0 могу имати произвољно велику магнитуду, а претпоставља се да је њихов носач[1] редак, али непознат.

У односу на недавна истраживања (видети [6]) која су се усредсредила на проблем попуњавања матрице, непознати носач грешака чини проблем још тежим. Проблем попуњавања матрице се односи на ситуацију када имамо непотпуне или недостајуће вредности у матрици, али имамо информације о њеном облику и структури. Са друге стране, у случају робусне методе, поред присуства грешака, имамо и незнање о њиховој тачној локацији и расподели.

Примене. Постоји много важних примена у којима се подаци који се проучавају могу природно моделовати као ретки доприноси ниског ранга. Све статистичке апликације, у којима се траже робусне главне компоненте, наравно одговарају нашем моделу. У наставку дајемо примере инспирисане савременим изазовима у рачунарству и напомињемо да у зависности од примене, или компонента ниског ранга или ретка компонента могу бити предмет интересовања:

- *Видео надзор.* С обзиром на низ рамова за видео надзор, често морамо да идентификујемо активности које се издвајају из позадине. Ако сложимо видео оквире као колоне матрице M, онда компонента ниског ранга L_0 природно одговара стационарној позадини, а ретка компонента S_0 снима покретне објекте у првом плану. Међутим, сваки оквир слике има хиљаде или десетине хиљада пиксела, а сваки видео фрагмент садржи стотине или хиљаде оквира. Било би немогуће разложити матрицу M

[1] скуп индекса који представљају локације ненултих елемената у матрици S_0

на такав начин осим ако немамо истински скалабилно решење за овај проблем.

- *Детекција аномалија.* У задацима детекције аномалија, подаци често показују комбинацију регуларних образаца (компоненте ниског ранга) и ретких, абнормалних догађаја (ретке компоненте). Идентификацијом и сепарацијом ретке компоненте, можемо открити и означити аномалије у подацима.

- *Системи препорука.* У алгоритмима препорука, подаци о интеракцији корисника и ставки могу се декомпоновати у компоненту ниског ранга која приказује основне корисничке преференце и карактеристике ставки, и ретку компоненту која представља случајни шум или изузетке у подацима. Сепарацијом ових компоненти, можемо пружити тачније и персонализоване препоруке корисницима.

- *Компримирајуће узорковање.* У области обраде сигнала, компримирајуће узорковање има за циљ да опорави ретке сигнале из малог броја мерења. Ретка компонента има кључну улогу у реконструкцији оригиналног сигнала на основу ограниченог броја мерења.

- *Биоинформатика.* У анализи биолошких података, као што су подаци о експресији гена или мреже интеракција протеина, компонента ниског ранга може приказати заједничке обрасце или латентне структуре, док ретка компонента може представљати ретке генетске варијације или изузетке.

Ово су само неки примери примена у којима се подаци могу природно моделирати као комбинација компоненте ниског ранга и ретке компоненте. Зависно од конкретне примене, или компонента ниског ранга или ретка компонента могу бити предмет интереса.

Изненађење

На први поглед, проблем раздвајања се чини немогућим за решавање јер је број непознатих за закључивање о L_0 и S_0 дупло већи од броја датих мерења у матрици $M \in \mathbf{R}^{n_1 \times n_2}$. Штавише, чини се застрашујућим што очекујемо да

ћемо поуздано добити матрицу ниског ранга L_0 са грешкама у S_0 произвољно велике величине.

Испоставља се да се овај проблем може решити и помоћу *поправљиве конвексне оптимизације*. Нека $\|M\|_\star := \sum_i \sigma_i(M)$ означава нуклеарну норму матрице, односно збир сингуларних вредности те матрице, и нека $\|M\|_1 = \sum_{i,j} |M_{ij}|$ представља l_1-норму матрице M посматране као вектор у $\mathbf{R}^{n_1 \times n_2}$. Може се показати да под прилично слабим претпоставкама, процена *Principal Component Pursuit* која решава следећи проблем

$$\begin{aligned} &\text{минимизовати } \|L\|_* + \lambda \|S\|_1 \\ &\text{под условом да је } L + S = M\,, \end{aligned} \tag{6.1}$$

тачно враћа L_0 ниског ранга и ретку матрицу S_0. Теоријски, ово гарантовано функционише чак и ако се ранг матрице L_0 готово линеарно повећава са димензијом матрице, а грешке у S_0 су до константног дела свих уноса. Алгоритамски, претходно наведени проблем се може решити ефикасним и скалабилним алгоритмима, по цени не толико већој од класичног метода главних компоненти. Емпиријски, симулације и експерименти указују да ово функционише под изненађујуће широким условима за разне врсте стварних података.

