УНИВЕРЗИТЕТ У БЕОГРАДУ
МАТЕМАТИЧКИ ФАКУЛТЕТ

Бојана Ђука

МЕШАВИНЕ ВИШЕДИМЕНЗИОНИХ НОРМАЛНИХ РАСПОДЕЛА И ЊИХОВА ПРИМЕНА У КЛАСТЕРОВАЊУ ПОДАТАКА

мастер рад

Београд, 2024.

Ментор:

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

Чланови комисије:

др Марко Обрадовић, доцент
Универзитет у Београду, Математички факултет

др Младен Николић, ванредни професор
Универзитет у Београду, Математички факултет

Датум одбране: 29.2.2024.

Садржај

Глава 1

Увод

У данашње време скоро да се не може наћи област у коју машинско учење није успело да продре и нађе неку примену. Кола која имају могућност аутономне вожње или наочаре за виртуелну реалност, само су неки од примера који приказују колико далеко је развој технологије одмакао и неоспориво приказују распрострањеност машинског учења у различитим сферама. Сваки од модела машинског учења заснива се на употреби података због чега је, за добијање што бољих резултата, изразито важно да подаци које поседујемо буду тачни и квалитетни, као и да их имамо у што већој мери. Све ово довело је до велике потребе за прикупљањем и складиштењем истих, услед чега можемо наићи на одређене препреке уколико, на пример, имамо ограничен меморијски простор за чување података.

Кластеровање је техника груписања података која је доста помогла у решавању овог проблема. Задатак кластеровања је подела података на групе тако да инстанце једног кластера буду што сличније међу собом, а што различитије уколико се ради о инстанцама различитих група. На овај начин, кластеровањем запажамо битне шаблоне у подацима што нам омогућава да чувамо мањи број података при том задржавајући главне законитости које карактеришу саме податке.

Захтевност у складиштењу података умножава се уколико се сусрећемо са задацима који се односе на коришћење звука или слике. Уколико сваки пиксел на слици посматрамо као један податак, употребом кластеровања можемо груписати пикселе различитих боја. Резултат кластеровања може се користити за одбацивање сувишних информације. На пример, како се не би чували подаци о боји сваког пиксела, могуће је све пикселе из једног кластера представити

истом бојом. Занемаривање непотребних података нам омогућава да значајно смањимо меморију потребну за чување слике, док и даље задржавамо основне карактеристике оригинала.

Ово је само једна од употреби кластеровања, а у наставку рада упознаћемо се и са другим ситуацијама када груписање података представља добар избор за рад с подацима. Такође, у овом раду биће описан један од најефикаснијих начина за кластеровање података, а његово функционисање заснива се на употреби мешавина вишедимензионих нормалних расподела.

Помињања мешавина расподела датирају од друге половине 19. века, међутим први значајни рад који је укључивао мешавину нормалних расподела био је рад Карла Пирсона [8], који је 1894. године ову расподелу искористио за описивање података о фамилијама ракова. Његов рад обухватао је мешавину две нормалне расподеле са различитим средњим вредностима и дисперзијама.

Тек шездесетих година 20. века посветила се пажња методу максималне веродостојности за оцењивање параметара мешавине и написани су радови који су допринели развијању овог приступа. Десетак година касније, 1977. године, у раду [1] Артур Демпстер искористио је метод максималне веродостијности за формализацију EM алгоритма који представља итеративни процес за одређивање параметара мешавине нормалних расподела. Након појаве овог алгоритма интересовање за мешавине нормалних расподела нагло је скочило, а са тим и број написаних радова везаних за ову тему.

Мешавине нормалних расподела почеле су се користити за кластеровање, а с развојем метода за одабир модела побољшале су се и методе за бирање броја кластера, што је чинило овај начин за груписање података све бољим и софистицираније.

У овом раду почећемо од упознавања вишедимензионих нормалних расподела, чија је употреба јако важна у раду са подацима који су смештени у више димензија. Концепт вишедимензионих нормалних расподела проширићемо на употребу њихових мешавина које имају кључну улогу у подели података на групе. Како бисмо што боље разумели ефикасност и примену мешавина нормалних расподела, објаснићемо идеју кластеровања података и одговорити на питање зашто је ово важан задатак у раду са стварним подацима. Осврнућемо се и на алгоритам К средина са крајњим циљем упоређивања перфомански два алгоритма за кластеровање. Након тога, фокус ће бити на представљању кластеровања коришћењем мешавина вишедимензионих нормалних расподела

и илустрацији рада овог начина кластеровања. Како је важан задатак у коришћењу мешавина оцењивање њихових параметара, биће уведен и описан EM алгоритам, како у случају мешавина, тако и у општем случају. Приказаћемо способност мешавина да кластерују податке сложенијег облика, као и поређење са алгоритмом K средина. Поред наведених тема, биће речи и о методама за бирање број кластера у подацима, као што су правило лакта, Акаикеов и Бајесов информациони критеријум.

Циљ рада је да читаоцу пружи теоријско разумевање мешавина нормалних расподела и скрене пажњу на њихову широку примену у машинском учењу и анализи података. У складу са тим, на крају рада биће приказани стварни примери употребе поменутог начина кластеровања који ће читаоцу омогућити да из другог угла сагледа моћ мешавина нормалних расподела.

Глава 2

Вишедимензиона нормална расподела и њене особине

Приликом решавања задатака који се баве обрадом и анализом података или моделовањем, нереално је очекивати да подаци поседују само једну варијаблу. У већини случајева, приликом прикупљања података, у било које сврхе, битно је забележити што више информација како би се посматрана појава што боље описала, а у обзир узео већи број различитих фактора који могу утицати на њено понашање. На пример, уколико бисмо желили да предвидимо колика је цена стана, није довољно да посматрамо само локацију стана. Иако никада не можемо прикупити све факторе у вези са овим проблемом, резултати предвиђања ће бити квалитетнији уколико укључимо и информације о старости стана, спрату на ком се налази, потреби за реновирањем... Из тог разлога јако је битно концепт једнодимензионих података проширити на више димензија.

Вишедимензиона нормална расподела представља уопштење нормалне расподеле у више димензија. Свака тачка ове расподеле је задата као вектор у посматраном n димензионом простору.

За случајан вектор $X = (X_1, ...X_n)^T$ кажемо да има вишедимензиону нормалну расподелу са средњом вредношћу $\mu = (\mu_1, ..., \mu_n)^T \in \boldsymbol{R^n}$ и симетричном, позитивно дефинитном коваријационом матрицом Σ димензије $n \times n$ ако је његова густина расподеле дата са:

$$f(x|\mu, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} \exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)), \tag{2.1}$$

ознака $|\Sigma|$ представља детерминанту матрице Σ, а израз $(x-\mu)^T\Sigma^{-1}(x-\mu)$ зовемо квадратном формом густине. Када желимо да назначимо да наша промен-

љива има вишедимензиону нормалну расподелу, користимо озанаку $X_{\mathcal{N}}(\mu, \Sigma)$.

Вишедимензиону нормалну расподелу чији је вектор средњих вредности једнак нула вектору, а коваријациона матрица јединична, називамо стандардна вишедимезиона нормална расподела.

Коваријациона матрица и њене особине

Коваријациона матрица Σ описује везу између компоненти расподеле, а њени елементи су задати формулом:

$$\Sigma_{ij} = Cov(X_i, X_j) = E[(X_i - E[X_i])(X_j - E[X_j])], \forall i, j \in \{1, ..., n\}.$$

С обзиром да важе следећи услови:

- $E(X_i) = \mu_i \ \forall i \in \{1, ..., n\}$,
- $\Sigma_{k,k} = E[(X_k - E[X_k])^2] = D(X_k), \ \forall k \in \{1, ..., n\}$,
- $\Sigma_{i,j} = Cov(X_i, X_j) = Cov(X_j, X_i) = \Sigma_{j,i}, \ \forall i, j \in \{1, ..., n\}$ и $i \neq j$,

можемо приметити да коваријациона матрица има следећи облик:

$$\Sigma = \begin{pmatrix} D(X_1) & Cov(X_1, X_2) & \dots & Cov(X_1, X_n) \\ Cov(X_1, X_2) & D(X_2) & \dots & Cov(X_2, X_n) \\ \vdots & \vdots & \dots & \vdots \\ Cov(X_1, X_n) & Cov(X_2, X_n) & \dots & D(X_n) \end{pmatrix}.$$

Дефиниција 2.0.1. *За матрицу A кажемо да је симетрична уколико је једнака свом транспонату, односно $A = A^T$.*

Дефиниција 2.0.2. *Симетрична матрица A димензије $n \times n$ је позитивно семидефинитна уколико за сваки реалан вектор a димензије $n \times n$ важи да је $a^T Aa \geq 0$.*

Дефиниција 2.0.3. *Симетрична матрица A димензије $n \times n$ је позитивно дефинитна уколико за сваки реалан вектор a димензије $n \times n$ важи да је $a^T Aa > 0$.*

Теорема 2.0.1. *Свака коваријациона матрица Σ произвољног случајног вектора X са очекивањем μ је симетрична и позитивно дефинитна.*

Доказ. *Симетричност следи из претходно истакнутих особина коваријације и може се уочити на приказу коваријационе матрице. За посматрани вектор X коваријациона матрица је дата са $\Sigma = E[(X-\mu)(X-\mu)^T]$, па за произвољан реални вектор $a \in \boldsymbol{R^n}$ важи:*

$$\begin{aligned} a^T\Sigma a &= a^T[E(X-\mu)(X-\mu)^T]a \\ &= E[a^T(X-\mu)(X-\mu)^T a] \\ &= E\left(((X-\mu)^T a)^T((X-\mu)^T a)\right) \\ &= E\left(((X-\mu)^T a)^2\right) \geq 0. \end{aligned}$$

Из претходне теореме можемо закључити, да би коваријациона матрица била исправно задата мора бити симетрична и позитивно семидефинитна, међутим у дефиницији вишедимензионе нормалне расподеле примећујемо да коваријациона матрица мора бити позитивно дефинитна. Овај додатни услов проистиче из захтева да, у случају вишедимензионих нормалних расподела, коваријациона матрица мора имати инверз.

Особине вишедимензионе нормалне расподеле

Теорема 2.0.2. *Нека је Z случајан вектор са стандардном вишедимензионом нормалном расподелом. Ако је Σ инвертибилна матрица димензије $n \times n$ и μ вектор димензије $n \times 1$ тада случајан вектор $X = \mu + \Sigma Z$ има вишедимензиону нормалну расподелу са параметрима μ и $\Sigma\Sigma^T$, односно, $X \sim f_{\mathcal{N}}(\mu, \Sigma\Sigma^T)$.*

Доказ. *Проверимо прво да ли $\Sigma\Sigma^T$ може представљати коваријациону матрицу. Из једнакости $(\Sigma\Sigma^T)^T = (\Sigma^T)^T\Sigma^T = \Sigma\Sigma^T$, можемо закључити да $\Sigma\Sigma^T$ задовољава услов симетричности. Такође, за произвољан реалан вектор a важи:*

$$a^T\Sigma\Sigma^T a = (\Sigma^T a)^T(\Sigma^T a) > 0,$$

стога можемо тврдити да је $\Sigma\Sigma^T$ позитивно дефинитна матрица.

Посматрамо трансформацију $z = \Sigma^{-1}(x-\mu)$. Јакобијева матрица ове трансформације дата је са:

$$J = \frac{\partial z}{\partial x} = \Sigma^{-1}.$$

Коришћењем дате трансформације, Јакобијана и чињенице да Z има стандардну нормалну расподелу, густину за X рачунамо на следећи начин:

$$\begin{aligned}
&f_X(x) = f_Z(\Sigma^{-1}(x-\mu))abs(|J|) = f_Z(\Sigma^{-1}(x-\mu))(abs|\Sigma|)^{-1} \\
&= \frac{1}{|abs(\Sigma)|}(2\pi)^{-\frac{n}{2}} exp\left(-\frac{1}{2}(\Sigma^{-1}(x-\mu))^T(\Sigma^{-1}(x-\mu))\right) \\
&= (abs|\Sigma|)^{-\frac{1}{2}}(abs|\Sigma|)^{-\frac{1}{2}}(2\pi)^{-\frac{n}{2}} exp\left(-\frac{1}{2}(x-\mu)^T(\Sigma^{-1})^T\Sigma^{-1}(x-\mu)\right) \\
&= (2\pi)^{-\frac{n}{2}}(abs(|\Sigma||\Sigma|))^{-\frac{1}{2}} exp\left(-\frac{1}{2}(x-\mu)^T(\Sigma^T)^{-1}\Sigma^{-1}(x-\mu)\right) \\
&= (2\pi)^{-\frac{n}{2}}(|\Sigma||\Sigma^T|)^{-\frac{1}{2}} exp\left(-\frac{1}{2}(x-\mu)^T(\Sigma^T\Sigma)^{-1}(x-\mu)\right) \\
&= (2\pi)^{-\frac{n}{2}}|\Sigma\Sigma^T|^{-\frac{1}{2}} exp\left(-\frac{1}{2}(x-\mu)^T(\Sigma^T\Sigma)^{-1}(x-\mu)\right).
\end{aligned}$$

Шеста једнакост користила је чињеницу да је детерминанта квадратне матрице једнака детерминанти њеног транспоната. Овим је показано да X има тражену густину, односно да важи $X \sim f_{\mathcal{N}}(\mu, \Sigma\Sigma^T)$.

Теорема 2.0.3. *Нека је $X = (X_1, ...X_n)^T$ случајан вектор такав да $X_{\mathcal{N}}(\mu, \Sigma)$ и A инвертибилна матрица димензије $n \times n$, тада за $Y = AX$ важи да $Y_{\mathcal{N}}(A\mu, A\Sigma A^T)$.*

Доказ. *Густину за Y ћемо одредити из густине $f(x|\mu, \Sigma)$ користећи смену $x = A^{-1}y$. Израчунајмо прво апсолутну вредност детерминанте Јакобијеве матрице ове трансформације.*

$$abs(|J|) = abs(|\frac{\partial x}{\partial y}|) = abs(|A^{-1}|) = \frac{1}{abs(|A|)} =$$
$$\sqrt{\frac{1}{|A|^2}} = \sqrt{\frac{|\Sigma|}{|A| \cdot |\Sigma| \cdot |A^T|}} = \frac{|\Sigma|^{1/2}}{|A\Sigma A^T|^{1/2}}.$$

Квадратна форма густине $f(x|\mu, \Sigma)$ након смене је дата са:

$$\begin{aligned}
&(x-\mu)^T\Sigma^{-1}(x-\mu) = \\
&(A^{-1}y-\mu)^T\Sigma^{-1}(A^{-1}y-\mu) = \\
&(A^{-1}y-A^{-1}A\mu)^T\Sigma^{-1}(A^{-1}y-A^{-1}A\mu) = \\
&(A^{-1}(y-A\mu))^T\Sigma^{-1}(A^{-1}(y-A\mu)) = \\
&(y-A\mu)^T(A^{-1})^T\Sigma^{-1}A^{-1}(y-A\mu) = \\
&(y-A\mu)^T(A\Sigma A^T)^{-1}(y-A\mu)
\end{aligned}$$

Коришћењем претходно изведено, добијамо да је густина f_y од Y дата са:

$$f_y(y) = f(A^{-1}y|\mu,\Sigma)abs(|J|) =$$
$$(2\pi)^{-\frac{n}{2}}|A\Sigma A^T|^{-\frac{1}{2}}\exp(-\frac{1}{2}(y-A\mu)^T(A\Sigma A^T)^{-1}(y-A\mu)) = f(y|A\mu,A^T).$$

Наредна теорема биће наведена без доказа и користиће као испомоћ за доказивање преосталих теорама у вези са особинама вишедимензионе нормалне расподеле.

Теорема 2.0.4. *Нека је R инвертибилна матрица коју заједно са њеним инверзом, R^{-1}, можемо поделити у 2×2 блокове на следећи начин:*

$$R = \begin{pmatrix} A & B \\ C & D \end{pmatrix}, \qquad R^{-1} = \begin{pmatrix} E & F \\ G & H \end{pmatrix}.$$

Ако су A,D и R инвертибилне матрице тада важи:

$$R^{-1} = \begin{pmatrix} A^{-1}+A^{-1}B(D-CA^{-1}B)^{-1}CA^{-1} & -A^{-1}B(D-CA^{-1}B)^{-1} \\ -(D-CA^{-1}B)^{-1}CA^{-1} & D^{-1}+D^{-1}C(A-BD^{-1}C)^{-1}BD^{-1} \end{pmatrix}.$$

Нека је $X = (X_1,...X_n)^T$ и $X_{\mathcal{N}}(\mu,\Sigma)$. Посматрајмо случајне величине $X_1,...,X_k$ и $X_{k+1},...,X_n$ које формирају случајне векторе димензије k и $m = n-k$, редом:

$$X^{(1)} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{pmatrix}, \qquad X^{(2)} = \begin{pmatrix} X_{k+1} \\ X_{k+2} \\ \vdots \\ X_n \end{pmatrix}.$$

Означимо сада очекивања, $E(X^{(1)}) = \mu^{(1)}$, $E(X^{(2)}) = \mu^{(2)}$, и коваријације са

$$E(X^{(1)}-\mu^{(1)})(X^{(1)}-\mu^{(1)})^T = \Sigma_{11},$$
$$E(X^{(2)}-\mu^{(2)})(X^{(2)}-\mu^{(2)})^T = \Sigma_{22},$$
$$E(X^{(1)}-\mu^{(1)})(X^{(2)}-\mu^{(2)})^T = \Sigma_{12}.$$

Можемо рећи да је вектор X подељен и да важи следеће:

$$X = \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix}, \qquad \mu = \begin{pmatrix} \mu^{(1)} \\ \mu^{(2)} \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

Теорема 2.0.5. *Маргинална расподела вектора $X^{(1)}$ димензије k је вишедимензиона нормална расподела са параметрима $\mu^{(1)}$ и Σ_{11} , односно важи $X^{(1)}_{\mathcal{N}}(\mu^{(1)}, \Sigma_{11})$.*

Доказ. *Густина расподеле за X и заједничка густина за $X^{(1)}$ и $X^{(2)}$ је дата са:*

$$f(x) = f(x_1, x_2) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right).$$

Означимо са $Q(x_1, x_2)$ квадратну форму $(x-\mu)^T\Sigma^{-1}(x-\mu)$, након чега густина расподеле за X гласи:

$$f(x) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} exp[-\frac{1}{2}Q(x_1, x_2)].$$

Означимо инверз коваријационе матрице на следећи начин:

$$\Sigma^{-1} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \Sigma^{11} & \Sigma^{12} \\ \Sigma^{21} & \Sigma^{22} \end{pmatrix}.$$

Користећи претходно наведену теорему 2.0.4, закључујемо да важе следеће везе:

$$\Sigma^{11} = \Sigma_{11}^{-1} + \Sigma_{11}^{-1}\Sigma_{12}(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1}\Sigma_{21}\Sigma_{11}^{-1},$$

$$\Sigma_{12} = -\Sigma_{11}^{-1}\Sigma_{12}(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1},$$

$$\Sigma_{21} = -(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1}\Sigma_{21}\Sigma_{11}^{-1},$$

$$\Sigma_{22} = \Sigma_{22}^{-1} + \Sigma_{22}^{-1}\Sigma_{21}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\Sigma_{12}\Sigma_{22}^{-1}.$$

Како је коваријациона матрица симетрична, знамо да важи да је $\Sigma_{21} = \Sigma_{12}^T$, $\Sigma_{11} = \Sigma_{11}^T$ и $\Sigma_{22} = \Sigma_{22}^T$. Користећи ове једнакости и особине транспоната матрице, може се доказати да је $\Sigma^{12} = (\Sigma^{21})^T$. На основу ове једнакости, претходне формуле ћемо записати на другачији начин.

$$\Sigma^{11} = \Sigma_{11}^{-1} + \Sigma_{11}^{-1}\Sigma_{12}(\Sigma_{22} - \Sigma_{12}^T\Sigma_{11}^{-1}\Sigma_{12})^{-1}\Sigma_{12}^T\Sigma_{11}^{-1},$$

$$\Sigma_{12} = -\Sigma_{11}^{-1}\Sigma_{12}(\Sigma_{22} - \Sigma_{12}^T\Sigma_{11}^{-1}\Sigma_{12})^{-1} = (\Sigma^{21})^T,$$

$$\Sigma_{22} = \Sigma_{22}^{-1} + \Sigma_{22}^{-1}\Sigma_{12}^T(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T)^{-1}\Sigma_{12}\Sigma_{22}^{-1}.$$

Забележимо сада квадратну форму Q на другачији начин.

$$
\begin{aligned}
&Q(x_1, x_2) = (x-\mu)^T \Sigma^{-1} (x-\mu) \\
&= [(x_1-\mu_1)^T, (x_2-\mu_2)^T] \begin{pmatrix} \Sigma^{11} & \Sigma^{12} \\ \Sigma^{21} & \Sigma^{22} \end{pmatrix} \begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix} = \\
&= (x_1-\mu_1)^T \Sigma^{11} (x_1-\mu_1) + 2(x_1-\mu_1)^T \Sigma^{12} (x_2-\mu_2) + (x_2-\mu_2)^T \Sigma^{22} (x_2-\mu_2) \\
&= (x_1-\mu_1)^T [\Sigma_{11}^{-1} + \Sigma_{11}^{-1}\Sigma_{12}(\Sigma_{22} - \Sigma_{12}^T \Sigma_{11}^{-1}\Sigma_{12})^{-1}\Sigma_{12}^T \Sigma_{11}^{-1}](x_1-\mu_1) \\
&\quad - 2(x_1-\mu_1)^T [\Sigma_{11}^{-1}\Sigma_{12}(\Sigma_{22} - \Sigma_{12}^T \Sigma_{11}^{-1}\Sigma_{12})^{-1}](x_2-\mu_2) \\
&\quad + (x_2-\mu_2)^T [\Sigma_{22}^{-1} + \Sigma_{22}^{-1}\Sigma_{12}^T(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T)^{-1}\Sigma_{12}\Sigma_{22}^{-1}](x_2-\mu_2) \\
&= (x_1-\mu_1)^T \Sigma_{11}^{-1} (x1-\mu 1) \\
&\quad + (x_1-\mu_1)^T [\Sigma_{11}^{-1}\Sigma_{12}(\Sigma_{22} - \Sigma_{12}^T \Sigma_{11}^{-1}\Sigma_{12})^{-1}\Sigma_{12}^T \Sigma_{11}^{-1}](x_1-\mu_1) \\
&\quad - 2(x_1-\mu_1)^T [\Sigma_{11}^{-1}\Sigma_{12}(\Sigma_{22} - \Sigma_{12}^T \Sigma_{11}^{-1}\Sigma_{12})^{-1}](x_2-\mu_2) \\
&\quad + (x_2-\mu_2)^T [\Sigma_{22}^{-1} + \Sigma_{22}^{-1}\Sigma_{12}^T(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T)^{-1}\Sigma_{12}\Sigma_{22}^{-1}](x_2-\mu_2).
\end{aligned}
$$

Да бисмо квадратну форму представили у другачијем облику, искористићемо наредно тврђење.

За произвољне векторе u,v и симетричну матрицу A важи:

$$
\begin{aligned}
&u^T A u - 2u^T A v + v^T A v = u^T A u - u^T A v - u^T A v + v^T A v = \\
&u^T A(u-v) - (u-v)^T A v = u^T A(u-v) - v^T A(u-v) = \\
&(u-v)^T A(u-v) = (v-u)^T A(v-u).
\end{aligned}
$$

Показану једнакост применићемо на израз за квадратну форму.

$$
\begin{aligned}
&Q(x_1, x_2) = (x_1-\mu_1)^T \Sigma_{11}^{-1}(x_1-\mu_1) \\
&\quad + [(x_2-\mu_2) - \Sigma_{12}^T \Sigma_{11}^{-1}(x_1-\mu_1)]^T (\Sigma_{22} - \Sigma_{12}^T \Sigma_{11}^{-1}\Sigma_1 2)^{-1} [(x_2-\mu_2) - \Sigma_{12}^T \Sigma_{11}^{-1}(x_1-\mu_1)].
\end{aligned}
$$

Дефинишимо ознаке b, A, $Q_1(x_1)$ и $Q_2(x_2)$ на следећи начин:

$$
\begin{aligned}
b &:= \mu_2 + \Sigma_{12}^T \Sigma_{11}^{-1}(x_1-\mu_1), \\
A &:= \Sigma_{22} - \Sigma_{12}^T \Sigma_{11}^{-1}\Sigma_{12}, \\
Q_1(x_1) &:= (x_1-\mu_1)^T \Sigma_{11}^{-1}(x_1-\mu_1), \\
Q_2(x_1, x_2) &= (x_2-b)^T A^{-1}(x_2-b).
\end{aligned}
$$

Заменом нових ознака у израз за Q добијамо да је

$$Q(x_1, x_2) = Q_1(x_1) + Q_2(x_1, x_2).$$

Користећи формулу $|M| = |A||D - CA^{-1}B|$ *која представља рачунање детерминанте блок матрице* M *чији су блокови* A, B, C *и* D *добијамо да је детерминанта матрице* Σ *дата са:*

$$|\Sigma| = |\Sigma_{11}||\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}| = |\Sigma_{11}||\Sigma_{22} - \Sigma_{12}^{T}\Sigma_{11}^{-1}\Sigma_{12}|.$$

Сада, узимајући у обзир све претходно изведено, за заједничку густину расподеле вектора $X^{(1)}$ *и* $X^{(2)}$ *можемо тврдити да важи:*

$$\begin{aligned}
f(x_1, x_2) &= \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} exp[-\frac{1}{2}Q(x_1, x_2)] \\
&= \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma_{11}|^{\frac{1}{2}}|\Sigma_{22} - \Sigma_{12}^{T}\Sigma_{11}^{-1}\Sigma_{12}|^{\frac{1}{2}}} exp[-\frac{1}{2}(Q_1(x_1) + Q_2(x_1, x_2))] \\
&= \frac{1}{(2\pi)^{\frac{k}{2}}|\Sigma_{11}|^{\frac{1}{2}}} exp\left(-\frac{1}{2}(x_1 - \mu_1)^{T}\Sigma_{11}^{-1}(x_1 - \mu_1)\right) \frac{1}{(2\pi)^{\frac{m}{2}}|A|^{\frac{1}{2}}} exp\left(-\frac{1}{2}(x_2 - b)^{T}A^{-1}(x_2 - b)\right) \\
&= f_{\mathcal{N}}(x_1, \mu_1, \Sigma_{11}) f_{\mathcal{N}}(x_2, b, A).
\end{aligned}$$

Маргинална расподела за $X^{(1)}$ *је:*

$$f_1(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2)dx_2 = \frac{1}{(2\pi)^{\frac{k}{2}}|\Sigma_{11}|^{\frac{1}{2}}} exp\left(-\frac{1}{2}(x_1 - \mu_1)^{T}\Sigma_{11}^{-1}(x_1 - \mu_1)\right).$$

Теорема 2.0.6. *Условна расподела* $X^{(2)}|X^{(1)}$ *је вишедимензиона нормална са параметрима* $\mu = \mu_2 + \Sigma_{12}^{T}\Sigma_{11}^{-1}(x_1 - \mu_1)$ *и* $\Sigma = \Sigma_{22} - \Sigma_{12}^{T}\Sigma_{11}^{-1}\Sigma_{12}$.

Доказ. *Користећи заједничку расподелу за* $X^{(1)}$ *и* $X^{(2)}$*, маргиналну расподелу вектора* $X^{(1)}$ *из доказа претходне теореме добијамо да важи:*

$$f_{2|1}(x_2|x_1) = \frac{f(x_1, x_2)}{f(x_1)} = \frac{1}{(2\pi)^{\frac{m}{2}}|A|^{\frac{1}{2}}} exp\left(-\frac{1}{2}(x_2 - b)^{T}A^{-1}(x_2 - b)\right).$$

Како је $b := \mu_2 + \Sigma_{12}^{T}\Sigma_{11}^{-1}(x_1 - \mu_1)$ *и* $A := \Sigma_{22} - \Sigma_{12}^{T}\Sigma_{11}^{-1}\Sigma_{12}$ *то је тврђење директно показано и важи* $X^{(2)}|X^{(1)} \sim f_{\mathcal{N}}(\mu_2 + \Sigma_{12}^{T}\Sigma_{11}^{-1}(x_1 - \mu_1), \Sigma_{22} - \Sigma_{12}^{T}\Sigma_{11}^{-1}\Sigma_{12})$.

Теорема 2.0.7. *Случајни вектори* $X^{(1)}$ *и* $X^{(2)}$ *су независни ако важи* $\Sigma_{12} = \Sigma_{21} = 0$. *Додатно,* $X_{\mathcal{N}}^{(2)}(\mu^{(2)}, \Sigma_{22})$.

Доказ. *Ова теорема представља специјалан случај теореме 2.0.4 па ће се и сам доказ ослањати на већ изведен доказ поменуте теореме.*

Претпоставимо да важи $\Sigma_{12} = \Sigma_{21} = 0$*, тада је инверз коваријационе матрице дат са:*

$$\Sigma^{-1} = \begin{pmatrix} \Sigma^{11} & 0 \\ 0 & \Sigma^{22} \end{pmatrix}.$$

Након нове претпоставке примећујемо да важи:

$$b = \mu_2 + \Sigma_{12}^T \Sigma_{11}^{-1}(x_1 - \mu_1) = \mu_2,$$

$$A = \Sigma_{22} - \Sigma_{12}^T \Sigma_{11}^{-1}\Sigma_{12} = \Sigma_{22},$$

$$|\Sigma^{-1}| = |\Sigma^{11}||\Sigma^{22}|.$$

Заједничка функција расподеле је сада дефинисана као:

$$\begin{aligned}
f(x_1, x_2) &= \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}Q(x_1, x_2)\right) \\
&= \frac{1}{(2\pi)^{\frac{k}{2}}|\Sigma_{11}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}Q_1(x_1)\right) \frac{1}{(2\pi)^{\frac{n-k}{2}}|\Sigma_{22}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}Q_2(x_1, x_2)\right) \\
&= \frac{1}{(2\pi)^{\frac{k}{2}}|\Sigma_{11}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x_1 - \mu_1)^T \Sigma_{11}^{-1}(x_1 - \mu_1)\right) \cdot \\
&\quad \frac{1}{(2\pi)^{\frac{m}{2}}|\Sigma_{22}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x_2 - \mu_2)^T \Sigma_{22}^{-1}(x_2 - \mu_2)\right) \\
&= f_{\mathcal{N}}(x_1, \mu_1, \Sigma_{11}) f_{\mathcal{N}}(x_2, \mu_2, \Sigma_{22}).
\end{aligned}$$

Сада можемо израчунати маргиналну густину за $X^{(2)}$.

$$f_2(x_2) = \int f(x_1, x_2)dx_1 = \frac{1}{(2\pi)^{\frac{m}{2}}|\Sigma_{22}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x_2 - \mu_2)^T \Sigma_{22}^{-1}(x_2 - \mu_2)\right).$$

Показали смо да важи да $X^{(2)}$ *има вишедимензиону нормалну расподелу с параметрима* μ_2 *и* Σ_{22}. *Како је заједничка густина производ маргиналних густина, то важи да су* $X^{(1)}$ *и* $X^{(2)}$ *независни чиме је доказ завршен.*

Изглед вишедимензионе нормалне расподеле

У овом делу рада биће описан облик вишедимензионих нормалних расподела у простору. Акценат ће бити на дводимензионим нормалним расподелама приказаним у тродимензионом простору. Овај приступ ће нам помоћи да боље разумемо простирање нормалних расподела у више димензија.

Посматрајмо $X = (X_1, X_2)$, чија је расподела дводимензиона нормална. Фокусираћемо се на налажења везе између параметара расподеле$(\mu_1, \sigma_{11}, \mu_2, \sigma_{22}, \sigma_{12})$ и њеног облика.

Кренућемо од чињенице да за тачке са једнаком густином расподеле важи да је $f(x_1, x_2) = k$. Расписивањем ове једнакости коришћењем формуле за густину дводимензионе нормалне расподеле добијамо следеће:

$$f(x_1, x_2) = k,$$
$$\frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22} - \sigma_{12}^2}} \exp\left(-\frac{\sigma_{22}(x_1-\mu_1)^2 + \sigma_{11}(x_2-\mu_2)^2 - 2\sigma_{12}(x_1-\mu_1)(x_2-\mu_2)}{2(\sigma_{11}\sigma_{22}-\sigma_{12}^2)}\right) = k,$$
$$\exp\left(-\frac{\sigma_{22}(x_1-\mu_1)^2 + \sigma_{11}(x_2-\mu_2)^2 - 2\sigma_{12}(x_1-\mu_1)(x_2-\mu_2)}{2(\sigma_{11}\sigma_{22}-\sigma_{12}^2)}\right) = 2k\pi\sqrt{\sigma_{11}\sigma_{22} - \sigma_{12}^2}$$
$$-\frac{\sigma_{22}(x_1-\mu_1)^2 + \sigma_{11}(x_2-\mu_2)^2 - 2\sigma_{12}(x_1-\mu_1)(x_2-\mu_2)}{2(\sigma_{11}\sigma_{22}-\sigma_{12}^2)} = \ln\left(2k\pi\sqrt{\sigma_{11}\sigma_{22} - \sigma_{12}^2}\right).$$

Можемо приметити да израз $-2(\sigma_{11}\sigma_{22} - \sigma_{12}^2)\ln\left(2k\pi\sqrt{\sigma_{11}\sigma_{22} - \sigma_{12}^2}\right)$ не зависи од x_1 и x_2. Имајући у виду чињеницу да је k позитиван број који мора бити мањи од максимума густине, који је у овом случају $\frac{1}{2k\pi\sqrt{\sigma_{11}\sigma_{22}-\sigma_{12}^2}}$, претходни израз можемо посматрати као константу $c > 0$. Односно,

$$c = -2(\sigma_{11}\sigma_{22} - \sigma_{12}^2)\ln\left(2k\pi\sqrt{\sigma_{11}\sigma_{22} - \sigma_{12}^2}\right).$$

Тада добијамо да све тачке с једнаком густином задовољавају услов:

$$\sigma_{22}(x_1-\mu_1)^2 + \sigma_{11}(x_2-\mu_2)^2 - 2\sigma_{12}(x_1-\mu_1)(x_2-\mu_2) = c.$$

Посматрајући различите случајеве можемо сагледати зависности контуре густине и параметара расподеле.

Случај када X_1 и X_2 нису корелисани

У овом случају знамо да важи $\sigma_{12} = 0$, па тачке са једнаком густином задовољавају следеће:

$$\sigma_{22}(x_1-\mu_1)^2 + \sigma_{11}(x_2-\mu_2)^2 = c,$$
$$\frac{(x_1-\mu_1)^2}{\frac{c}{\sigma_{22}}} + \frac{(x_2-\mu_2)^2}{\frac{c}{\sigma_{11}}} = 1.$$

У случају да су σ_{11} и σ_{22} једнаке, претходна једнакост ће гласити:

$$(x_1-\mu_1)^2 + (x_2-\mu_2)^2 = \frac{c}{\sigma_{11}}.$$

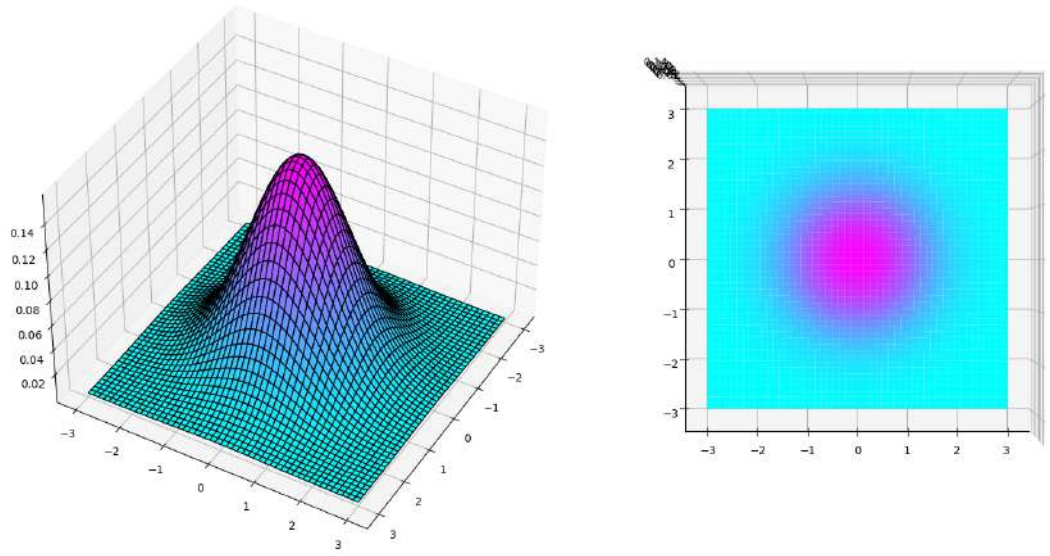
Можемо закључити да у овом случају тачке са једнаком густином формирају кругове полипречника $\sqrt{\frac{c}{\sigma_{11}}}$ и центром у (μ_1, μ_2).

За случај када важи да је $\sigma_{11} > \sigma_{22}$ добијамо елипсе са истим центром чија се дужа оса поклапа са x осом. Ротацијом ове елипсе за 90 степени добијамо изглед контура када је $\sigma_{11} < \sigma_{22}$.

У наставку су приказане илустрације дводимензионих нормалних расподела у простору, као и изглед из птичије перспективе на основу ког се може посматрати распршеност тачака око центра расподеле. Уколико није другачије наведено, центри су у тачки $\mu = (0, 0)$. Циљ приказа је илустрација различитих облика и изгледа дводимензионих нормалних расподела у простору који могу варирати услед промене коваријационих матрица.

На слици 2.1 приказана је стандардна дводимензиона нормална расподела са јединичном коваријационом матрицом која представља пример расподела где тачке са истом густином формирају кругове око центра расподеле.

Слика 2.2 пример је расподеле с параметрима $\sigma_{11} = 2, \sigma_{22} = 8, \sigma_{12} = 0$. Како важи да је $\sigma_{22} > \sigma_{11}$ то се дужа оса елипсе поклапа са y осом.

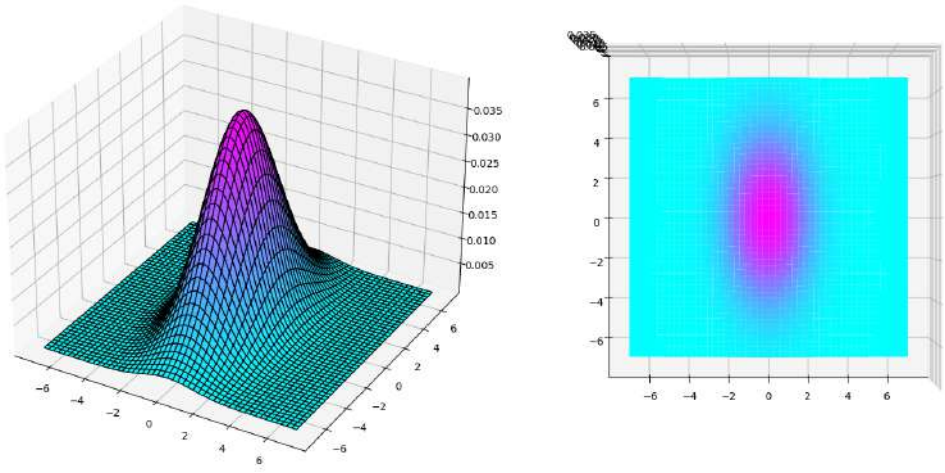


Слика 2.1: Приказ стандардне дводимензионе нормалне расподеле

Случај када су X_1 и X_2 корелисани

У овом делу посматраћемо две могућности.

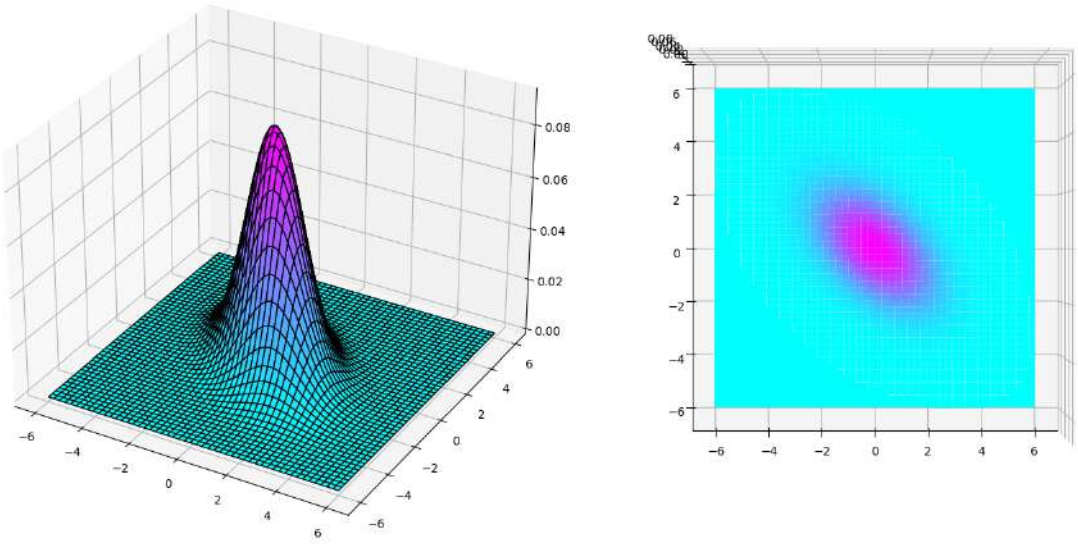
Први подслучај користи услове $\sigma_{12} \neq 0$ и $\sigma_{11} = \sigma_{22}$ услед чега се добија да тачке са истом густином формирају елипсе ротиране за 45 степени у односу



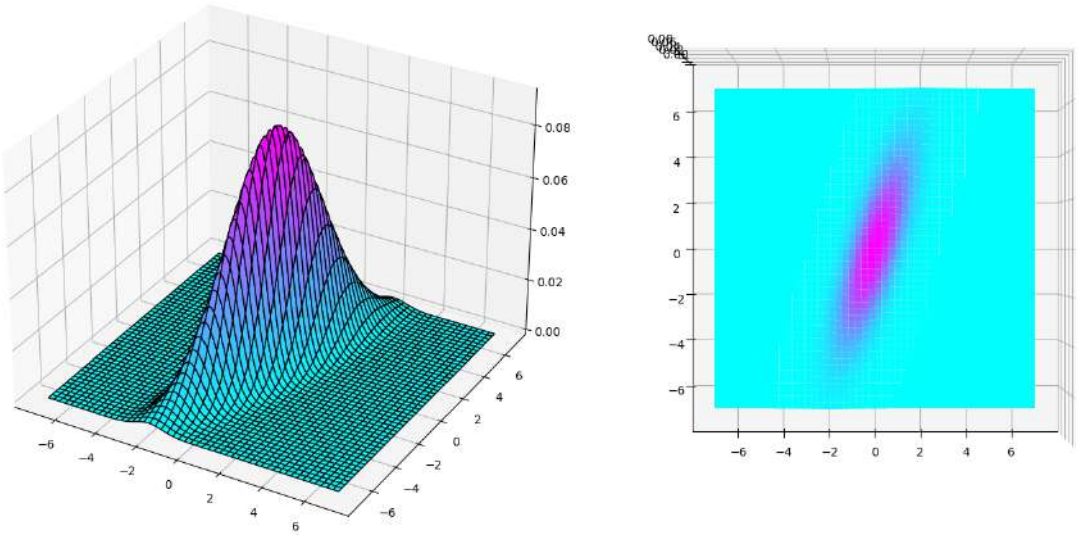
Слика 2.2: Приказ дводимензионих расподела код којих компоненте нису корелисане

на x осу. Уколико важи да је $\sigma_{12} > 0$, велика оса елипсе поклапаће се с правом $y = x$, а у супротном с правом $y = -x$. Случај када је $\sigma_{12} < 0$ може се уочити на слици 2.3, где је приказана вишедимензиона нормална расподела са параметрима $\sigma_{11} = 2, \sigma_{22} = 2, \sigma_{12} = -1$.