Када раздвајање има смисла?

Чини се да нема довољно информација да би се савршено раздвојиле компоненте ниског ранга и ретке компоненте. И заиста, у овоме има неке истине, пошто очигледно постоји проблем идентификације. На пример, претпоставимо да је матрица M једнака $e_1 e_1^T$ (ова матрица има јединицу у горњем левом углу и нуле свуда другде). Онда пошто је M и ретка и ниског ранга, како можемо одлучити шта је од ово двоје? Да би проблем био смислен, морамо наметнути да компонента ниског ранга L_0 није ретка. За проблем комплетирања матрице користићемо општи појам некохерентности (ово је претпоставка која се тиче сингуларних вектора компоненте нижег ранга). Записаћемо декомпозицију сингуларне вредности матрице $L_0 \in \mathbf{R}^{n_1 \times n_2}$ као

$$L_0 = U \Sigma V^T = \sum_{i=1}^{r} \sigma_i u_i v_i^T\,,$$

где је r ранг матрице, $\sigma_1, ..., \sigma_r$ су позитивне сингуларне вредности, а $U = [u_1, ..., u_r]$ и $V = [v_1, ..., v_r]$ су матрице левих и десних сингуларних вектора.

Тада услов некохерентности са параметром μ изражава

$$\max_i \|U^T e_i\|^2 \leq \frac{\mu r}{n_1}\,,\ \max_i \|V^T e_i\|^2 \leq \frac{\mu r}{n_2} \tag{6.2}$$

и

$$\|UV^T\|_\infty \leq \sqrt{\frac{\mu r}{n_1 n_2}}\,. \tag{6.3}$$

Овде и у даљем тексту, $\|M\|_\infty = \max_{i,j} |M_{ij}|$, то јест представља l_∞ норму матрице M посматране као дуги вектор. Пошто је ортогонална пројекција P_U на простор колона U дата са $P_U = UU^T$, онда је (5.2) еквивалентно са $\max_i \|P_U e_i\|^2 \leq \frac{\mu r}{n_1}$, и слично за P_V. Услов некохерентности тврди да су за мале вредности μ, сингуларни вектори разумно распоређени – другим речима, нису ретки.

Још један проблем у вези са идентификацијом јавља се ако ретка матрица има низак ранг. Ово ће се десити ако се, рецимо, сви уноси различити од нуле налазе у колони или у неколико колона. Претпоставимо, на пример, да је прва колона од S_0 супротна првој колони матрице L_0, а да све остале колоне од S_0 нестају. Тада је јасно да не бисмо могли да повратимо L_0 и S_0 ниједном методом јер би $M = L_0 + S_0$ имао простор колона једнак, или укључен у простор колона матрице L_0. Да бисмо избегли такве бесмислене ситуације, претпоставићемо да је образац реткости ретке компоненте изабран равномерно насумично.

Главни резултат

Изненађујуће је да под овим минималним претпоставкама, *PCP* (енг. *Principal Component Pursuit*) решење савршено враћа компоненте ниског ранга и ретке компоненте, под условом, наравно, да ранг компоненте ниског ранга није превелик, и да је ретка компонента разумно ретка. У наставку, $n_{(1)} = \max(n_1, n_2)$ и $n_{(2)} = \min(n_1, n_2)$.

Теорема 6.1. *Нека је L_0 матрица формата $n \times n$ која испуњава услове* (5.2) *и* (5.3)*, и нека је скуп подршке матрице S_0 униформно распоређен међу свим скуповима кардиналности m. Тада постоји константа c таква да са вероватноћом већом од $1 - cn^{-10}$ (при избору подршке за S_0) PCP* (5.1) *са $\lambda = \frac{1}{\sqrt{n}}$ је тачна, односно важи $\hat{L} = L_0$ и $\hat{S} = S_0$ под условом да је*

$$rang(S_0) \leq \rho_r n \mu^{-1} (\log(n))^{-2} \ \text{ и } \ m \leq \rho_s n^2\,,$$

где су ρ_r и ρ_s позитивне константе. У општем правоугаоном случају где је L_0 формата $n_1 \times n_2$, PCP са $\lambda = \frac{1}{\sqrt{n_{(1)}}}$ успева са вероватноћом већом од $1 - cn_{(1)}^{-10}$, под условом да $rang(L_0) \leq \rho_r n_{(2)} \mu^{-1} (\log n_{(1)})^{-2}$ и $m \leq \rho_s n_1 n_2$.