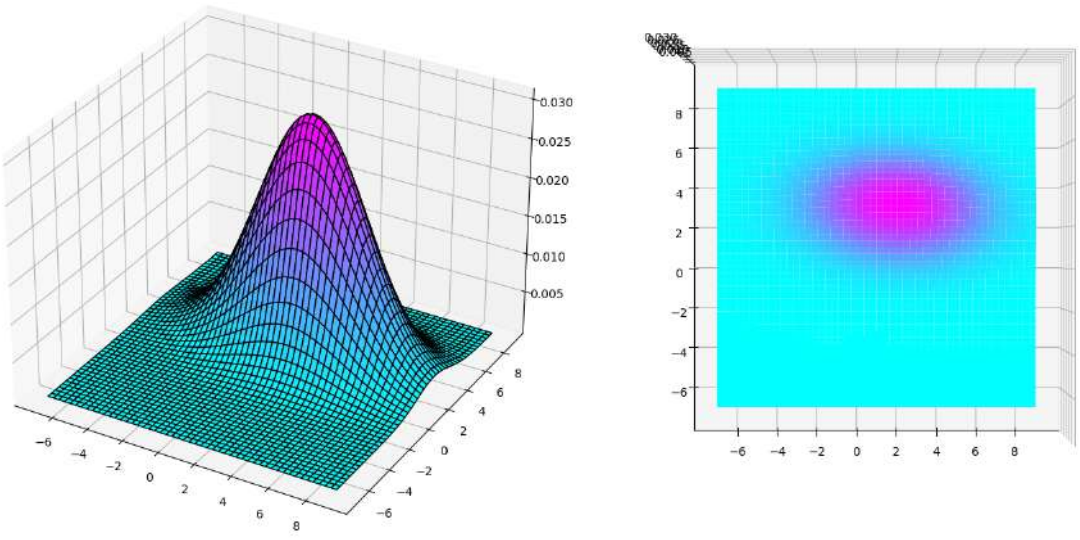
Други подслучај претпоставља да важи $\sigma_{12} \neq 0$ и $\sigma_{11} \neq \sigma_{22}$. Контуре норалних расподеле чији параметри матрице коваријације задовољавају ове услове имају облик елипсе која може заклапати произвољан угао са x осом. Примери везани за расподеле из ове групе могу се видети на сликама 2.4 и 2.5, где прва расподела има параметре $\sigma_{11} = 1, \sigma_{22} = 7, \sigma_{12} = 2$, а друга $\sigma_{11} = 9, \sigma_{22} = 3, \sigma_{12} = -0.4, \mu_1 = 2, \mu_2 = 3$. Као што се да приметити, тачке могу бити сконцетрисане у облику произвољне елипсе, где угао који велика оса елипсе заклапа с x осом, њен положај и облик зависе од параметара расподеле. Подсећања ради, свака од до сада приказаних нормалних расподела може имати центар у произвољној тачки, као што је приказано на слици 2.5 где расподела има вектор средњих вредности $\mu = (2, 3)$.



Слика 2.3: Приказ дводимензионих расподела код којих су компоненте корелисане и важи да су дисперзије компоненти једнаке



Слика 2.4: Приказ дводимензионих расподела код којих су компоненте позитивно корелисане и дисперзије компоненти различите



Слика 2.5: Приказ дводимензионих расподела код којих су компоненте негативно корелисане и дисперзије компоненти различите

У вишим димензијама вишедимензионе нормалне расподеле формирају елипсоиде у n димензионом простору чији величина и облик зависе од сопствених вредности и вектора коваријационе матрице.

Глава 3

Мешавина нормалних расподела

Иако због својих корисних својстава нормална расподела представља једну од најкоришћенијих, приликом моделирања реалних података она може имати одређена ограничења. Неретко се у пракси дешава да наши подаци имају више мода због чега се појавила потреба за увођењем мешавине нормалних расподела. Као што јој и само име каже, мешавина нормалних рапсодела представља комбинацију више Гаусових расподела и користи се за описивање података који имају сложенију структуру. Претпоставимо да нам је дата тежина за сваку особу из узорка без тога да знамо да ли је у питању мушкарац или жена. С обзиром на то да мушкарци и жене имају различите просечне тежине очекујемо да бисмо приликом посматрања густине наших података уочили мешавину две нормалне расподеле и из тог разлога би придруживање обичне нормалне расподеле било погрешно. Овај пример илуструје значај увођења мешавине нормалних расподела приликом рада са подацима који у себи садрже различите подгрупе. Препознавање подгрупа и образаца у нашим подацима нам омогућава боље разумевање података, а самим тим и могућност за прилагођавање одређеном проблему који решавамо и постизање бољих резултата.

Густина мешавине K нормалних расподела дефинише се изразом:

$$f_m(x|\mu, \Sigma, \pi) = \sum_{k=1}^{K} \pi_k f_{\mathcal{N}}(x; \mu_k, \Sigma_k),$$

где $f_{\mathcal{N}}(x; \mu_k, \Sigma_k)$ представља густину сваке од нормалних расподела која учествује у мешавини, а π_k вероватноћу да произвољна опсервација из мешавине долази из $k-$те расподеле. Вредност π_k често називамо тежином, односно уделом који свака од појединачних нормалних расподела има у мешавини. Дакле,

свака мешавина нормалних расподела јединствено је одређена вектором μ који представља вектор средњих вредости, вектором коваријационих матрица Σ и вектором π који садржи вредности π_k за сваку од расподела. Такође, обратимо пажњу да приликом коришћења вероватноћа π_k мора да важи:

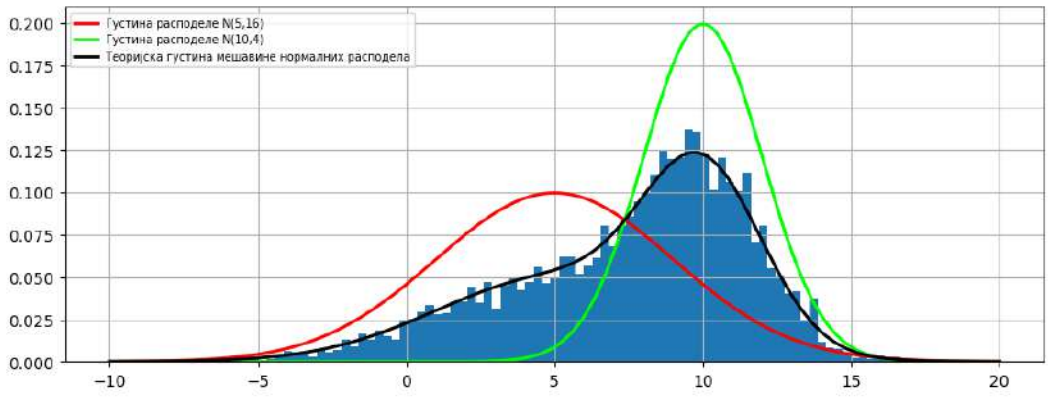
$$\sum_{k=1}^{K} \pi_k = 1.$$

У наредном примеру приказаћемо изглед узорка из мешавине две једнодимензионе нормалне расподела.

Пример 3.0.1. *Посматрајмо две једнодимензионе нормалне расподеле са средњим вредностима 5 и 10 и стандардним девијацијама 4 и 2, односно расподеле* $f_{\mathcal{N}}(5, 16)$ *и* $f_{\mathcal{N}}(10, 4)$*. Генерисаћемо узорак* X *који се састоји од 5000 тачака из мешавине са једнаким уделима две поменуте расподеле. Густина наше мешавине дефинисана је следећим изразом:*

$$f_m(x|\mu, \Sigma, \pi) = \frac{1}{2} f_{\mathcal{N}}(5, 16) + \frac{1}{2} f_{\mathcal{N}}(10, 4).$$

Слика 3.1 илуструје изглед генерисаних података, криве густина појединачних нормалних расподела, као и густину саме мешавине. Као што можемо приметити у овом примеру, иако је узорак генерисан из мешавине две нормалне расподеле, две подгрупе у подацима нису одмах уочљиве тако да треба имати на уму да подгрупе не морају бити увек очигледне.



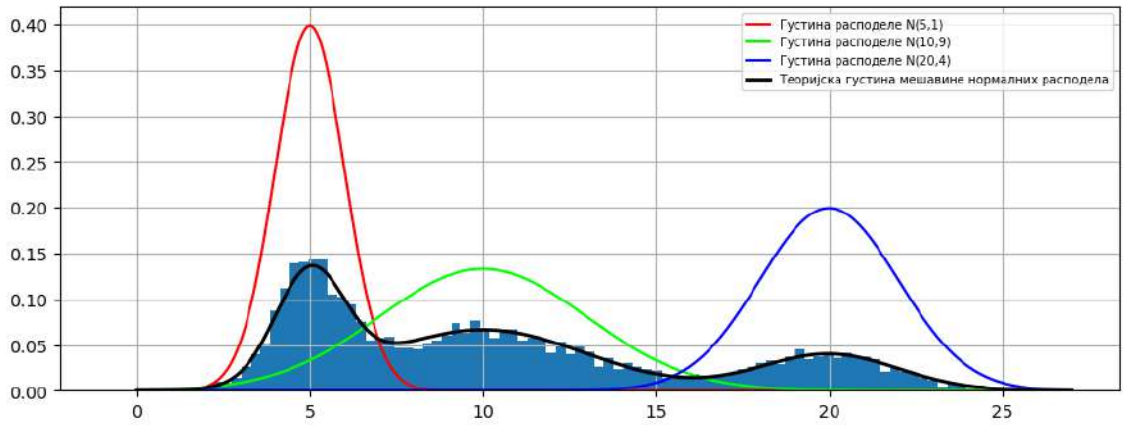
Слика 3.1: Приказ густина расподела

Промена броја нормалних расподела и њихових учесталости у мешавини нам омогућава описивање података различите сложености. Мењањем тежина

можемо да утичемо на значај одређене подгрупе у подацима и самим тим да истакнемо жељене обрасце у истим.

У наредном примеру генерисаћемо сложенију структуру података и приказати мешавину три једнодимензионе нормалне расподеле.

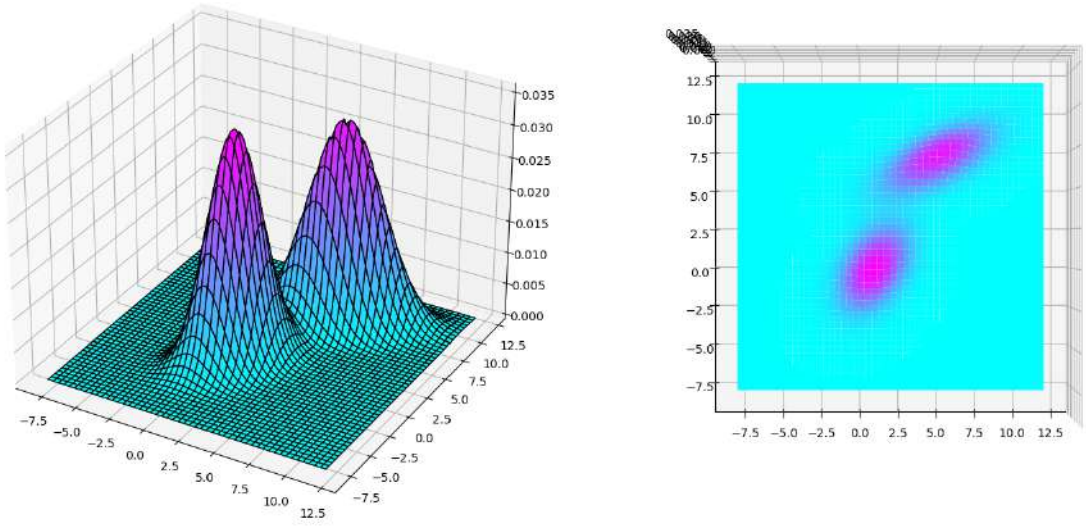
Пример 3.0.2. *Посматрајмо узорак од 5000 тачака генерисан из мешавине три нормалне расподеле,* $f_{\mathcal{N}}(5,1)$*,* $f_{\mathcal{N}}(10,9)$ *и* $f_{\mathcal{N}}(20,4)$*. У овом примеру компоненте унутар мешавине нису једнако расподељене и вектор* π *задат је тако да важи* $\pi = (0.3, 0.5, 0.2)$*. На слици 3.2 запажамо густину мешавине, као и густине њене три компоненте. За разлику од претходног примера, јасно је уочљиво постојање три моде, односно све три нормалне расподеле су лако видљиве.*



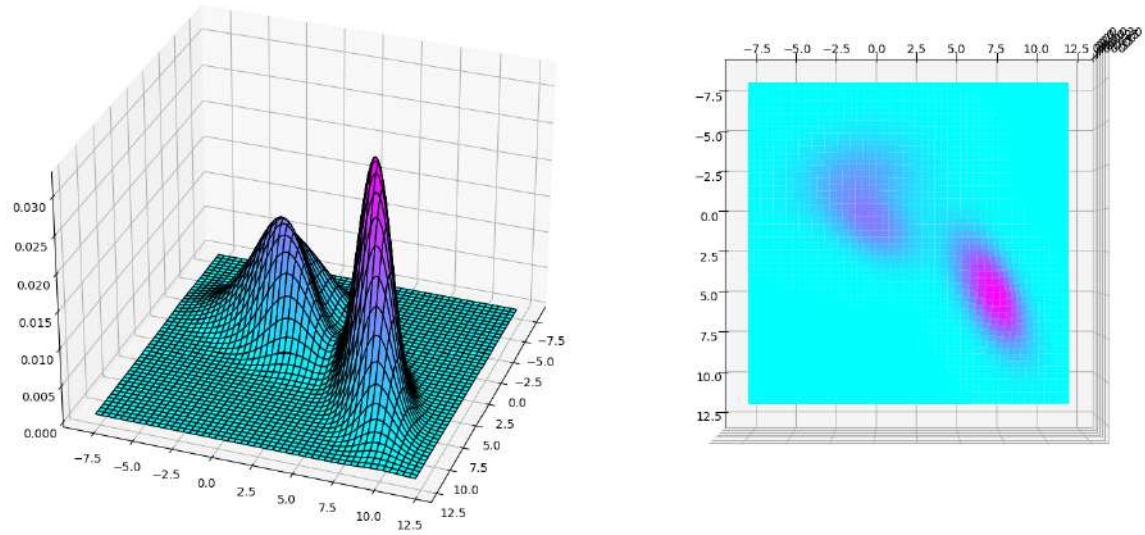
Слика 3.2: Приказ густина расподела

У практичним проблемима учесталије долази до рада са вишедимензионим подацима због чега се много чешће јављају мешавине вишедимензионих нормалних расподела. Иако је у претходним примерима приказан рад са једнодимензионим, исти концепт се примењује и на вишедимензионе нормалне расподеле.

На сликама 3.3 и 3.4 налазе се илустрације мешавина две и три дводимензионе нормалне расподеле. Приказ лево представља изглед расподела у простору, док десно можемо посматрати контуре расподела у две димензије. На слици 3.3 расподеле имају једнак удео, па су обе лако уочљиве, док нам илустрација испод приказује како у простору изгледа мешавина чије се подгрупе преплићу.



Слика 3.3: Приказ густине мешавине 2 дводимензионе нормалне расподеле са једнаким уделима у расподели.



Слика 3.4: Приказ густине мешавине 3 дводимензионе нормалне расподеле са уделима $\pi = \{0.2, 0.5, 0.3\}$.

Мешавине нормалних расподела наилазе на примену у разним областима. Уколико у нашим подацима уочавамо одређене групе, односно кластере, које можемо описати различитим нормалним расподалама тада можемо претпоставити да подаци имају мешавину нормалних расподела. Ова претпоставка нам омогућава коришћење мешавине нормалних расподела за кластеровање података, о чему ће бити више речи у наредном поглављу.

Глава 4

Кластеровање коришћењем мешавине нормалих расподела

4.1 Ненадгледано учење

За разлику од надгледаног учења где поседујемо крајње резултате које желимо да наш модел научи, код ненадгледаног учења ово није случај. Ненадгледано учење је вид машинског учења који има за циљ препознавање веза међу подацима, без присуства циљне променљиве. Коришћењем алгоритама ненадгледаног учења, можемо стећи увид у структуру података, као и у обрасце и законитости који владају међу њима. Ова сазнања нам могу омогућити боље упознавање с подацима, као и корисне информације за даљу анализу и процесе доношења одлука. Постоје различити алгоритми који се заснивају на ненадгледаном учењу, а најчешћи су алгоритми кластеровања и алгоритми за смањивање димензије података. Због недостатка циљне променљиве алгоритми ненадгледаног учења имају ограничење у стварима које могу да науче и из тог разлога се често користе за претпроцесирање података, након чега се примењују алгоритми надгледаног учења.

4.2 Кластеровање

Кластеровање је метода ненадгледаног учења и представља поделу података у групе тако да су припадници једне групе сличнији (по неком критеријуму) међу собом него што су слични са члановима других група. Критеријум по ком вршимо кластеровање није једнозначно одређен и зависи од нашег задатка. Та-

кође, број кластера на који желимо да поделимо податке може варирати, компликујући додатно решавање овог проблема. Кластеровање има широку примену у обради података и њиховој анализи почевши од помоћи у организовању и разумевању узорка до могућности смањивања потребне количине података за извршавање неког задатка. Само неки од практичних проблема у којима се кластеровање показало као добро решење су: препознавање различитих врста ткива на медицинским снимцима, уочавање сумњивог понашања корисника кредитних картица које може указивати на одређени вид превара, као и груписање муштерија које послодавцима омогућава прављење другачијих пословних планова и маркетинга за различите групе. Такође, у случају када је потребно радити са великим базама, а постоје ограничени ресурси за складиштење података, кластеровање се показало као веома корисно решење, јер омогућава чување само репрезентативних чланова кластера. Вршења различитих врста истраживања на великом узорку такође се могу испоставити скупа због чега је корисно истраживања вршити само на одређеном броју представника сваког кластера уместо на целом скупу. Кластеровање је често корисно чак и када нам подела података није главни задатак, јер може представљати помоћ у претпроцесирању података. У случајевима где се сусрећемо са неизбалансираним групама унутар скупа за тренирање претпроцесирање приликом ког би се подаци груписали позитивно би утицало на смањење преприлагођавања и боље перформансе модела јер би се моделу обезбедиле разноврсне инстанце на којима ће бити трениран. Такође, у подацима могу постојати и одређене законитости и групе које човек није у стању да уочи, тако да би у овом случају кластеровање помогло и за боље разумевање података и налажење нових начина за приступање решавању проблема обраде података.

Постоје различити алгоритми за кластеровање и од самог проблема зависи који од алгоритама ће се најбоље показати. Ипак, у овом раду ћемо, у контексту кластеровања, приказати рад мешавина вишедимензионих нормалних расподела, које представљају уопштење алгоритма К средина.

4.3 Алгоритам К средина

Због своје једноставности, алгоритам К средина је широко распрострањен у кластеризационим проблемима. За потребе разумевања рада алгоритма, неопходно је увести појам центроиде кластера. Сваки кластер k има једну цен-

троиду, μ_k, која се, у случају овог алгоритма, добија упросечавањем елемената посматраног кластера. Користећи растојања тачака од центроида, алгоритам К средина групише податке у К кластера.

Претпоставимо да имамо скуп података $X = (x_1, ..., x_N)$ који желимо да поделимо на K кластера. Број K је унапред одређен, а начини за одређивање броја кластера биће описани у наставку рада. Нека је тачки x_n, $\forall n \in \{1, ..., N\}$, придружен вектор $r_n = \{r_{n1}, ..., r_{nK}\}$ тако да је $\forall k \in \{1, ..., K\}$ и $\forall n \in \{1, ..., N\}$ $r_{nk} \in \{0, 1\}$ дефинисан као бинарни индикатор који нам указује на то ком кластеру тачка x_n припада, односно, уколико x_n припада кластеру k важи $r_{nk} = 1$ и $r_{nj} = 0$ за $j \neq k$. Пошто свака тачка припада тачно једном кластеру $\forall n \in \{1, ..., N\}$ важи:

$$\sum_{k=1}^{K} r_{nk} = 1.$$

Алгоритам К средина има за циљ распоређивање центроида тако да растојање између сваке тачке и центроиде њој додељеног кластера буде минимално. Поставља се питање на који начин одредити центре кластера тако да се задовољи овај услов и осигура да свака тачка буде у непосредној близини одговарајућег центра. Математички гледано овај проблем се може записати на следећи начин. Посматрајмо функцију

$$\sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} ||x_n - \mu_k||_2^2.$$

Дефинисана функција представља збир квадрата Еуклидског растојања сваке тачке од центра њој додељеног кластера. Растојања између тачака и центроида ће бити најмања за r_{nk} и μ_k за које важи да је дата функција минимална. Минимизација ове функције постиже се итеративно.

За почетак се врши иницијализација центроида тако што се произвољно изабере K тачака из узорка. Након тога, кораци (1) і (2) се понављају све док не дође до конвергенције, односно до тренутка кад је промена вредности центроида занемарљиво мала, или док не дође до максималног броја итерација у алгоритму.

(1) У овом кораку, свакој тачки додељујемо најближи кластер тако што вршимо израчунавање удаљености тачака од сваке центроиде, односно кластер чија центроида има најмању удаљеност од посматране тачке сматра се њеним кластером. У општем случају, који ће бити приказан, за одређивање дистанце користи се Еуклидско растојање, међутим, у зависности од ситуације и типа

задатка могуће је користити и друге метрике за рачунање растојања. Формално написано, за посматрано μ и за свако n рачунамо r_n по формули:

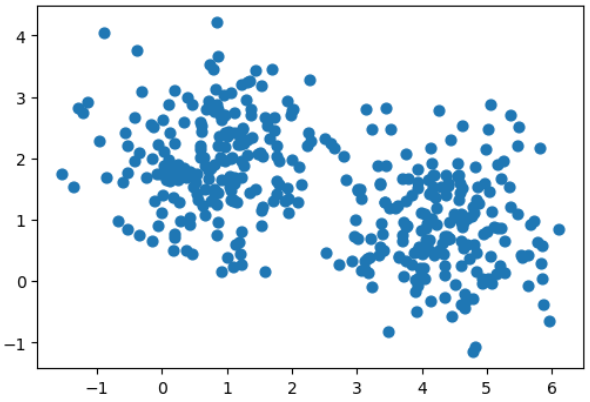
$$r_{nk} = \begin{cases} 1, & \text{ако } k = \text{argmin}_{j=1,\dots,K} ||x_n - \mu_j||_2^2 \\ 0, & \text{иначе} \end{cases}$$

(2) С обзиром да сада свака инстанца има додељен кластер, у овом кораку врши се поновно израчунавање центроида упросечавањем вредности у сваком од кластера. Односно, за фиксирано r_{nk},

$$\mu_k = \frac{\sum_{n=1}^{N} r_{nk} x_n}{\sum_{n=1}^{N} r_{nk}}, \forall k \in \{1, ..., K\}.$$

Јасно је да именилац представља број инстанци у посматраном кластеру k, а бројилац збир свих вредности које припадају том кластеру.

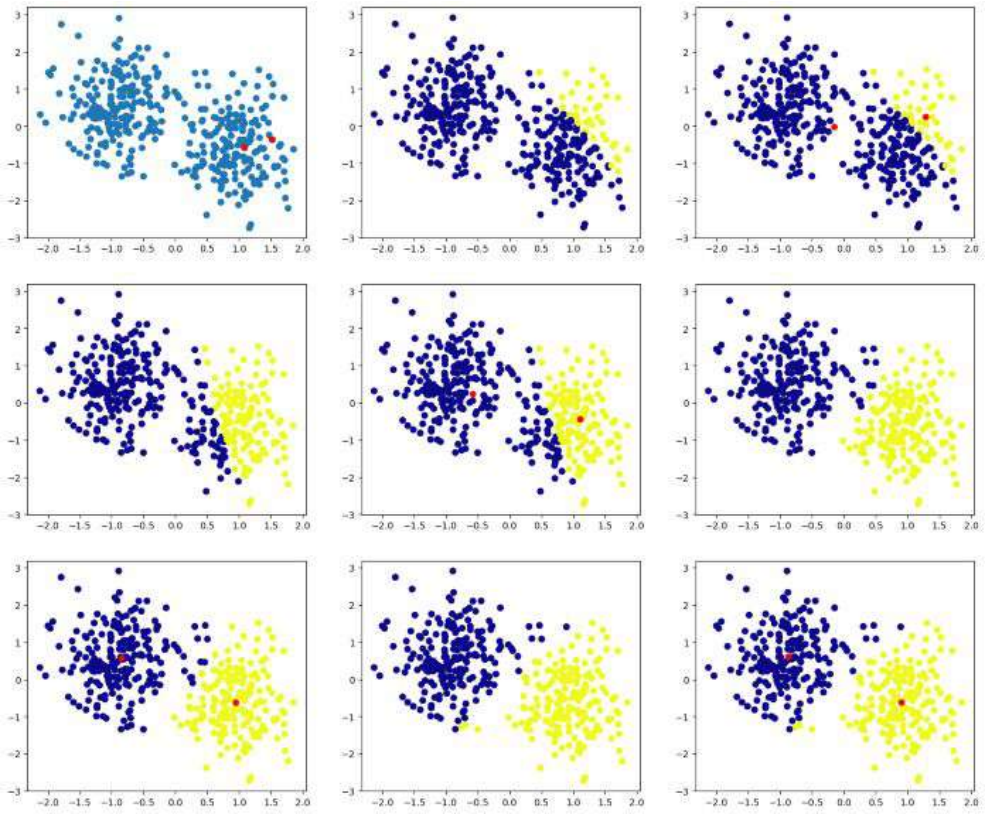
Пример 4.3.1. *Генеришимо скуп података X који ће, ради једноставности приказа, садржати два кластера. Дати скуп илустрован је на слици 4.1. На*



Слика 4.1: Приказ рада алгоритма К средина

слици 4.2 можемо сагледати сам рад алгоритма К средина. За почетак произвољно се бирају две тачке скупа које ће представљати почетне центроиде кластера. Оне су на првој слици означене црвеном бојом. Након тога, као што алгоритам налаже, свака тачка добија свој кластер у односу на удаљеност од датих центроида. Тачке једног кластера приказане су љубичастом бојом, док је други кластер обележен жутом бојом. На трећој сличици означени су нови центри кластера, добијени на основу просечних вредности тачака из посматраних кластера. Наставак илуструје смењивање претходно наведених корака

док се не дође до конвергенције и финална два кластера која се могу видети на последњој сличици.



Слика 4.2: Приказ рада алгоритма K средина

Поред своје једноставности и применљивости овај алгоритам поседује и одређене мане. То су:

- Због минимизације Еуклидског растојања алгоритам претпоставља да су кластери сферног облика што није увек погодно у раду са реалним подацима.
- Немамо никакву информацију о томе колико је алгоритам сигуран да одређена тачка припада датом кластеру.
- Коришћење квадрата растојања доводи до превеликог утицаја удаљенијих тачака на центроиде што чини овај алгоритам осетљивим при раду са одударајућим подацима.

- Како се полазне центроиде бирају насумично, перформансе алгоритма могу зависити од почетног одабира. Из тог разлога се често алгоритам покреће више пута како би се изабрале различите почетне тачке и добили што бољи резултати.

4.4 Примена мешавине нормалних расподела у кластеровању

Мотивација за кластеровање које употребљава мешавине вишедимензионих нормалних расподела настала је из потребе да кластери не буду искључиво сферног облика, већ да, уколико се појави потреба, могу представљати и елипсоиде у простору. Такође, корисно је да за сваку инстанцу постоји вероватноћа припадања сваком од кластера. На овај начин за сваку тачку знамо колико је наше кластеровање прецизно и имамо могућност да уочимо инстанце које се налазе на границама кластера.

Пре него што се упустимо у теоријску страну овог алгоритма, описаћемо идеју принципа његовог рада. Како се ради о проблему кластеровања следи да поседујемо скуп података који желимо да поделимо на одређене групе. Идеја алгоритма је да оценимо параметре вишедимензионе нормалне расподеле тако да она што боље одговара нашим подацима. Колико кластера желимо да имамо, толико нормалних расподела учествује у мешавини. На основу тога, свака од компоненти мешавине представља по једну групу, а све тачке са највећом вероватноћом припадања одређеној нормалној расподели биће сврстане у исти кластер.

Посматрајмо узорак $(\boldsymbol{X_1}, ..., \boldsymbol{X_N})$ који желимо да поделимо на K кластера. Циљ је наћи мешавину K вишедимензионих нормалних расподела која ће најбоље описивати посматране податке. Потребно је да свакој тачки $\boldsymbol{X_i}$ из података доделимо вредност $Z_i \in \{1, ..., K\}, \forall i \in 1, ..., N$ која представља показатељ којој нормалној расподели из мешавине $\boldsymbol{X_i}$ припада. Вектор ових вредности означимо са Z на следећи начин: $Z = (Z_1, ..., Z_N)$. Пошто приликом посматрања података вредности променљиве Z не можемо уочити, ову променљиву зовемо латентна или скривена променљива. С обзиром да тежина π_k представља вероватноћу да произвољно одабрана опсервација из мешавине припада $k-$тој

компоненти, примећујемо да важи:

$$P(Z_i = k) = \pi_k.$$

Како смо, ради тражења најбољих параметра, претпоставили да узорак $\boldsymbol{X} = (\boldsymbol{X_1}, ..., \boldsymbol{X_N})$ долази из мешавине K вишедимензионих нормалних расподела, следи да $\boldsymbol{X_i}|Z_i = k$ има одговарајућу вишедимензиону нормалну расподелу, односно, $\boldsymbol{X_i}|Z_i = k \sim f_{\mathcal{N}}(\mu_k, \Sigma_k)$. Другим речима, густина за $\boldsymbol{X_i}$ је дата са:

$$f_{\boldsymbol{X_i}}(x) = \sum_{k=1}^{K} \pi_k f_{\mathcal{N}}(x; \mu_k, \Sigma_k).$$

Посматрајмо сада вероватноћу $P(Z_i = k|\boldsymbol{X_i} = \boldsymbol{x_i})$ и означимо је са $\gamma_{Z_i}(k)$. Користећи Бајесову формулу добијамо следећу једнакост:

$$\gamma_{Z_i}(k) = P(Z_i = k|\boldsymbol{X_i} = \boldsymbol{x_i}) = \frac{P(Z_i = k) f_{\boldsymbol{X_i}|Z_i=k}(\boldsymbol{x_i})}{f_{\boldsymbol{X_i}}(\boldsymbol{x_i})} = \frac{\pi_k f_{\mathcal{N}}(\boldsymbol{x_i}; \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j f_{\mathcal{N}}(\boldsymbol{x_i}; \mu_j, \Sigma_j)}.$$

Као што је већ поменуто, главни циљ приликом коришћења оваквог начина кластеровања је оценити параметаре $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ и $\boldsymbol{\pi}$ како бисмо свакој опсервацији придодали кластер за који има највећу вероватноћу да се у њему налази. Да бисмо оценили параметре користимо метод максималне веродостојности за који нам је потребно да израчунамо функцију веродостојности $L(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})$. Пошто се ради о независним и једнако расподељеним опсервацијама, важи:

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = f_m(\boldsymbol{x_1}, ..., \boldsymbol{x_n}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \prod_{i=1}^{N} f_m(\boldsymbol{x_i}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k f_{\mathcal{N}}(\boldsymbol{x_i}|\mu_k, \Sigma_k).$$

Применом логаритамске функције на фукцију веродостојности добијамо:

$$l(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \ln(L(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})) = \sum_{i=1}^{N} \ln \left[\sum_{k=1}^{K} \pi_k f_{\mathcal{N}}(\boldsymbol{x_i}|\mu_k, \Sigma_k) \right].$$

Компликације са налажењем максимума ове функције настају из разлога што логаритам више не упрошћава густину нормалне расподеле, као што је то случај када имамо једну нормалну расподелу. Ипак, покушаћемо да максимизујемо ову функцију помоћу класичног начина изједначавања извода по непознатим параметрима са нулом.

Након рачунања извода логаритма функције веродостојности по μ_k и његовог изједначавања са нулом, добијамо:

$$0 = -\sum_{i=1}^{N} \frac{\pi_k f_{\mathcal{N}}(\boldsymbol{x_i}; \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j f_{\mathcal{N}}(\boldsymbol{x_i}; \mu_j, \Sigma_j)} \Sigma_k(\boldsymbol{x_i} - \mu_k) = -\sum_{i=1}^{N} \gamma_{Z_i}(k)\Sigma_k(\boldsymbol{x_i} - \mu_k).$$

С обзиром да Σ_k и μ_k не зависе од i можемо записати следеће:

$$\begin{aligned} 0 &= -\Sigma_k \sum_{i=1}^{N} \gamma_{Z_i}(k)(\boldsymbol{x_i} - \mu_k), \Big/ \cdot \Sigma_k^{-1} \\ 0 &= -\sum_{i=1}^{N} \gamma_{Z_i}(k)(\boldsymbol{x_i} - \mu_k), \\ 0 &= -\sum_{i=1}^{N} \gamma_{Z_i}(k)\boldsymbol{x_i} + \mu_k \sum_{i=1}^{N} \gamma_{Z_i}(k). \end{aligned}$$

Множење прве једнакости са Σ_k^{-1} могуће је из разлога што је Σ_k инвертибилна матрица, стога њен инверз сигурно постоји.

Даље важи,

$$\mu_k = \frac{\sum_{i=1}^{N} \gamma_{Z_i}(k)\boldsymbol{x_i}}{\sum_{i=1}^{N} \gamma_{Z_i}(k)} = \sum_{i=1}^{N} \frac{\gamma_{Z_i}(k)}{\sum_{i=1}^{N} \gamma_{Z_i}(k)} \boldsymbol{x_i}.$$

Приметимо да изведена средња вредности кластера k има смисла јер представља тежински просек свих инстанци, где у изградњи тежина учествује удео припадања инстанце посматраном кластеру.

Уколико уведемо ознаку $N_k = \sum_{i=1}^{N} \gamma_{Z_i}(k)$, претходну изведену једнакост можемо записати као:

$$\mu_k = \frac{1}{N_k} \sum_{i=1}^{N} \gamma_{Z_i}(k)\boldsymbol{x_i}.$$

Понављањем истог поступка за извод по Σ_k добија се формула за рачунање ове вредности.

$$\Sigma_k = \frac{1}{N_k} \sum_{i=1}^{N} \gamma_{Z_i}(k)(\boldsymbol{x_i} - \mu_k)(\boldsymbol{x_i} - \mu_k)^T.$$

Приликом тражења извода по π_k морамо узети у обзир и услов $\sum_{k=1}^{K} \pi_k = 1$. Из тог разлога, потребно је користити Лагранжове множиоце и наћи максимум израза:

$$\ln(L(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})) + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right).$$

Максимална вредност посматране функције следи из рачуна:

$$0 = \sum_{i=1}^{N} \frac{f_{\mathcal{N}}(\boldsymbol{x_i}|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j f_{\mathcal{N}}(\boldsymbol{x_i}|\mu_j, \Sigma_j)} + \lambda =$$
$$\sum_{i=1}^{N} \frac{\pi_k f_{\mathcal{N}}(\boldsymbol{x_i}|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j f_{\mathcal{N}}(\boldsymbol{x_i}|\mu_j, \Sigma_j)} + \pi_k\lambda =$$
$$\sum_{i=1}^{N} \gamma_{Z_i}(k) + \pi_k\lambda = N_k + \pi_k\lambda.$$

Када претходну једнакости просумирамо по k добијамо да важи:

$$0 = \sum_{k=1}^{K} N_k + \lambda \sum_{k=1}^{K} \pi_k = \sum_{k=1}^{K} N_k + \lambda,$$
$$\lambda = -\sum_{k=1}^{K} N_k = -\sum_{k=1}^{K}\sum_{i=1}^{N} \gamma_{Z_i}(k) = -\sum_{i=1}^{N} \underbrace{\sum_{k=1}^{K} P(Z_i = k|\boldsymbol{X_i} = \boldsymbol{x_i})}_{1}.$$

Јасно је да важи да је $\lambda = -N$, а како смо показали да је $0 = N_k + \pi_k\lambda$ закључујемо да следи:

$$\pi_k = \frac{N_k}{N}.$$

Може се приметити да након извођења ових формула нисмо дошли до експлицитних вредности параметара, јер μ_k, Σ_k и π_k сви зависе од $\gamma_{Z_i}(k)$, док $\gamma_{Z_i}(k)$ зависи од μ_k, Σ_k и π_k. За решавање овог проблема употребљава се ЕМ алгоритам који се често користи за налажење максимума функције веродостојности у случају када имамо латентне променљиве.

ЕМ алгоритам у кластеровању помоћу мешавине нормалних расподела

ЕМ алгоритам је итеративни метод који се састоји од четири корака и употребљава се за максимизацију функције веродостојности.

Објаснићемо ЕМ алгоритам на конкретном примеру мешавине нормалних расподела. Први корак представља иницијализацију непознатих параметара мешавина, односно, вектора средњих вредности, коваријационих матрица и тежина. Након тога, почињемо са другим кораком који зовемо Е корак. Приликом Е

корака тренутне вредности за μ_k, Σ_k и π_k користимо за рачунање нове вредности за $\gamma(Z_{ik})$ коју добијамо преко већ изведене формуле:

$$\gamma_{Z_i}(k) = \frac{\pi_k f_{\mathcal{N}}(\boldsymbol{x_i}; \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j f_{\mathcal{N}}(\boldsymbol{x_i}; \mu_j, \Sigma_j)}.$$

Трећи корак, односно М корак, користи последњу добијену вредност $\gamma_{Z_i}(k)$ за поновно изражавање μ_k, Σ_k и π_k преко формула:

$$\mu_k = \frac{1}{N_k} \sum_{i=1}^{N} \gamma_{Z_i}(k) \boldsymbol{x_i},$$

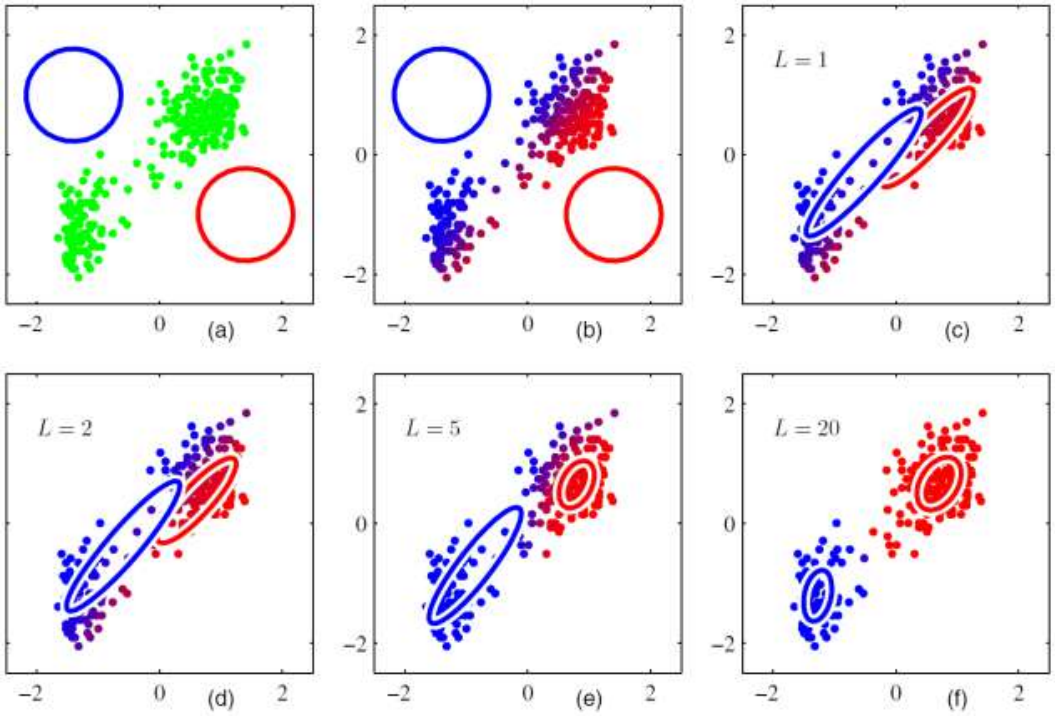
$$\Sigma_k = \frac{1}{N_k} \sum_{i=1}^{N} \gamma_{Z_i}(k)(\boldsymbol{x_i} - \mu_k)(\boldsymbol{x_i} - \mu_k)^T,$$

$$\pi_k = \frac{N_k}{N}.$$

Крајњи корак односи се на проверу да ли алгоритам треба прекинути или наставити са побољшањем оцена. Постоји више начина услед којих алгоритам престаје са радом. Првенствено, врши се провера да ли долази до конвергенције параметара или функције веродостојности ка некој вредности. Уколико је то случај, алгоритам се прекида и за оцене наших параметара узимамо њихове последње вредности. Такође, могуће је поставити и фиксан број итерација који ће служити за обустављање алгоритма уколико конвергенција не буде постигнута пре тога. У случајевима када ниједан од критеријума за прекидање алгоритма није задовољен, поново се извршавају сви кораци алгоритма почевши од Е корака.