Другим речима, матрице L_0 чији су сингуларни вектори — или главне компоненте — разумно распоређени, могу се повратити са вероватноћом скоро један из произвољних и потпуно непознатих образаца корупције (све док су они насумично распоређени). У ствари, ово функционише за велике вредности ранга, тј. реда $\frac{n}{(\log n)^2}$ када μ није превелико. Треба нагласити да се једина „случајност" у нашим претпоставкама односи на локације ненултих уноса матрице S_0, све остало је детерминистичко. Конкретно, све што захтевамо у вези са L_0 је да његови сингуларни вектори нису шиљасти. Такође, не правимо никакве претпоставке о величинама или знацима ненултих уноса матрице S_0. Да бисмо избегли било какву двосмисленост, модел за S_0 је следећи: узмемо произвољну матрицу S и поставимо на нулу њене уносе на случајном скупу Ω^c, то даје S_0.

Прилично изузетна чињеница је да у алгоритму не постоји подешавајући параметар. Под претпоставкама из теореме, минимизација

$$\|L\|_* + \frac{1}{\sqrt{n_{(1)}}} \|S\|_1 \, , \, n_{(1)} = \max(n_1, n_2)$$

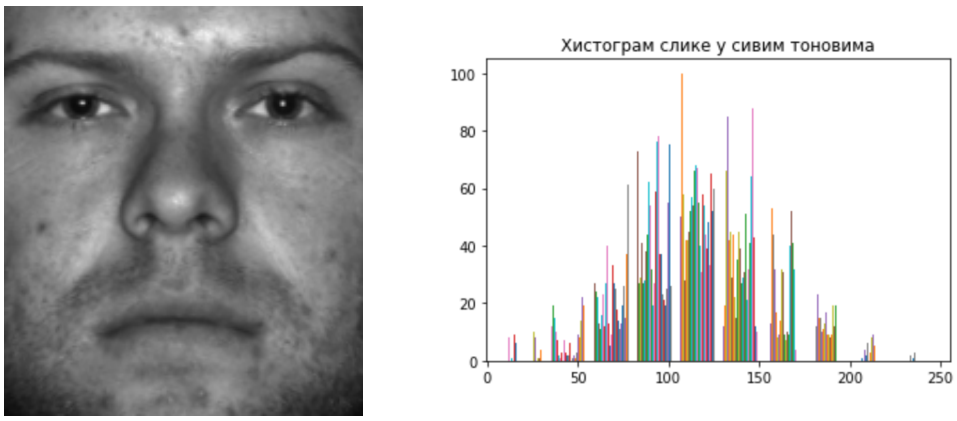
увек враћа тачан одговор. Ово је изненађујуће јер би се могло очекивати да треба изабрати прави скалар λ како би се на одговарајући начин избалансирала два члана у $\|L\|_* + \lambda \|S\|_1$ (можда у зависности од њиховог релативног односа). То, међутим, очигледно није случај. У том смислу, избор $\lambda = \frac{1}{\sqrt{n_{(1)}}}$ је универзалан. Даље, није априори јасно зашто је $\lambda = \frac{1}{\sqrt{n_{(1)}}}$ исправан избор без обзира на то како изгледају L_0 и S_0. Математичка анализа открива исправност ове вредности. У ствари, доказ теореме даје читав низ исправних вредности, а ми можемо одабрати довољно једноставну вредност у том опсегу.

Други коментар је да се могу постићи резултати са већим вероватноћама успеха, тј. облика $1 - O(n^{-\beta})$ (или $1 - O(n_{(1)}^{-\beta})$) за $\beta > 0$, уз смањењење вредности ρ_r.