На слици 4.3 приказана је илустрација ЕМ алгоритма са двадесет итерација у случају када податке треба да групишемо у два кластера. На слици (а) иницијализовани су параметри кластера, а почетни кластери представљрни су на слици као два круга различите боје. Након тога, за све тачке се рачунају вероватноће припадања сваком од кластера и приписује им се онај кластер за који је вероватноћа припадања била већа. Овај корак приказан је бојењем тачака у складу са одабраним кластером. Сада се за сваки од кластера поново рачунају параметри што доводи до приметне промене облика и положаја кластера. Понављањем Е и М корака кроз 20 итерација јасно примећујемо како подаци видно постају груписани.

Треба обратити пажњу да нас ЕМ алгоритам доводи до локалног максимума функције веродостојности који се може разликовати од глобалног. Ова особи-



Слика 4.3: Илустрација EM алгоритма. Слика преузета из [4]

на може проузроковати различите поделе за различите иницијалне вредности параметара.

Иницијализација може бити урађена на различите начине укључујући и те да за центре наших кластера одаберемо произвољне тачке из података или произвољне тачке генерално. Међутим, овакав избор није баш практичан пошто, у том случају, алгоритму углавном треба више итерација да би конвергирао. Из тог разлога метод K средина се често користи за иницијализацију приликом коришћења EM алгоритма.

Остало је појаснити порекло назива овог алгоритма за шта је потребно да га сагледамо на другачији начин. Видели смо да се максимизација функције веродостојности не може урадити на класичан начин услед њеног компликованог записа и немогућности логаритма да прође кроз суму. Превазилажење овог проблема могуће је коришћењем логаритма фукције веродостојности у односу на заједничку расподелу X и Z, $\sum_{i=1}^{N} \ln f(x_i, z_i|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})$. Како је променљива Z скривена, а њене вредности непознате, ову функцију не можемо израчунати. Уместо тога можемо користити њено условно очекивање, чијом максимизацијом добијамо тражене вредности параметара. Формално речено, приликом E кора-

ка на основу тренутних вредности параметара тражимо очекивање логаритма функције веродостојности заједничке расподеле за X и Z, које ћемо означити са $Q(\boldsymbol{\theta^t}, \boldsymbol{\theta^{t-1}})$, где $\boldsymbol{\theta^t}$ представља тренутни вектор параметара, а $\boldsymbol{\theta^{t-1}}$ параметре у претходној итерацији.

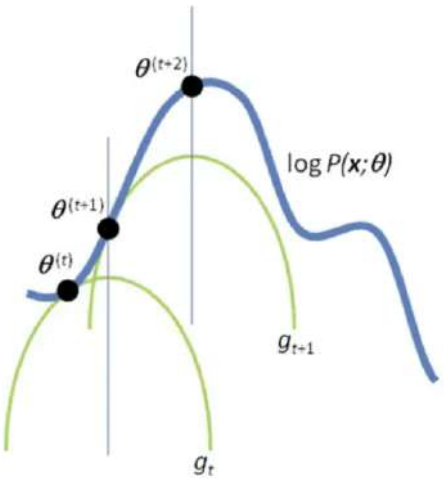
$$Q(\boldsymbol{\theta^t}, \boldsymbol{\theta^{t-1}}) = \sum_z p(z|\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) \ln \left(f(\boldsymbol{x}, z|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) \right).$$

Овај део алгоритма добио је назив Е корак јер долази до рачунања очекивања. Наредни корак, М, представља максимизацију приказаног израза, на основу чега је и добио име.

ЕМ алгоритам у општем случају

Приказали смо употребу ЕМ алгоритма за решавање нашег проблема у тражењу максимума функције веродостојности код кластеровања коришћењем мешавине вишедимензионих нормалних расподела, међутим у овом делу ћемо се приближити раду алгоритма у општем случају и сагледати га из другог угла.

Нека је $L(\boldsymbol{\theta})$ функција веродостојности коју треба максимизовати, а $\boldsymbol{\theta}$ вектор параметара које желимо да оценимо.



Слика 4.4: Приказ одабира параметара ЕМ алгоритма. Слика преузета из [3]

Посматрајмо слику 4.4 како бисмо боље разумели идеју рада ЕМ алгоритма. Нека је плавом бојом представљена функција коју желимо да максимизујемо. Вектор $\boldsymbol{\theta^{(t)}}$ представља вектор параметара у посматраном тренутку t. Идеја алгоритма је да у датој тачки $\boldsymbol{\theta^{(t)}}$ конструишемо помоћну функцију (на приказу

означена зеленом бојом) која представља доње ограничење почетне функције. Након тога, тражимо тачку у којој помоћна функција достиже максимум и ту вредност узимамо као следећу тачку $\boldsymbol{\theta^{(t+1)}}$ за коју понављамо поступак. Ова процедура омогућава долазак до апроксимације траженог максимума функције.

Приметимо да алгоритам има два корака, први је конструисање помоћне функције, који одговара слову „E" у називу алгоритма, а други је тражење њеног максимума, који одговара слову „M".

Дефинишимо сада претходно описани алгоритам математичким формулацијама.

Ознаком $l(\theta|\theta^{(t)})$ обележићемо помоћну функцију приликом чега θ представља аргумент функције, а $\theta^{(t)}$ тачку у којој је помоћна функција дефинисана. У првом кораку потребно је конструисати ову функцију тако да важе услови да она буде доње ограничење главне фукнције, као и да обе функције имају једнаку вредност у тачки $\theta^{(t)}$. Односно, желимо да важи следеће:

1. $L(\boldsymbol{\theta}) \geq l(\boldsymbol{\theta}|\boldsymbol{\theta^{(t)}})$,
2. $L(\boldsymbol{\theta^{(t)}}) = l(\boldsymbol{\theta^{(t)}}|\boldsymbol{\theta^{(t)}})$.

Други корак резервисан је за рачунање нових параметара, односно вредности у којој помоћна функција достиже максимум. Овај корак се може представити формулом:

$$\boldsymbol{\theta^{(t+1)}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\ l(\boldsymbol{\theta}|\boldsymbol{\theta^{(t)}}).$$

Нека је $\boldsymbol{X} = (\boldsymbol{X_1}, ..., \boldsymbol{X_N})$ случајан узорак са заједничком густином $f(\boldsymbol{x}; \boldsymbol{\theta})$. Како је функција веродостојности заједничка густина посматрана као функција параметра $\boldsymbol{\theta}$, њен логаритам, $L(\boldsymbol{\theta})$, дефинишемо са:

$$L(\boldsymbol{\theta}) = \ln f(\boldsymbol{x}|\boldsymbol{\theta}).$$

С обзиром да је циљ максимизовање функције $L(\boldsymbol{\theta})$, желимо да ново $\boldsymbol{\theta}$ које тражимо задовољава услов:

$$L(\boldsymbol{\theta}) > L(\boldsymbol{\theta^{(t)}}),$$

односно, желимо да добијемо што већу вредност разлике

$$L(\boldsymbol{\theta}) - L(\boldsymbol{\theta^{(t)}}).$$

Нека је Z дискретна скривена променљива, а z њена реализација. Користећи формулу потпуне вероватноће, заједничку густину f можемо представити преко латентне променљиве на следећи начин:

$$f(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_z f(\boldsymbol{x}|z, \boldsymbol{\theta}) p(z|\boldsymbol{\theta}).$$

У датој формули, како радимо са комбинацијом апсолутно непрекидних и дискретиних случајних величина, f представља ознаку густине, а p функцију расподеле.

Коришћењем претходно показаног изводимо наредну једнакост.

$$L(\boldsymbol{\theta}) - L(\boldsymbol{\theta^{(t)}}) = \ln f(\boldsymbol{x}|\boldsymbol{\theta}) - \ln f(\boldsymbol{x}|\boldsymbol{\theta^{(t)}}) = \ln \left(\sum_z f(\boldsymbol{x}|z, \boldsymbol{\theta}) p(z|\boldsymbol{\theta}) \right) - \ln f(\boldsymbol{x}|\boldsymbol{\theta^{(t)}}).$$

Лема 4.4.1. *Јенсенова неједнакост: Нека је f конвексна функција на интервалу I. Ако су $x_1, x_2, ..., x_n \in I$ и $\lambda_1, \lambda_2, ..., \lambda_n \geq 0$ такви да $\sum_{i=1}^n \lambda_i = 1$ онда:*

$$f(\sum_{i=1}^n \lambda_i x_i) \leq \sum_{i=1}^n \lambda_i f(x_i).$$

Доказ. *Ову лему доказаћемо уз помоћ индукције. За $n = 1$ важи једнакост и случај је тривијалан. За $n = 2$ неједнакост следи из дефиниције конвексности. Претпоставимо да лема важи за n и докажимо да важи за $n + 1$.*

$$\begin{aligned}
f(\sum_{i=1}^{n+1} \lambda_i x_i) &= f(\lambda_{n+1} x_{n+1} + \sum_{i=1}^{n} \lambda_i x_i) \\
&= f(\lambda_{n+1} x_{n+1} + (1 - \lambda_{n+1}) \frac{1}{1 - \lambda_{n+1}} \sum_{i=1}^{n} \lambda_i x_i) \\
&\leq \lambda_{n+1} f(x_{n+1}) + (1 - \lambda_{n+1}) f(\frac{1}{1 - \lambda_{n+1}} \sum_{i=1}^{n} \lambda_i x_i) \\
&= \lambda_{n+1} f(x_{n+1}) + (1 - \lambda_{n+1}) f(\sum_{i=1}^{n} \frac{\lambda_i}{1 - \lambda_{n+1}} x_i) \\
&\leq \lambda_{n+1} f(x_{n+1}) + (1 - \lambda_{n+1}) \sum_{i=1}^{n} \frac{\lambda_i}{1 - \lambda_{n+1}} f(x_i) \\
&= \lambda_{n+1} f(x_{n+1}) + \sum_{i=1}^{n} \lambda_i f(x_i) \\
&= \sum_{i=1}^{n+1} \lambda_i f(x_i).
\end{aligned}$$

Коришћењем ове леме и чињенице да је $-\ln x$ конвексна функција добијамо да важи:

$$\ln \sum_{i=1}^{n} \lambda_i x_i \geq \sum_{i=1}^{n} \lambda_i \ln(x_i).$$

С обзиром да важи $p(z|\boldsymbol{X}, \boldsymbol{\theta^{(t)}}) \geq 0$ и $\sum_z p(z|\boldsymbol{X}, \boldsymbol{\theta^{(t)}}) = 1$ претходно доказану Јенсенову неједнакост ћемо применити у извођењу које следи.

$$\begin{aligned}
L(\boldsymbol{\theta}) - L(\boldsymbol{\theta^{(t)}}) &= \ln \sum_z f(\boldsymbol{x}|z, \boldsymbol{\theta}) p(z|\boldsymbol{\theta}) - \ln f(\boldsymbol{x}|\boldsymbol{\theta^{(t)}}) \\
&= \ln \sum_z p(z|\boldsymbol{X}, \boldsymbol{\theta^{(t)}}) \frac{f(\boldsymbol{x}|z, \boldsymbol{\theta}) p(z|\boldsymbol{\theta})}{p(z|\boldsymbol{X}, \boldsymbol{\theta^{(t)}})} - \ln f(\boldsymbol{x}|\boldsymbol{\theta^{(t)}}) \\
&\geq \sum_z p(z|\boldsymbol{X}, \boldsymbol{\theta^{(t)}}) \ln \left(\frac{f(\boldsymbol{x}|z, \boldsymbol{\theta}) p(z|\boldsymbol{\theta})}{p(z|\boldsymbol{X}, \boldsymbol{\theta^{(t)}})} \right) - \ln f(\boldsymbol{x}|\boldsymbol{\theta^{(t)}}) \\
&= \sum_z p(z|\boldsymbol{X}, \boldsymbol{\theta^{(t)}}) \ln \left(\frac{f(\boldsymbol{x}|z, \boldsymbol{\theta}) p(z|\boldsymbol{\theta})}{p(z|\boldsymbol{X}, \boldsymbol{\theta^{(t)}}) f(\boldsymbol{x}|\boldsymbol{\theta^{(t)}})} \right).
\end{aligned}$$

Дефинишимо сада:

$$\Delta(\boldsymbol{\theta}|\boldsymbol{\theta^{(t)}}) := \sum_z p(z|\boldsymbol{X}, \boldsymbol{\theta^{(t)}}) \ln \left(\frac{f(\boldsymbol{x}|z, \boldsymbol{\theta}) p(z|\boldsymbol{\theta})}{p(z|\boldsymbol{X}, \boldsymbol{\theta^{(t)}}) f(\boldsymbol{x}|\boldsymbol{\theta^{(t)}})} \right).$$

Јасно је да важи:

$$L(\boldsymbol{\theta}) \geq L(\boldsymbol{\theta^{(t)}}) + \Delta(\boldsymbol{\theta}|\boldsymbol{\theta^{(t)}}).$$

Такође, нашу помоћну функцију $l(\boldsymbol{\theta}, \boldsymbol{\theta^{(t)}})$ ћемо дефинисати на следећи начин:

$$l(\boldsymbol{\theta}, \boldsymbol{\theta^{(t)}}) := L(\boldsymbol{\theta^{(t)}}) + \Delta(\boldsymbol{\theta}|\boldsymbol{\theta^{(t)}}).$$

Потребно је проверити да ли конструисана функција задовољава услове које треба да задовољава помоћна функција, односно треба проверити да ли важи:

1. $L(\boldsymbol{\theta}) \geq l(\boldsymbol{\theta}|\boldsymbol{\theta^{(t)}})$,
2. $L(\boldsymbol{\theta^{(t)}}) = l(\boldsymbol{\theta^{(t)}}|\boldsymbol{\theta^{(t)}})$.

Први услов директно следи из начина на који смо дефинисали функцију $l(\boldsymbol{\theta}|\boldsymbol{\theta^{(t)}})$.

Проверимо сада други услов:

$$\begin{aligned}
l(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) &= L(\boldsymbol{\theta}^{(t)}) + \Delta(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) \\
&= L(\boldsymbol{\theta}^{(t)}) + \sum_z p(z|\boldsymbol{X}, \boldsymbol{\theta}^{(t)}) \ln\left(\frac{f(\boldsymbol{x}|z, \boldsymbol{\theta}^{(t)}) p(z|\boldsymbol{\theta}^{(t)})}{p(z|\boldsymbol{X}, \boldsymbol{\theta}^{(t)}) f(\boldsymbol{x}|\boldsymbol{\theta}^{(t)})}\right) \\
&= L(\boldsymbol{\theta}^{(t)}) + \sum_z p(z|\boldsymbol{X}, \boldsymbol{\theta}^{(t)}) \ln\left(\frac{f(\boldsymbol{x}, z|\boldsymbol{\theta}^{(t)})}{f(\boldsymbol{x}, z|\boldsymbol{\theta}^{(t)})}\right) \\
&= L(\boldsymbol{\theta}^{(t)}) + \sum_z p(z|\boldsymbol{X}, \boldsymbol{\theta}^{(t)}) \ln 1 \\
&= L(\boldsymbol{\theta}^{(t)}).
\end{aligned}$$

Овим је доказано да су функције $l(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ и $L(\boldsymbol{\theta})$ једнаке за $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$.

Сумирајмо сада досадашње резултате, уверили смо се да је функција $l(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ доње ограничење функције $L(\boldsymbol{\theta})$, као и да у тачки $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$ важи једнакост између ових функција. Оно што можемо закључити из овога је да свако $\boldsymbol{\theta}$ које повећава вредност функције $l(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$, такође повећава и вредност функције $L(\boldsymbol{\theta})$. Из тог разлога максимизовањем наше помоћне функције долазимо до вредности параметара, $\boldsymbol{\theta}^{(t+1)}$, које ћемо користити за наредне кораке.

$$\begin{aligned}
\boldsymbol{\theta}^{(t+1)} &= \operatorname*{argmax}_{\boldsymbol{\theta}} \; l(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \\
&= \operatorname*{argmax}_{\boldsymbol{\theta}} \left(L(\boldsymbol{\theta}^{(t)}) + \sum_z p(z|\boldsymbol{X}, \boldsymbol{\theta}^{(t)}) \ln\left(\frac{f(\boldsymbol{x}|z, \boldsymbol{\theta}) p(z|\boldsymbol{\theta})}{p(z|\boldsymbol{X}, \boldsymbol{\theta}^{(t)}) f(\boldsymbol{x}|\boldsymbol{\theta}^{(t)})}\right)\right) \\
&= \operatorname*{argmax}_{\boldsymbol{\theta}} \left(\sum_z p(z|\boldsymbol{X}, \boldsymbol{\theta}^{(t)}) \ln\left(f(\boldsymbol{x}|z, \boldsymbol{\theta}) p(z|\boldsymbol{\theta})\right)\right) \\
&= \operatorname*{argmax}_{\boldsymbol{\theta}} \left(\sum_z p(z|\boldsymbol{X}, \boldsymbol{\theta}^{(t)}) \ln\left(\frac{f(\boldsymbol{x}, z, \boldsymbol{\theta})}{p(z, \boldsymbol{\theta})} \frac{p(z, \boldsymbol{\theta})}{p(\boldsymbol{\theta})}\right)\right) \\
&= \operatorname*{argmax}_{\boldsymbol{\theta}} \left(\sum_z p(z|\boldsymbol{X}, \boldsymbol{\theta}^{(t)}) \ln\left(f(\boldsymbol{x}, z|\boldsymbol{\theta})\right)\right).
\end{aligned}$$

Потребно је нагласити да приликом тражења максимума по параметру $\boldsymbol{\theta}$ можемо занемарити делове израза који зависе само од $\boldsymbol{\theta}^{(t)}$, будући да се они у овом случају понашају као константе и нису пресудни за крајњи резултат.

Алгоритам не гарантује конвергенцију до глобалног максимума функције веродостојности, међутим, морамо приметити битно својство да функција ве-

родостојности расте у сваком наредном кораку алгоритма. Доказ овог тврђења следи у наредним редовима.

Знамо да важе следеће две једнакости:

$$\boldsymbol{\theta}^{(t+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \ (L(\boldsymbol{\theta}^{(t)}) + \Delta(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \ \left(\Delta(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})\right),$$
$$\Delta(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) = 0.$$

Пошто смо параметре за наредни корак изабрали максимизацијом израза $\Delta(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$, можемо закључити да важи

$$\Delta(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) \geq \Delta(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) = 0.$$

Како је:

$$L(\boldsymbol{\theta}^{(t+1)}) - L(\boldsymbol{\theta}^{(t)}) \geq \Delta(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) \geq 0,$$

јасно је да долази до повећања функције веродостојности кроз итерације.

Општи EM алгоритам из угла мешавина вишедимензионих нормалних расподела

Користећи општи алгоритам изведимо формуле за специјалан случај мешавина вишедимензионих нормалних расподела.

Наш проблем базира се на максимизацији логаритма функције веродостојности. Како бисмо испоштовали нотацију општег алгоритма ову функцију ћемо означити на следећи начин:

$$L(\theta) = \sum_{i=1}^{N} \ln \left[\sum_{k=1}^{K} \pi_k f_{\mathcal{N}}(\boldsymbol{x_i}|\mu_k, \Sigma_k) \right] = \sum_{i=1}^{N} L_i(\theta).$$

Пратећемо правила општег EM алгоритма, како бисмо дошли до крајњег циља.