6.1 Пример

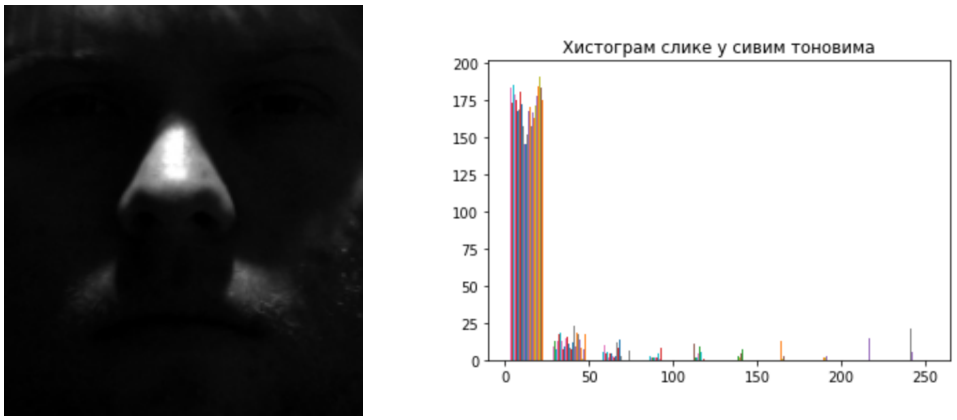
За овај пример ћемо користити слике из базе *Yale Face Database B*, која садржи 5760 слика сиве боје са једним извором светлости, које приказују 10

субјеката од којих се сваки види под 576 услова гледања. Рачунари нису као ми и немогу разумети слике као што ми то чинимо. Међутим, они разумеју матрице и ми можемо користити матрицу са различитим вредностима интезитета за сваки пиксел како бисмо представили црно-белу слику. Слике у боји имају још три додатна канала као што су RGB или неки други стандард. Ми ћемо радити са прве четири слике из фолдера *yaleB01*. Све слике су поравнате, и свака слика има димензију 168×192 или 32256 пиксела. Вредности могу варирати између 0 и 255, од тамнијих ка светлијим тоновима. Колико је слика тамна или светла, можемо сазнати посматрајући њен хистограм.



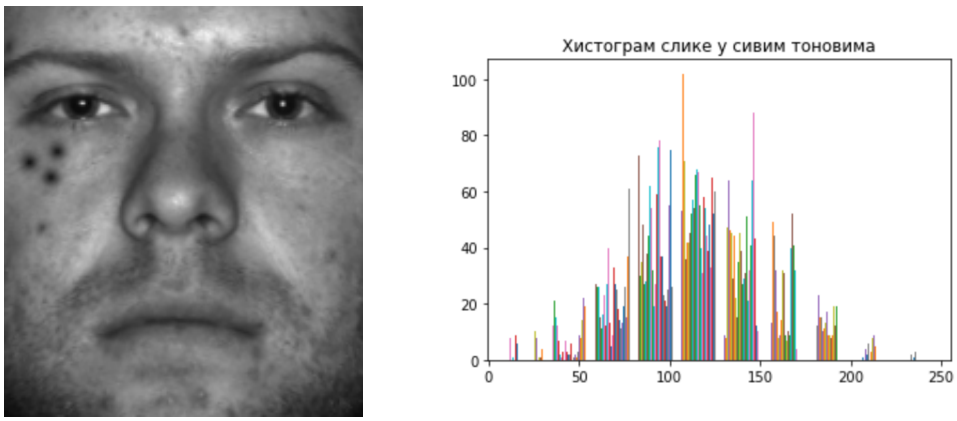
Слика 6.1: Прва слика из фолдера и њен хистограм

Као што видимо са слике 6.1, ова слика није превише тамна нити светла, већ је негде између, односно средње осветљена. Ако погледамо слику 6.2, видимо да овај хистограм има много више вредности близу нуле, што је очигледно тамније како можемо видети.



Слика 6.2: Четврта слика из фолдера и њен хистограм

Коришћење робусне анализе главних компоненти за откривање изузетака. Изабраћемо прву слику и на њу ћемо додати неке изузетке. Ова слика има мање веома тамних и веома светлих вредности. Хајде да додамо 3 црне сузе, за шта можемо користити разне уређиваче слика. Нова слика и њен хистограм приказани су на слици 6.3. Сада ћемо на ову слику да примениму робусну анализу главних компоненти и да видимо колико добро се носи са изузецима. Параметар λ, који у једнакости (6.1) представља проценат ретких грешака, има висок утицај на репрезентације ниског ранга. Иако га можемо израчунати користећи Теорему 6.1, то није оптимално. У пракси, иницијализујемо λ помоћу Теореме 6.1, а затим га постепено повећавамо или смањујемо корак по корак у интервалима од 0.005 како бисмо изабрали најбољу вредност. У овом примеру, ми ћемо се здражати на вредности коју нам даје Теорема 6.1.



Слика 6.3: Прва слика са изузецима и њен хистограм

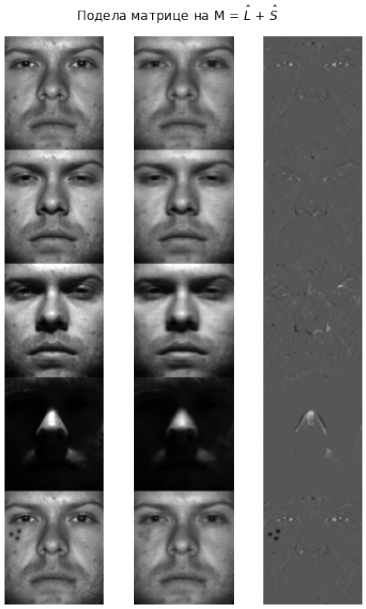
На слици 6.4, примећује се значајно побољшање, иако није савршено. Међутим, поставља се питање како би резултати изгледали да имамо знатно више података о стварној дистрибуцији? Ако применимо робусну анализу главних компоненти на скуп који садржи прве четири слике из скупа *yaleB01*, посматрајући слику 6.5 можемо да закључимо да додатни подаци побољшавају перформансе овог приступа, омогућавајући боље препознавање и обраду изузетака на слици.

На слици 6.6 можемо да видимо како су високодимензиони подаци подељени на линеарну комбинацију две матрице. Приказана је матрица ниског ранга $\hat{L}$ и ретка матрица $\hat{S}$, добијене као решење конвексног програма. Прва колона приказује оригиналне слике са променама осветљења и израза лица. Друга

Слика 6.4: Резултати након примене робусне анализе главних компоненти на основу полазне слике

Слика 6.5: Резултати након примене робусне анализе главних компоненти на основу све четири слике

колона приказује ниске рангове и приближне слике оригиналних слика. Последња, трећа колона приказује ретке слике грешака оригиналних слика, које представљају разлику између оригиналне и приближне слике.



Слика 6.6: Пример примене робусне анализе главних компоненти на *Yale B* бази лица

Након примене робусне анализе главних компоненти, примећујемо да су три црне тачке постале блеђе на матрици ниског ранга. Ова промена у осветљењу указује на то да је метод успешно препознао и издвојио ове аутлајере као нискофреквентне компоненте, што је карактеристично за аутлајере који

се не придржавају облика и варијација главних компоненти. Овај резултат потврђује способност методе да идентификује неправилности у подацима чак и када се оне вештачки додају. Наглашава се његова робусност у детекцији аутлајера који се разликују од остатка података, чиме се омогућава боље разумевање структуре података и боља интерпретација њихових карактеристика.

У будућим истраживањима, могуће је размотрити даље прилагођавање параметара робусне методе и анализирати различите врсте аутлајера како би се дубље истражила ефикасност овог приступа. Такође, могуће је комбиновати робусну анализу главних компоненти са другим техникама за додатну унапређеност детекције и обраде аутлајера. У целини, примена робусне методе представља корак напред у разумевању и руковању изузецима у скупу података, пружајући корисне увиде за даљу анализу и доношење информисаних одлука.

Глава 7

Закључак

У овом раду, дубоко смо истражили методе смањења димензије темељене на анализи главних компоненти. Наша анализа је указала на кључну улогу коју анализа главних компоненти и сродне технике играју у обради и интерпретацији високодимензионих података. Кроз имплементацију и експерименте, успели смо да размотримо њихову ефикасност и практичну применљивост у различитим контекстима.

Наша истраживања показују да стандардна метода пружа моћан механизам за трансформацију простора података, омогућавајући нам да идентификујемо доминантне обрасце варијације међу атрибутима. Овај процес смањења димензије не само да олакшава визуализацију података, већ и помаже у суштинском разумевању структуре података. Осим тога, размотрили смо и варијације традиционалне методе, као што су анализа главних компоненти са кернел трансформацијама, анализа главних компоненти проређених података и робусна анализа главних компоненти. Метод са кернел трансформацијама нас је провео кроз врата нелинеарних простора, омогућавајући анализу сложених веза међу подацима. Анализа главних компоненти проређених података је осветлила пут према ретким структурама и бољој интерпретацији, док је робусна анализа истакла своју вредност у откривању образаца упркос присуству шума. Свака од ових варијација доноси своје предности и компромисе, што их чини адекватним изборима за различите типове података и аналитичке циљеве.