За почетак изаберемо иницијалне вредности параметара, односно одаберемо $\theta^{(1)}$. У тренутку t вредност параметара је задата са $\theta^{(t)}$. Да бисмо одредили наредну вредност $\theta^{(t+1)}$ потребно је дефинисати функцију у тачки t која ће представљати доње ограничење функције L, а тачка у којој помоћна функција достиже максимум биће тражена вредност $\theta^{(t+1)}$. Користећи Јенсенову неједнакост за логаритамску функцију,

$$\ln(\sum_{i=1}^{n} \lambda_i x_i) \geq \sum_{i=1}^{n} \lambda_i \ln(x_i),$$

можемо ограничити функцију $L_i(\theta)$ на следећи начин:

$$\ln \left[\sum_{k=1}^{K} \pi_k f_{\mathcal{N}}(\boldsymbol{x_i}|\mu_k, \Sigma_k) \right] \geq \sum_{k=1}^{K} \lambda_{ki} \ln \frac{\pi_k f_{\mathcal{N}}(\boldsymbol{x_i}|\mu_k, \Sigma_k)}{\lambda_{ki}}.$$

Да би Јенсенова неједнакост могла да важи потребно је да буде испуњено: $\sum_{k=1}^{K} \lambda_{ki} = 1$ и $\lambda_{1i}, \lambda_{2i}, ..., \lambda_{Ki} \geq 0$. Нека је

$$\lambda_{ki} = P(Z_i = k|\boldsymbol{X_i} = \boldsymbol{x_i}) = \frac{P(Z_i = k) f_{\boldsymbol{X_i}|Z_i=k}(\boldsymbol{x_i})}{f_{\boldsymbol{X_i}}(\boldsymbol{x_i})} =$$
$$\frac{\pi_k^{(t)} f_{\mathcal{N}}(\boldsymbol{x_i}|\mu_k^{(t)}, \Sigma_k^{(t)})}{\sum_{j=1}^{K} \pi_j^{(t)} f_{\mathcal{N}}(\boldsymbol{x_i}|\mu_j^{(t)}, \Sigma_j^{(t)})} = \gamma_{Z_i}(k).$$

Како су све λ_{ki} рачунате у односу на нашу тренутну позицију, односно у вредности $\theta^{(t)}$, користимо ознаку $^{(t)}$ да то нагласимо.

Како је овако дефинисано λ_{ki} вероватноћа, следи да за свако k важи $\lambda_{ki} \geq 0$. Проверимо сада важење другог потребног услова:

$$\sum_{k=1}^{K} \lambda_{ki} = \sum_{k=1}^{K} \frac{\pi_k^{(t)} f_{\mathcal{N}}(\boldsymbol{x_i}|\mu_k^{(t)}, \Sigma_k^{(t)})}{\sum_{j=1}^{K} \pi_j^{(t)} f_{\mathcal{N}}(\boldsymbol{x_i}|\mu_j^{(t)}, \Sigma_j^{(t)})} = \frac{\sum_{k=1}^{K} \pi_k^{(t)} f_{\mathcal{N}}(\boldsymbol{x_i}|\mu_k^{(t)}, \Sigma_k^{(t)})}{\sum_{j=1}^{K} \pi_j^{(t)} f_{\mathcal{N}}(\boldsymbol{x_i}|\mu_j^{(t)}, \Sigma_j^{(t)})} = 1.$$

Уверили смо се да је овакав одабир λ_{ki} задовољавајући и да је функција која представља доње ограничење функције $L_i(\theta)$ дата са:

$$l_i(\boldsymbol{\theta}, \boldsymbol{\theta^{(t)}}) = \sum_{k=1}^{K} \lambda_{ki} \ln \frac{\pi_k f_{\mathcal{N}}(\boldsymbol{x_i}|\mu_k, \Sigma_k)}{\lambda_{ki}},$$

где је $\lambda_{ki} = \gamma_{Z_i}(k)$. Подсетимо се, други аргумент представља тачку у којој је функција l_i дефинисана.

Остало је још проверити да важи једнакост $L_i(\boldsymbol{\theta^{(t)}}) = l_i(\boldsymbol{\theta^{(t)}}, \boldsymbol{\theta^{(t)}})$.

$$\begin{aligned}
l_i(\boldsymbol{\theta^{(t)}}, \boldsymbol{\theta^{(t)}}) &= \sum_{k=1}^{K} \frac{\pi_k^{(t)} f_{\mathcal{N}}(\boldsymbol{x_i}|\mu_k^{(t)}, \Sigma_k^{(t)})}{\sum_{j=1}^{K} \pi_j^{(t)} f_{\mathcal{N}}(\boldsymbol{x_i}|\mu_j^{(t)}, \Sigma_j^{(t)})} \ln \frac{\frac{\pi_k^{(t)} f_{\mathcal{N}}(\boldsymbol{x_i}|\mu_k^{(t)}, \Sigma_k^{(t)})}{1}}{\frac{\pi_k^{(t)} f_{\mathcal{N}}(\boldsymbol{x_i}|\mu_k^{(t)}, \Sigma_k^{(t)})}{\sum_{j=1}^{K} \pi_j^{(t)} f_{\mathcal{N}}(\boldsymbol{x_i}|\mu_j^{(t)}, \Sigma_j^{(t)})}} \\
&= \sum_{k=1}^{K} \frac{\pi_k^{(t)} f_{\mathcal{N}}(\boldsymbol{x_i}|\mu_k^{(t)}, \Sigma_k^{(t)})}{\sum_{j=1}^{K} \pi_j^{(t)} f_{\mathcal{N}}(\boldsymbol{x_i}|\mu_j^{(t)}, \Sigma_j^{(t)})} \ln \left(\sum_{j=1}^{K} \pi_j^{(t)} f_{\mathcal{N}}(\boldsymbol{x_i}|\mu_j^{(t)}, \Sigma_j^{(t)}) \right) \\
&= \ln \left(\sum_{j=1}^{K} \pi_j^{(t)} f_{\mathcal{N}}(\boldsymbol{x_i}|\mu_j^{(t)}, \Sigma_j^{(t)}) \right) \frac{\sum_{k=1}^{K} \pi_k^{(t)} f_{\mathcal{N}}(\boldsymbol{x_i}|\mu_k^{(t)}, \Sigma_k^{(t)})}{\sum_{j=1}^{K} \pi_j^{(t)} f_{\mathcal{N}}(\boldsymbol{x_i}|\mu_j^{(t)}, \Sigma_j^{(t)})} \\
&= \ln \left(\sum_{j=1}^{K} \pi_j^{(t)} f_{\mathcal{N}}(\boldsymbol{x_i}|\mu_j^{(t)}, \Sigma_j^{(t)}) \right) = L_i(\boldsymbol{\theta^{(t)}}).
\end{aligned}$$

Пошто смо дефинисали функцију у тачки $\theta(t)$ која представља доње ограничење, наредни корак алгоритма је тражење нове вредности параметра.

$$\begin{aligned}
\boldsymbol{\theta^{(t+1)}} &= \max_{\boldsymbol{\theta}} \sum_{i=1}^{N} l_i(\boldsymbol{\theta}, \boldsymbol{\theta^{(t)}}) \\
&= \max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \sum_{k=1}^{K} \lambda_{ki} \ln \left(\frac{\pi_k f_{\mathcal{N}}(\boldsymbol{x_i}|\mu_k, \Sigma_k)}{\lambda_{ki}} \right) \\
&= \max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \sum_{k=1}^{K} \left(\lambda_{ki} \ln \left(\pi_k f_{\mathcal{N}}(\boldsymbol{x_i}|\mu_k, \Sigma_k) \right) - \lambda_{ki} \ln \lambda_{ki} \right).
\end{aligned}$$

Како је $\theta^{(t)}$ фиксирано и максимизујемо само по параметру θ, а λ_{ki} зависи само од $\theta^{(t)}$, следи да део израза који садржи $-\lambda_{ki} \ln \lambda_{ki}$ можемо занемарити јер

не зависи од θ. Односно:

$$\boldsymbol{\theta^{(t+1)}} = \max_{\boldsymbol{\theta}} \sum_{i=1}^{N}\sum_{k=1}^{K} \lambda_{ki}(\ln \pi_k + \ln f_{\mathcal{N}}(\boldsymbol{x_i}|\mu_k, \Sigma_k)).$$

Изједначавањем извода ове функције по параметрима μ, Σ^{-1} и π са нулом, добијамо формуле за рачунање парамерара. Такође, и у овом случају, приликом рачунања извода по параметру π треба узети у обзир услов који ови параметри морају да задовољавају. Након рачунања извода добијене формуле гласе:

$$\mu_k^{(t+1)} = \frac{\sum_{i=1}^{N} \lambda_{ki} x_i}{\sum_{i=1}^{N} \lambda_{ki}} = \frac{\sum_{i=1}^{N} \gamma_{Z_i}(k) x_i}{\sum_{i=1}^{N} \gamma_{Z_i}(k)},$$

$$\Sigma_k^{(t+1)} = \frac{\sum_{i=1}^{N} \lambda_{ki}(x_i - \mu_k^{(t+1)})(x_i - \mu_k^{(t+1)})^T}{\sum_{i=1}^{N} \lambda_{ki}} = \frac{\sum_{i=1}^{N} \gamma_{Z_i}(k)(x_i - \mu_k^{(t+1)})(x_i - \mu_k^{(t+1)})^T}{\sum_{i=1}^{N} \gamma_{Z_i}(k)},$$

$$\pi_k^{(t+1)} = \frac{\sum_{i=1}^{N} \lambda_{ki}}{N} = \frac{\sum_{i=1}^{N} \gamma_{Z_i}(k)}{N}.$$

Можемо закључити да су изведене формуле у складу са формулама представљеним у првом делу овог поглавља, што указује да оба начина извођења пружају идентичне резултате.

4.5 Веза између кластеровања коришћењем K средина и мешавине вишедимензионих нормалних расподела

Задовољавањем одређених услова, кластеровање које користи мешавине нормалних расподела се своди на алгоритам K средина. Уколико би у кластеровању мешавинама следећи услови били задовољени:

- $\Sigma_k = \sigma^2 I, \quad \forall k \in \{1, ..., K\}$,
- $\pi_k = \frac{1}{K}, \quad \forall k \in \{1, ..., K\}$,
- $\gamma_{Z_i}(k) = \begin{cases} 1, & \text{ако је тачки } x_i \text{ најближи центар } \mu_k, \\ 0, & \text{иначе,} \end{cases}$

односно, уколико би кластери имали једнак удео у подацима, били сферног облика и припадање кластеру била бинарна вредност која би се одређивала на основу најмање удаљености од центроида, тада би се добило кластеровање алгоритмом K средина (видети [12]).

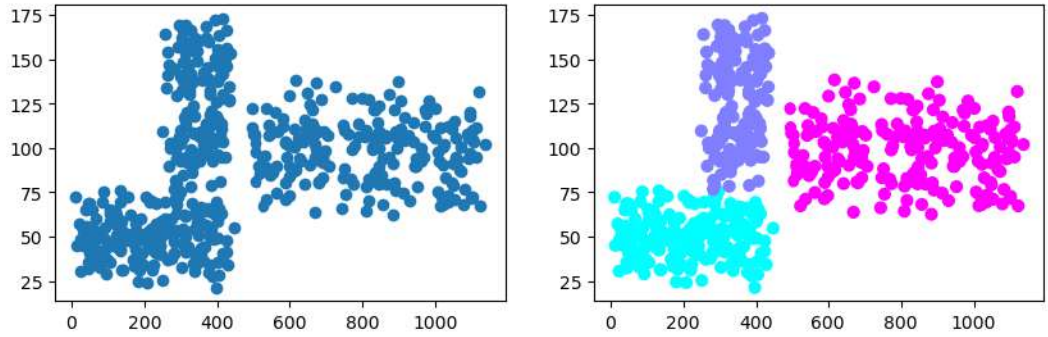
Из претходно наведеног можемо закључити да је кластеровање мешавинама општије, односно може се примењивати како на кластере сферног облика тако и на елипсоидне кластере.

Наредни пример приказаће како фокусирање на сферне кластере утиче на перформансе алгоритма K средина и како су ови проблеми превазиђени његовом надоградњом на кластеровање употребом мешавина.

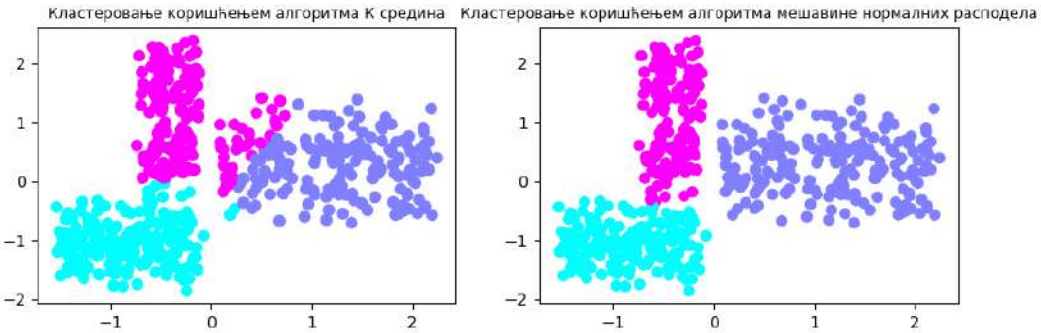
Пример 4.5.1. *Посматрајмо вештачки генерисан узорак, са три очигледне подгрупе, приказан на слици 4.5. Јасно се може приметити да подскупови нису сферног облика. Пре почетка сваког процеса кластеровања потребно је извршити скалирање скупа које ће обезбедити да тачке буду на истој скали како би рачунање растојања међу њима имало смисла. На посматраном узорку скалирање је вршено тако што је од сваке тачке одузета просечна вредност свих тачака, услед чега су ови бројеви подељени укупном стандардном девијацијом узорка. Након скалирања, примењена су оба алгоритма кластеровања како бисмо упоредили резултате и видели да ли кластери који нису кружног облика проузрокују неке разлике у алгоритмима. Пошто се ради о вештачки генерисаном скупу за потребе илустровања рада алгоритама, кластеровање*

смо вршили коришћењем три кластера. О начину одабира броја кластера у општем случају биће речи у наредном поглављу.

Слика 4.6 приказује резултате кластеровања, односно, тачке приказане истом бојом су тачке за које су алгоритми сматрали да треба да се нађу у истом кластеру. Можемо приметити да постоји видна разлика у облицима добијених кластера. Док је кластеровање мешавинама успело тачно да уочи дугуљасте кластере, алгоритам К средина је ипак тежио да кластери буду што сфернијег облика што је проузроковало погрешно груписање одређених инстанци.



Слика 4.5: Приказ вештачки генерисаних тачака и њихова припадност кластерима



Слика 4.6: Приказ резултата коришћења оба алгоритма

Претходни пример уверио нас је у ограничења алгоритма К средина и приказао боље сналажење мешавине нормалних расподела у кластеровању општих података.

Глава 5

Начини одређивања броја кластера

С обзиром да кластеровање представља ненадгледани вид учења, а у решавању проблема није увек експилицно наглашено колико кластера треба имати, јасно је да овај број може варирати од задатка до задатка. Важно је напоменути да не постоји један универзални начин за одређивање оптималног броја кластера. Избор броја кластера зависи од природе података и циља анализе. Поседовање доменског знања може бити од велике помоћи приликом одабира, међутим уколико то није случај експериментисање са различитим бројевима кластера и коришћењем неких од техника евалуације можемо одредити количину која даје најбоље резултате.

Вероватно би већини људи, који су се сустретали са кластеровањем, на начин одређивања броја кластера прва помисао била правило лакта, међутим у случају кластеровања помоћу мешавина нормалних расподела оно није много корисно. Наиме, правило лакта се базира на опадању инерције која представља збир квадрата растојања сваке тачке од њој најближе центроиде. Како оваква техника иде у прилог кластерима сферног облика, не можемо је користити за наш проблем јер приликом кластеровања коришћењем мешавина нормалних расподела кластери могу бити елипсоидног облика, односно тачке не морају да имају што мање растојање од својих центроида. Из тог разлога, за решавање овог проблема користе се другачије методе као што су силуета и Бајесов или Акаикеов информациони критеријум које дају боље резултате. Ипак, треба имати на уму да различити начини за одређивање броја кластера могу давати различите резултате, те тако поред ових метода приликом доношења одлуке о броју кластера треба познавати и проблем који се решава, као и саме податке са којима се ради.

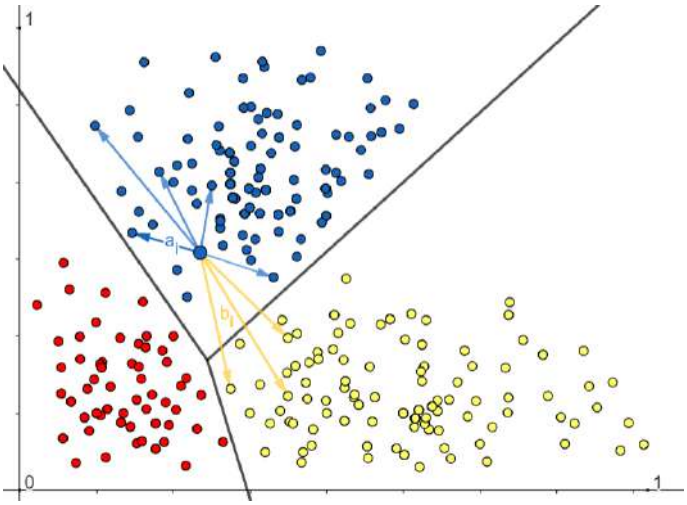
Силуета

Силуета представља меру за одређивање квалитета кластеровања. Приликом рачунања вредности овог коефицијента узима се у обзир растојање између инстанци унутар истог кластера, као и растојање између тачака једног кластера и припадника најближег суседног кластера. Овако задата мера нам омогућава да нађемо што бољу вредност броја кластера за коју ће важити да су припадници једног кластера близу, док је међусобна удаљеност између кластера што већа. Силуету за инстанцу i означавамо са $s(i)$ и рачунамо на следећи начин:

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))},$$

где $a(i)$ представља средњу вредност раздаљина између инстанце i и свих осталих припадника њеног кластера, а вредност $b(i)$ средњу вредност раздаљина између инстанце i и инстанци њој најближег кластера. Начин на који одређујемо инстанци најближи кластер је тај да израчунамо просечно растојање инстанце са инстанцама сваког кластера, а затим одаберемо кластер за који је овај број најмањи.

На слици 5.1 приказане су раздаљине потребне за рачунање силуете.



Слика 5.1: Приказ дужина коришћених за рачунање силуете

Крајњи коефицијент, за узорак од N тачака, је један број који добијамо коришћењем форуме:

$$s = \frac{1}{N} \sum_{i=1}^{N} s(i).$$

Вредности силуете налазе се у интервалу $[-1, 1]$, где различите вредности указују на различит квалитет кластеровања. Коефицијент који тежи вредности 1 представља најбољи квалитет и значи да су подаци добро груписани, а кластери добро раздвојени. Вредности у околини нуле нам поручују да су кластери близу и да не можемо бити сигурни у добијене резултате, а вредности које теже ка -1 указују да је кластеровање лоше јер долази до преклапања кластера.

С обзиром да се налажење овог коефицијента заснива на коришћењу растојања, важно је нагласити да се он може користити како за Еуклидско тако и за било које друго растојање.

Акаикеов информациони критеријум

Акаикеов информациони критеријум (AIC) је мера која се користи за упоређивање модела са циљем да се изабере модел који боље одговара подацима. Формула за рачунање овог критеријума гласи:

$$AIC = -2 \ln L(\theta) + 2k,$$

где је L функција веродостојности, а k број параметара модела. Први део формуле односи се на меру колико модел добро објашњава податке, док је други део везан за сложеност модела. Тачније, Акаикеов информациони критеријум кажњава моделе који имају превелики број параметара, јер то представља ризик за преприлагођавање. С друге стране, даје предност моделима који најбоље одговарају подацима. Комбинујући ова два дела, Акаикеов информациони критеријум представља меру која узима у обзир однос између квалитета и сложености модела. Модел са најмањом вредношћу овог критеријума се сматра најбољим јер, у поређењу са осталим, представља најједноставнији модел који успева да ухвати највише законитости у подацима. У контексту одређивања броја кластера, AIC можемо користити тако што ћемо груписати податке за различите бројеве кластера,а потом за свако од кластеровања израчунати вредност овог критеријума. За кластеровање са најмањом вредношћу овог критеријума сматрамо да је најбоље и његов број кластера узимамо као тражени број.

Бајесов информациони критеријум

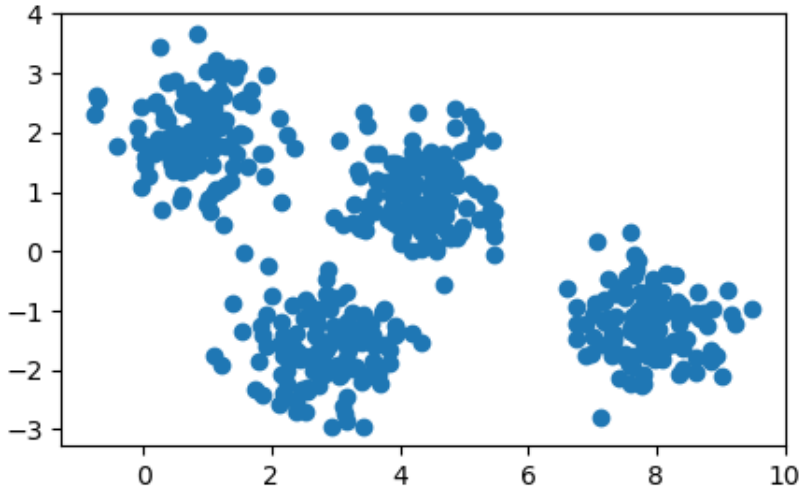
Бајесов информациони критеријум (BIC) је такође мера која се користи за упоређивање модела. Настао је као потреба за већим кажњавањем модела са више параметара стога представља строжу надоградњу претходног критеријума. Формула за рачунање овог критеријума гласи:

$$BIC = -2 \ln L(\theta) + ln(n)k,$$

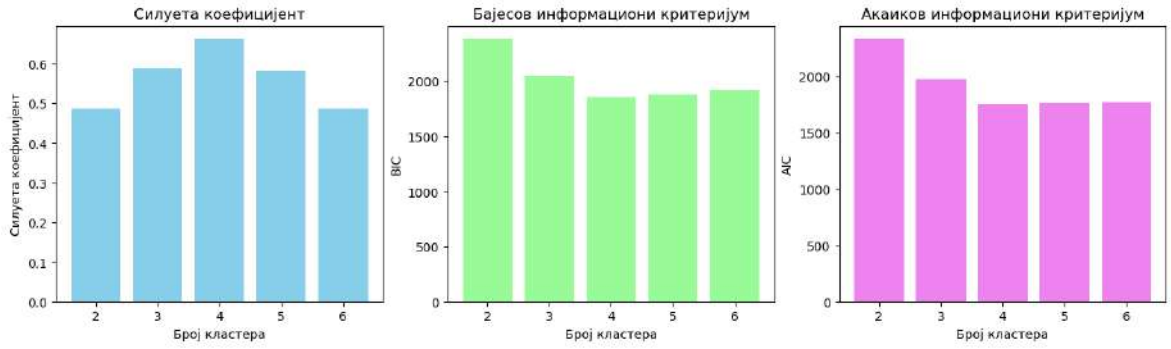
где је L функција веродостојности, k број параметара модела, а n број тачака у узорку. За разлику од претходног критеријума, Бајесов информациони критеријум узима у обзир и величину узорка користећи га за повећавање сабирка који служи за спречавање преприлагођавања. Како је $ln(n) > 2$ за $n > 8$ можемо се уверити у већ поменуту тврдњу, да BIC више узима у обзир сложеност модела. Принцип рада са Бајесовим критеријумом је идентичан као са претходним, модел са најмањом вредношћу овог критеријума представља најбољи баланс између ефикасности рада и сложености.

У наредним примерима приказаћемо рад ових начина за одређивање броја кластера.

Пример 5.0.1. *Генеришимо вештачки узорак који ће у себи садржати четири кластера и испробајмо различите критеријуме за налажење броја кластера. На слици 5.2 можемо јасно уочити 4 кластера, али и то да постоје неке тачке за које не можемо бити тачно сигурни ком кластеру припадају. Након скалирања података и тренирања мешавине нормалних расподела са различитим бројем кластера (од два до шест) покренули смо претходно поменуте начине како бисмо утврдили одговарајући број кластера. На слици 5.3 можемо посматрати добијене резултате. Силуета има највишу вредност за четири кластера. Слично, у случају Акаикеовог и Бајесовог информационог критеријума најмања вредност је за четири кластера, међутим за разлику од силуете ове вредности су јако сличне за пет кластера. Узимајући у обзир комбинацију ова три критеријума, свакако бисмо се одлучили за четири кластера.*



Слика 5.2: Приказ вештачки генерисаног узорка са 4 кластера

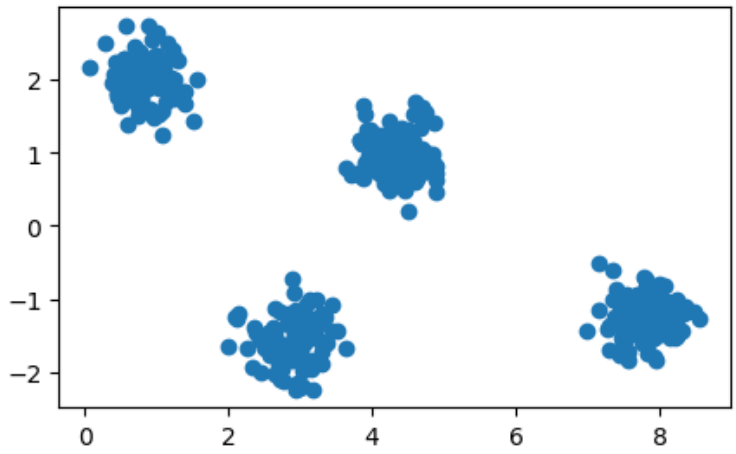


Слика 5.3: Приказ резултата

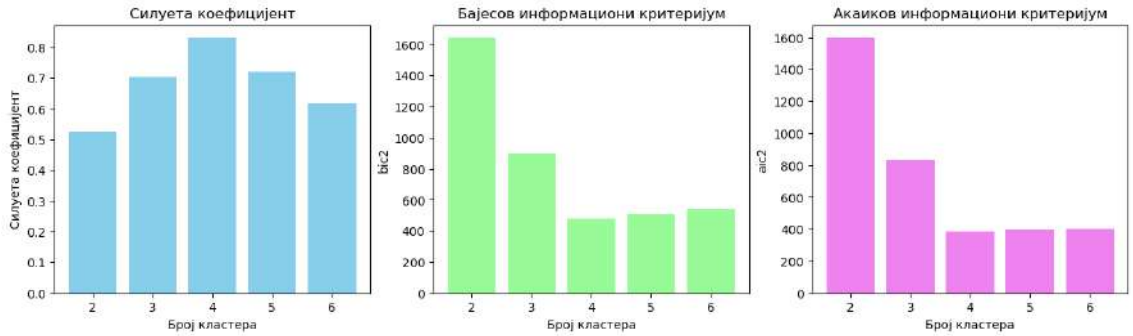
Посматрајмо пример у ком су наши кластери боље раздвојени, како бисмо видели какво је понашање критеријума у овом случају.

Пример 5.0.2. *Слично као у претходном примеру генерисаћемо узорак са четири кластера, међутим кластери ће бити боље раздвојени. Као што се може приметити са слике 5.4, више немамо тачке за које нисмо сигурни ком кластеру припадају. Поновићемо исти поступак као у претходном примеру и посматрати резултате са слике 5.5. И даље, силуета је најизраженија за 4 кластера, међутим можемо приметити промену у његовој вредности. У претходном примеру вредност овог коефицијента била је нешто већа од 0,6*

док је сада изнад 0,8. Разлог за овај скок долази од веће сигурности да модел добро групише податке, односно модел има мање тачака за које није сигуран ком кластеру припадају. Што се тиче Акаикеовог и Бајесовог информационог критеријума, иако су вредности остале сличне за четири и пет кластера, примећујемо да модел прави знатно већу разлику између кластеровања са два или три кластера и кластеровања са четири кластера.



Слика 5.4: Приказ вештачки генерисаног узорка са 4 кластера

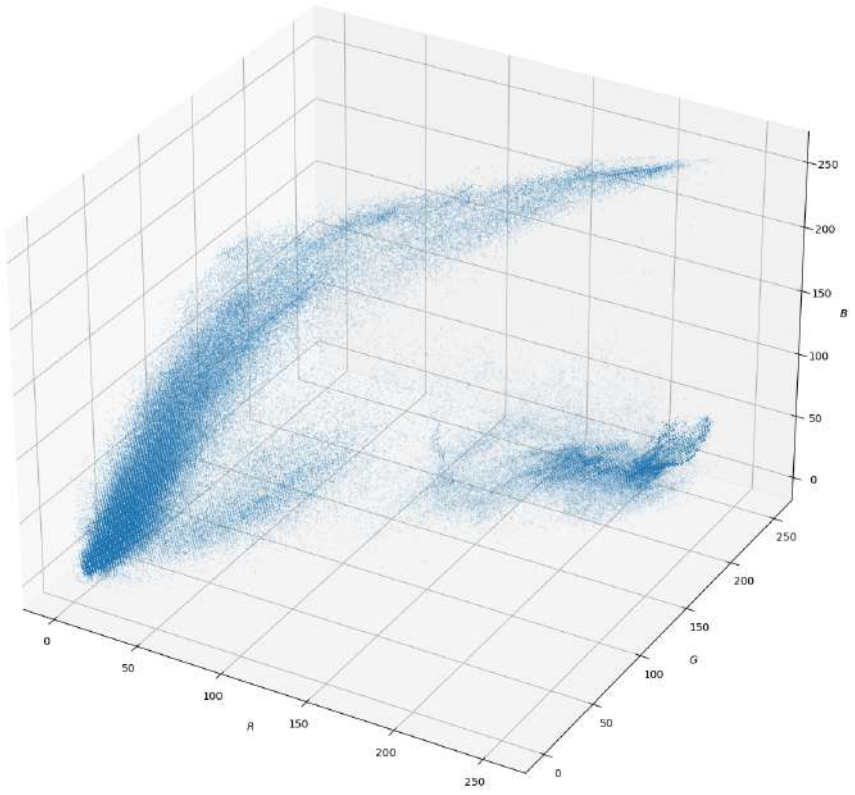


Слика 5.5: Приказ вештачки генерисаног узорка са 4 кластера

Наредни пример приказаће рад метода за одабир броја кластера у случају тродимензионог узорка. У остатку рада биће речи о кластеровању слика и њиховој репрезентацији у простору основних боја, због чега ћемо у овом примеру слику посматрати само као тродимензионе податке на које желимо да применимо поменуте методе.

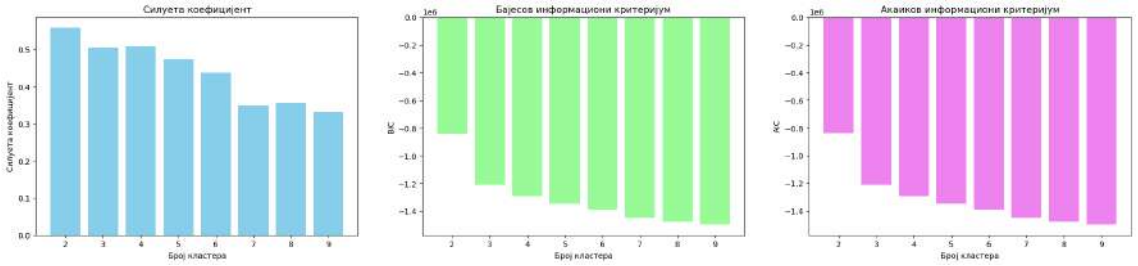
Пример 5.0.3. *Нека слика 5.6 представља фотографију за коју желимо да знамо колико кластера треба применити приликом кластеровања, док слика 5.7 представља репрезентацију фотографије у тродимензионом простору.*

Слика 5.6: Фотографија за коју желимо да одредимо број кластера



Слика 5.7: Приказ слике 5.6 у тродимензионом простору

На основу 3Д приказа можемо закључити да пригодан број кластера није лако уочљив и да немамо јасно одвојене групације у простору. Након испробавања мешавине вишедимензионих нормалних расподела за различите бројеве кластера и рачунања метрика за одређивање броја кластера, добили смо следеће резултате:



Слика 5.8: Резултати кластеровања са различитим бројем кластера

Можемо приметити да овог пута методе нису усаглашене као што је то био случај са ранијим примерима. Док метод силуете предлаже два кластера, ситуација је доста другачија за Бајесов и Акаикеов информацији критеријум који сугеришу да би за број кластера требало одабрати неки од бројева седам, осам или девет.

Као што можемо видети у овом примеру, дешава се да овакав приступ неће загарантовано дати један број који ћемо искористи, већ за одабир крајњег броја треба узети у обзир и друге информације.

Од велике је важности да разумемо задатак којим се бавимо и да на основу њега одаберемо смислен број кластера. Такође, треба имати на уму да се задати проблем можда не може решити једном итерацијом кластеровања, те тако најбољи број кластера и не постоји. У оваквим случајевима почнемо од неког броја кластера, па на основу резултата вршимо додатне итерације за кластере којима је то потребно. Уколико нам ресурси дозвољавају, некад број кластера можемо одабрати и испробавањем различитих бројева на основу чијих резултата бисмо донели крајњу одлуку.

С обзиром да задатак одабира броја кластера може бити компликован и зависи од више фактора, не треба слепо веровати методама, чак и у случајевима када су усаглашене и дају исти број. Ове методе треба користити као смернице, а на уму увек имати ширу слику и резултате које желимо да постигнемо.

Глава 6

Примери коришћења мешавине нормалних расподела за кластеровање података

Иако мешавине нормалних расподела имају широку примену у различитим задацима, у овом поглављу акценат ће бити на приказу ефикасности мешавина у кластеровању података. Решавање овог проблема од велике је важности у многим областима, па тако мешавине нормалних расподела могу наћи примену у различитим сферама укључујући маркетингу, банкарство или медицину.

За свако добро пословање неопходно је осигурати да производ који се нуди буде што боље прилагођен купцима, односно циљној групи којој је намењен. Тачније, за компаније које нуде различите производе и услуге битно је ускладити понуде са интересима одређене групе потрошача. Овакав вид рекламирања изразито је важан у данашње време и доводи до постизања бољих резултата компанија. За решавање овог проблема кластеровање представља неизоставни корак, о чему ће бити више речи у наставку овог поглавља.

Још једна од области у којој кластеровање има незаобилазну улогу јесте област компјутерске визије која је стекла огромну популарност последњих година, а чије се нове примене непрестано проналазе. Задатак компјутерске визије представља и сегментација слике за коју се често користе мешавине нормалних расподела.

Наставак рада има за циљ да кроз практичне примере илуструје примену мешавина нормалних расподела у поменутим областима. Ове области нису одабране само због приказивања употребне вредности мешавина кроз стварне

задатке, већ и са намером да се скрене пажња на различитост њихове примене и широк спектар проблема за које могу бити коришћене.

6.1 Кластеровање корисника

Приказаћемо сада један једноставан пример кластеровања корисника који ће илустровати на који начин овакав приступ рекламирању може побољшати резултате пословања.

Пример 6.1.1. *Замислимо да неки мобилни оператер жели да побољша своје резултате тако што ће промовисати нове мобилне пакете који ће одговарати различитим групама потрошача. Компанија поседује податке о корисницима заједно са просечним бројем месечних позива и просечном количином искоришћених мегабајта, минута и роминг минута током месеца. Како бисмо груписали потрошаче сходно њиховом коришћењу услуга, извршили смо кластеровање помоћу мешавине нормалних расподела и добили следеће статистике:*

	count	mean	std	min	10%	25%	50%	75%	90%	95%	max
Интернет_потрошња	987.0	0.081733	0.206411	0.0	0.0	0.0	0.0	0.0	0.31	0.35	2.48
Број_искоришћених_минута	987.0	133.138602	34.761982	2.6	84.5	113.2	140.0	159.3	170.78	177.84	193.00
Број_позива	987.0	102.945289	18.677715	48.0	79.0	90.0	103.0	116.0	127.00	134.00	165.00
Месечни_рачун	987.0	40.563019	6.855696	15.7	31.0	37.0	41.0	45.9	49.00	50.00	55.00
Број_искоришћених_роминг_минута	987.0	11.111246	2.145727	6.3	8.5	9.6	11.0	12.5	13.94	14.80	18.90

Слика 6.1: Статистике првог кластера

	count	mean	std	min	10%	25%	50%	75%	90%	95%	max
Интернет_потрошња	768.0	2.982435	0.596483	1.7	2.24	2.570	2.94	3.3575	3.78	4.050	5.4
Број_искоришћених_минута	768.0	184.846615	52.024877	40.9	115.97	149.850	185.50	220.1000	251.84	271.255	322.4
Број_позива	768.0	100.703125	19.741363	35.0	77.00	88.000	101.00	114.0000	126.00	135.000	163.0
Месечни_рачун	768.0	79.139453	10.825799	49.0	65.87	71.675	79.00	86.4250	93.13	97.000	111.3
Број_искоришћених_роминг_минута	768.0	11.044010	2.209513	6.3	8.30	9.500	10.90	12.4250	14.00	15.000	20.0

Слика 6.2: Статистике другог кластера

	count	mean	std	min	10%	25%	50%	75%	90%	95%	max
Интернет_потрошња	877.0	0.069829	0.129008	0.0	0.0	0.0	0.0	0.0	0.31	0.34	0.51
Број_искоришћених_минута	877.0	225.977993	33.866129	166.4	188.9	198.7	220.4	245.5	272.90	289.92	350.80
Број_позива	877.0	104.420753	17.987050	58.0	81.0	92.0	105.0	117.0	127.00	134.00	157.00
Месечни_рачун	877.0	56.940023	6.907428	36.0	49.0	52.0	56.0	61.0	66.50	70.00	79.00
Број_искоришћених_роминг_минута	877.0	11.317788	2.088412	6.9	8.8	9.8	11.1	12.7	14.10	15.00	18.40

Слика 6.3: Статистике трећег кластера

	count	mean	std	min	10%	25%	50%	75%	90%	95%	max
Интернет_потрошња	701.0	0.412111	0.684487	0.0	0.0	0.0	0.0	0.35	1.76	1.97	2.32
Број_искоришћених_минута	701.0	182.079458	47.367897	0.0	124.3	152.5	182.1	211.10	240.80	261.70	346.80
Број_позива	701.0	91.623395	22.073985	0.0	63.0	77.0	92.0	106.00	120.00	129.00	160.00
Месечни_рачун	701.0	52.658916	9.731147	14.0	41.0	46.0	52.4	59.00	65.10	68.40	84.00
Број_искоришћених_роминг_минута	701.0	6.771184	2.126956	0.0	4.4	5.8	6.8	8.00	9.10	10.00	13.30

Слика 6.4: Статистике четвртог кластера

Приликом посматрања статистика, можемо уочити да највећу потрошњу интернета имају корисници који су се нашли у другом кластеру. Такође, ова варијабла варира унутар четвртог кластера, чинећи да се у њему налазе особе које користе интернет, али и они који то јако ретко чине. Како бисмо добили хомогеније резултате урадићемо још једну итерацију кластеровања поменутог кластера. Кластере из ове итерације означићемо са 4_1 *и* 4_2*, како би било јасно да они воде порекло из четвртог кластера. Овим поступком, гледајући добијене резултате са слика 6.5 і 6.6, можемо закључити да смо успели да одвојимо кориснике интернет услуга и на тај начин добили крајњих пет кластера.*

	count	mean	std	min	10%	25%	50%	75%	90%	95%	max
Интернет_потрошња	584.0	0.132791	0.282059	0.0	0.00	0.000	0.00	0.240	0.35	0.4485	1.76
Број_искоришћених_минута	584.0	186.601884	46.989970	0.0	132.03	158.775	185.55	214.325	245.05	264.5650	346.80
Број_позива	584.0	90.410959	22.762227	0.0	61.00	75.000	90.00	105.000	120.70	128.8500	160.00
Месечни_рачун	584.0	50.744007	9.073449	14.0	40.00	45.000	51.00	56.000	62.10	65.6700	84.00
Број_искоришћених_роминг_минута	584.0	6.787842	2.286602	0.0	4.20	5.600	6.90	8.200	9.20	10.2000	13.30

Слика 6.5: Статистике првог кластера добијеног кластеровањем четвртог

	count	mean	std	min	10%	25%	50%	75%	90%	95%	max
Интернет_потрошња	117.0	1.806325	0.272149	1.13	1.448	1.59	1.81	2.03	2.142	2.21	2.32
Број_искоришћених_минута	117.0	159.505983	42.724269	58.80	105.080	127.20	161.20	192.30	210.920	226.24	264.30
Број_позива	117.0	97.675214	17.088930	59.00	75.200	87.00	98.00	107.00	117.400	129.20	138.00
Месечни_рачун	117.0	62.217094	6.860246	44.20	53.400	57.30	62.30	66.70	70.660	73.24	80.00
Број_искоришћених_роминг_минута	117.0	6.688034	1.006630	4.20	5.360	5.90	6.70	7.50	7.940	8.20	8.60

Слика 6.6: Статистике другог кластера добијеног кластеровањем четвртог

Анализирањем и упоређивањем статистика кластера можемо закључити следеће:

- *Потрошачи из кластера 1 не користе интернет, троше најмање минута током месеца, у поређењу са другим кластерима, али имају велики број позива, што нас наводи на закључак да припадници овог кластера обављају доста кратких позива. Овим корисницима би се највише исплатило понудити пакете које карактерише бесплатно успостављање везе.*

- *Кластер 2 одликује највећа потрошња интернета и просечно коришћење месечних позива и минута. Сходно већој активности, рачуни ових корисника су највећи, па би се у понуди овој групи нашли и мегабајти интернета и бесплатни минути.*

- *У кластеру 3 налазе се корисници који имају изражну потрошњу месечних минута и позива, односно корисници који обављају пуно дугих позива. Будући да не користе интернет, овим потрошачима били би понуђени пакети који обухватају бесплатне минуте.*

- *Потрошачи који су се нашли у кластеру* 4_1 *имају најмање позива у односу на потрошаче из осталих кластера, троше просечан број минута и не користе интернет. Можемо закључити да ови корисници обављају дуге позиве, али то не чине са истом учесталошћу као припадници претходног кластера, па би за њих најбоље било обезбедити повољнији пакет са мањим бројем бесплатиних минута.*

- *Последњи кластер издвојио се из првобитног четвртог кластера по коришћењу интернета. Што се осталих особина тиче, можемо приметити да је потрошња минута исподпросечна, а број позива уобичајен. Пакети ових корисника морали би да садрже бесплатне мегабајте интернета, са евентуалним додатком бесплатне успоставе везе.*

Како се у овом примеру подаци налазе у петодимензионом простору, то за сваки од почетна четири кластера имамо петодимензиони вектор центроиде на основу ког можемо сагледати положај сваког од кластера.

$$\begin{aligned}
\boldsymbol{\mu_1} &= [-0.57, 0.83, 0.16, 0.04, 0.29]^T, \\
\boldsymbol{\mu_2} &= [1.69, 0.12, 0.01, 1.40, 0.29]^T, \\
\boldsymbol{\mu_3} &= [-0.57, -0.78, 0.15, -0.90, 0.34]^T, \\
\boldsymbol{\mu_4} &= [-0.32, -0.07, -0.39, -0.29, -1.06]^T.
\end{aligned}$$

Што се тиче вектора тежина, он је након прве итерације задат са:

$$\boldsymbol{\pi} = [0.26, 0.23, 0.29, 0.22]^T.$$

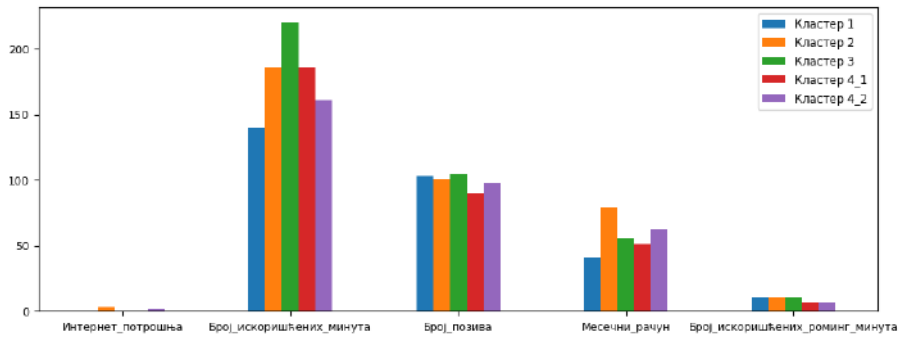
Можемо закључити да је трећи кластер незнатно доминантнији, али да свака од компоненти има јако сличан удео у мешавини.