Наши експерименти су потврдили да одговарајући избор параметара и пажљиво препознавање претпоставки о подацима играју кључну улогу у постизању оптималних резултата. Такође смо приметили да је важно постићи

баланс између смањења димензије и задржавања што више информација. Док смањење димензије може поједноставити анализу, треба бити опрезан како не бисмо изгубили битне аспекте података.

Кроз ово истраживање, схватили смо да нема универзалног приступа који би био најбољи за све ситуације. Избор технике зависи од природе података и циљева анализе. Интеграција ових техника или истраживање нових варијанти може отворити врата ка још дубљем разумевању података и бољем доношењу одлука.

На основу ових сазнања, будући рад може се фокусирати на даљу оптимизацију ових техника или на њихову комбинацију са другим методама анализе података. Без обзира на то, закључујемо да су анализа главних компоненти и њене модификације кључне компоненте у области обраде података, које обећавају наставак иновација у пољу истраживања података и статистичког учења.

Библиографија

[1] Two case studies in the application of principal component analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 16(3):225–236, 1967.

[2] Методе за смањење димензије података засноване на анализи главних компоненти. `https://github.com/JelenaRadojevic/master_primeri`, 2023.

[3] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

[4] Theodore Wilbur Anderson, Theodore Wilbur Anderson, Theodore Wilbur Anderson, and Theodore Wilbur Anderson. *An introduction to multivariate statistical analysis*, volume 2. Wiley New York, 1958.

[5] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37, 2011.

[6] Emmanuel J Candes and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

[7] Raymond B Cattell. The scree test for the number of factors. *Multivariate behavioral research*, 1(2):245–276, 1966.

[8] Rosember Guerra-Urzola, Katrijn Van Deun, Juan C Vera, and Klaas Sijtsma. A guide for sparse pca: model comparison and applications. *psychometrika*, 86(4):893–919, 2021.

[9] Heiko Hoffmann. Kernel pca for novelty detection. *Pattern recognition*, 40(3):863–874, 2007.

[10] Iain M Johnstone and Arthur Yu Lu. Sparse principal components analysis, 2009.

[11] Ian T Jolliffe. *Principal component analysis for special types of data*. Springer, 2002.

[12] Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

[13] Ian T Jolliffe, Nickolay T Trendafilov, and Mudassir Uddin. A modified principal component technique based on the lasso. *Journal of computational and Graphical Statistics*, 12(3):531–547, 2003.

[14] Henry F Kaiser. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3):187–200, 1958.

[15] George P McCabe. Principal variables. *Technometrics*, 26(2):137–144, 1984.

[16] Mladen Nikolić and Anđelka Zečević. Mašinsko učenje. *Beograd: Matematički fakultet*, 2019.

[17] Sebastian Raschka, P Linear, R Gaussian, and Locally-Linear Embedding LLE. Kernel tricks and nonlinear dimensionality reduction via rbf kernel pca. *Blog, September*, 2014.

[18] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Kernel principal component analysis. In *International conference on artificial neural networks*, pages 583–588. Springer, 1997.

[19] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319, 1998.

[20] Carlos Oscar Sánchez Sorzano, Javier Vargas, and A Pascual Montano. A survey of dimensionality reduction techniques. *arXiv preprint arXiv:1403.2877*, 2014.

[21] Quan Wang. Kernel principal component analysis and its applications in face recognition and active shape models. *arXiv preprint arXiv:1207.3538*, 2012.

[22] Huan Xu, Constantine Caramanis, and Sujay Sanghavi. Robust pca via outlier pursuit. *Advances in neural information processing systems*, 23, 2010.

[23] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 2005.

[24] Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.

Биографија аутора

Јелена Радојевић је рођена 19. мај 1998. године у Ужицу. Ту је завршила и основну школу „Душан Јерковић“. Школовање је наставила у Ужичкој гимназији на природно-математичком смеру. Гимназију завршава 2017. године као носилац дипломе „Вук Караџић“. Исте године уписује Математички факултет Универзитета у Београду, на смеру Статистика, актуарска и финансијска математика. Основне академске студије завршава у септембру 2021. године са просечном оценом 9.50, а у октобру исте године уписује и мастер студије на истом смеру.