Приликом коришћења мешавина нормалних расподела у кластеровању могуће је одабрати један од четири типа коваријационе матрице који нам помажу уколико желимо унапред да одредимо облике мешавина. Потпуна (eng. full) коваријациона матрица представља општи случај приликом ког све компоненте мешавине имају своју коваријациону матрицу и могу заузети било какав облик у простору. Везана коваријациона матрица (eng. tied) користи се у случају када компоненте имају исти облик па деле исту коваријациону матрицу. Уколико се за тип коваријационе матрице користи параметар „diag", ради се о дијагоналној коваријационој матрици за сваку компоненту која обезбеђује да се мешавине простиру дуж координатних оса простора. Последњи тип представљају сферне матрице (eng. spherical) где свака од компоненти има истоимени облик. У овом примеру искоришћен је сферни облик коваријационих матрица тако да за сваки од кластера има своју дисперзију која се може видети у вектору $\boldsymbol{\Sigma} = [0.44, 0.65, 0.45, 0.68]^T$. *Крајњи параметри мешавина за другу итерацију кластеровања дати су са:*

$$\boldsymbol{\mu_{4_1}} = [-0.40, 0.07, -0.05, -0.21, 0.01]^T,$$
$$\boldsymbol{\mu_{4_2}} = [2, -0.35, 0.26, 1.09, -0.05]^T,$$
$$\boldsymbol{\pi} = [0.84, 0.16]^T,$$
$$\boldsymbol{\Sigma} = [0.85, 0.48]^T.$$

Слика 6.7 илуструје однос медијана колона које су учествовале у кластеровању и омогућава лакше сагледавање понашања сваког од кластера.

Овај приступ представља један једноставан пример кластеровања потрошача, а у зависности од разноврсности расположивих података могу се извршити различите занимљиве анализе корисника.



Слика 6.7: Медијане колона од интереса приказане за различите кластере

У овом примеру приметили смо како је било потребно извршити две итерације кластеровања да бисмо добили жељене резултате. Некада повећавање броја кластера неће обезбедити гранулирање нама потребног кластера, већ додатни кластер може представљати комбинацију елеманата из више различитих кластера.

Овакав случај можемо посматрати у резултатима кластеровања коришћењем пет кластера. Анализирајући статистике кластера, уочавамо да су оне сличне статистикама добијеним коришћењем четири кластера. Односно, да се приметити да унутар нових пет кластера постоје четири кластера јако слична кластерима из првог решења. Коришћењем оваквог приступа, статистике четвртог кластера остају непромењене, резултујући задржавањем нехомогеног кластера. Додатно, новонастали кластер, кластер три, никако не доприноси повећавању броја група које садрже кориснике интернет услуга.

	count	mean	std	min	10%	25%	50%	75%	90%	95%	max
Интернет_потрошња	719.0	0.071433	0.130247	0.0	0.00	0.00	0.0	0.00	0.31	0.341	0.51
Број_искоришћених_минута	719.0	234.266759	33.069484	170.4	194.80	208.35	229.8	255.45	278.42	294.910	350.80
Број_позива	719.0	107.723227	16.018875	61.0	87.00	97.00	107.0	118.00	128.00	135.000	156.00
Месечни_рачун	719.0	58.437552	6.731486	39.0	50.08	54.00	57.8	63.00	68.00	70.550	79.00
Број_искоришћених_роминг_минута	719.0	10.890125	2.146111	5.8	8.20	9.40	10.7	12.30	13.80	14.610	18.30

Слика 6.8: Статистике првог кластера

	count	mean	std	min	10%	25%	50%	75%	90%	95%	max
Интернет_потрошња	787.0	2.953494	0.625280	0.0	2.19	2.54	2.92	3.32	3.78	4.05	5.4
Број_искоришћених_минута	787.0	185.034562	52.523638	40.9	116.42	149.55	185.40	220.10	252.90	272.85	346.8
Број_позива	787.0	100.601017	19.818538	35.0	77.00	88.00	101.00	114.00	126.00	135.00	163.0
Месечни_рачун	787.0	78.897078	10.855138	49.0	65.22	70.95	78.90	86.15	93.10	96.97	111.3
Број_искоришћених_роминг_минута	787.0	10.970394	2.252084	6.1	8.10	9.40	10.90	12.40	14.00	15.00	20.0

Слика 6.9: Статистике другог кластера

	count	mean	std	min	10%	25%	50%	75%	90%	95%	max
Интернет_потрошња	715.0	0.093245	0.260336	0.0	0.00	0.00	0.0	0.0	0.31	0.36	2.48
Број_искоришћених_минута	715.0	131.700839	38.577259	2.6	75.62	107.25	138.5	160.2	177.02	182.79	209.20
Број_позива	715.0	113.981818	14.771024	68.0	96.00	104.00	113.0	123.0	133.60	141.00	165.00
Месечни_рачун	715.0	40.361818	7.402682	15.7	30.00	36.00	41.0	46.0	49.20	51.00	57.30
Број_искоришћених_роминг_минута	715.0	10.559161	2.158297	4.6	7.90	9.10	10.3	11.9	13.30	14.30	18.20

Слика 6.10: Статистике трећег кластера

	count	mean	std	min	10%	25%	50%	75%	90%	95%	max
Интернет_потрошња	579.0	0.417582	0.675112	0.0	0.00	0.00	0.0	0.35	1.73	1.92	2.32
Број_искоришћених_минута	579.0	179.026943	47.658620	0.0	117.76	149.35	180.5	209.60	236.26	254.94	332.90
Број_позива	579.0	91.027634	19.584683	0.0	67.00	79.00	92.0	103.00	115.00	122.00	151.00
Месечни_рачун	579.0	52.158549	9.873034	14.0	40.00	46.00	52.0	59.00	65.00	67.91	84.00
Број_искоришћених_роминг_минута	579.0	6.380138	1.926604	0.0	4.18	5.50	6.6	7.70	8.50	8.80	11.70

Слика 6.11: Статистике четвртог кластера

	count	mean	std	min	10%	25%	50%	75%	90%	95%	max
Интернет_потрошња	533.0	0.069606	0.128743	0.0	0.00	0.0	0.0	0.0	0.31	0.35	0.46
Број_искоришћених_минута	533.0	163.804315	33.878210	35.1	119.04	140.7	165.7	190.1	206.46	215.42	242.60
Број_позива	533.0	82.409006	12.086844	45.0	66.00	74.0	84.0	92.0	97.80	100.00	107.00
Месечни_рачун	533.0	45.962477	6.818549	22.0	37.00	41.0	46.0	50.0	54.60	57.00	63.30
Број_искоришћених_роминг_минута	533.0	12.032458	1.972884	8.4	9.70	10.6	11.8	13.2	14.70	15.64	18.90

Слика 6.12: Статистике петог кластера

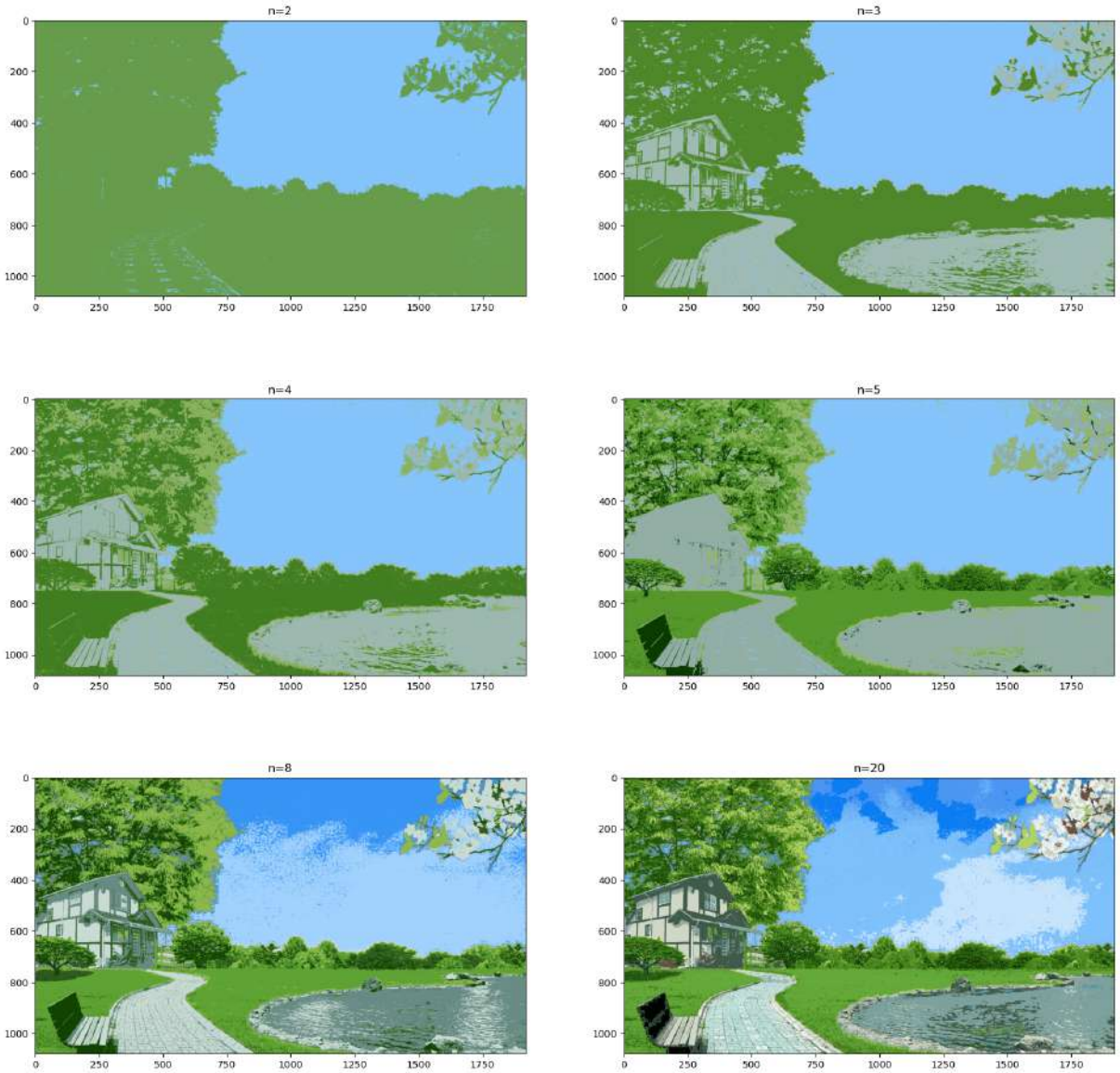
6.2 Сегментација слике

Сегментација слике представља задатак поделе слике на више сегмената, односно целина, омогућавајући лакше уочавање објеката или детектовање ивица на слици. Због упрошћавања слике, овај поступак неопходан је за ефикаснију анализу и процесирање слика и користи се у многим проблемима компјутерског вида. Идеја сегментације је груписати пикселе сличних карактеристика, односно извршити кластеровање и сваком пикселу доделити групу којој припада.

Илуструјмо сада коришћење мешавине нормалних расподела у решавању овог проблема.

Пример 6.2.1. *На слици 6.8 налази се оригинална слика за коју ћемо илустровати различите резултате кластеровања. Сегментација слике постиже се груписањем пиксела, након чега се сваком од пиксела из одређене групе додељује иста вредност, најчешће средња вредност или медијана елемената кластера. Овакав поступак резултује истом бојом пиксела унутар једног кластера, чинећи лакше уочавање регија на слици. Слика 6.9 приказује резултате кластеровања коришћењем мешавине нормалних расподела за различити одабир броја кластера. Уочавамо да су главне информације које слика пружа остале сачуване, при чему приликом повећавања броја кластера добијамо слику која поседује више детаља и сличнија је оригиналу.*

Слика 6.13: Почетна слика на коју желимо да применимо процес кластеровања

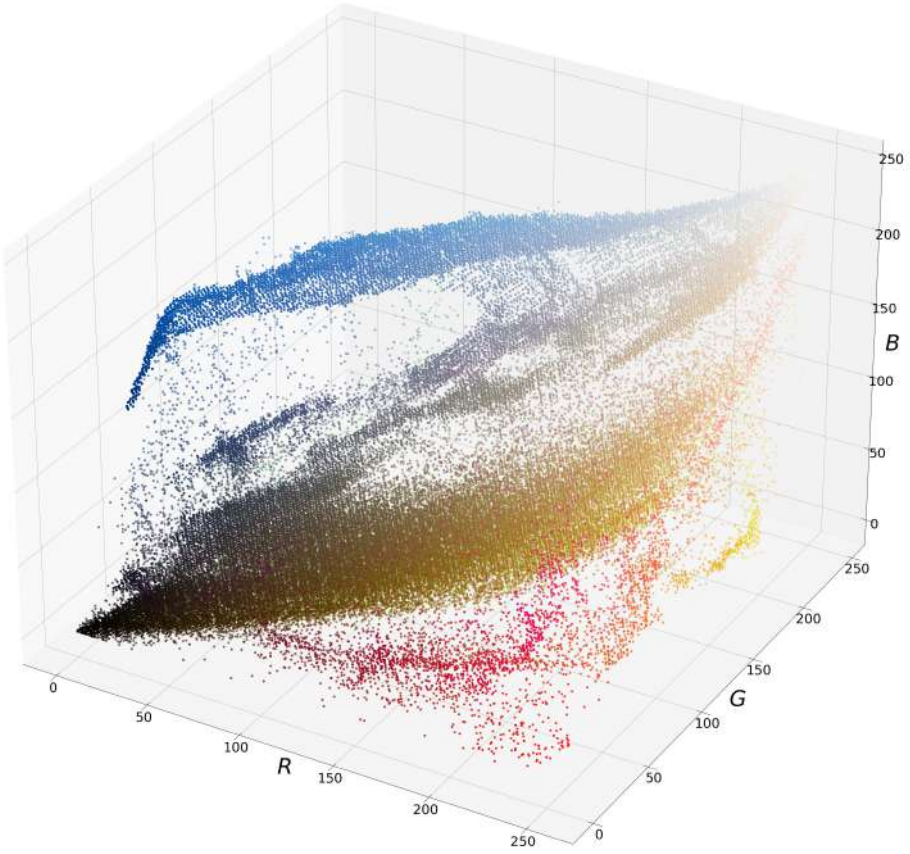


Слика 6.14: Резултати кластеровања слике коришћењем мешавине нормалних расподела са различитим бројем кластера

Свака боја може се описати помоћу три основне боје (црвене, зелене и плаве), комбинујући њихове различите уделе. С обзиром да је основна карактеристика пиксела његова боја, следи да све пикселе можемо сместити у простор три основне боје (РГБ простор), односно тродимензиони простор чија свака оса одговара интензитету једне основне боје. Наредни пример илустроваће кластеровање слике посматрано из угла овог простора, као и разлику између сегментације слике помоћу мешавине нормалних расподела и К средина.

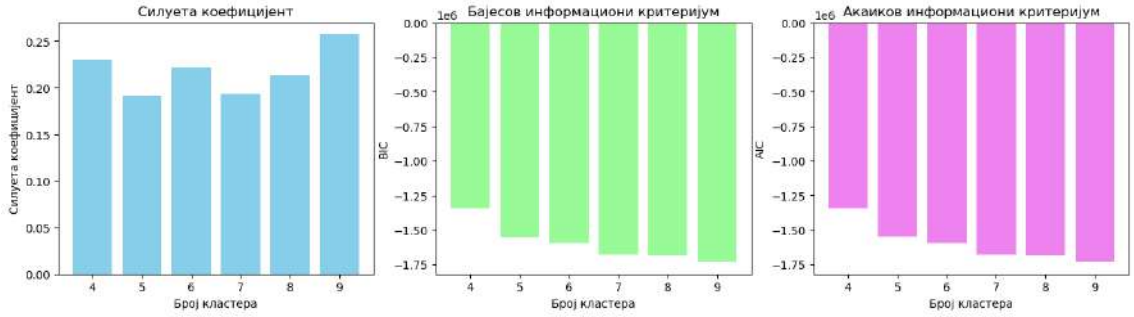
Пример 6.2.2. *Посматрајмо фотографију (слика 6.10) коју желимо да кластерујемо. На слици 6.11 налази се приказ почетне фотографије смештене у простор основних боја. Сви пиксели представљени су својом бојом, а положај пиксела у простору зависи од појединачних удела основних боја у његовој репрезентација. Скала за сваку боју иде од 0 до 255, па је тако црвена боја у овом простору задата координатама (255,0,0), док црна и бела боја имају координате (0,0,0) и (255,255,255). У складу са тим, примећујемо да тамније боје доминирају ближе координатном почетку, док све више бледе приближавајући се вишим вредностима за сваку од оса.*

Слика 6.15: Почетна слика на коју желимо да применимо процес кластеровања



Слика 6.16: Слика приказана у простору основних боја

Упоредимо сада резултате кластеровања добијене помоћу мешавине нормалних расподела и методе K средина. Сагледавањем резултата метода за одабир броја кластера, објашњених у претходном поглављу, који се могу видети на слици 6.12, закључујемо да, од понуђених бројева, девет кластера даје најбоље резултате. Такође, након извршавања кластеровања све тачке једног кластера представили смо бојом медијане додељеног кластера.

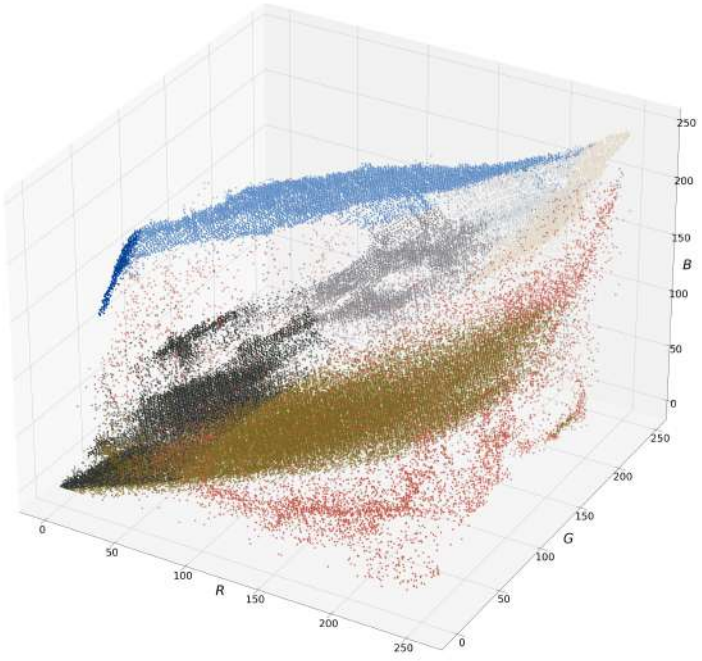


Слика 6.17: Резултати метода за одабир броја кластера

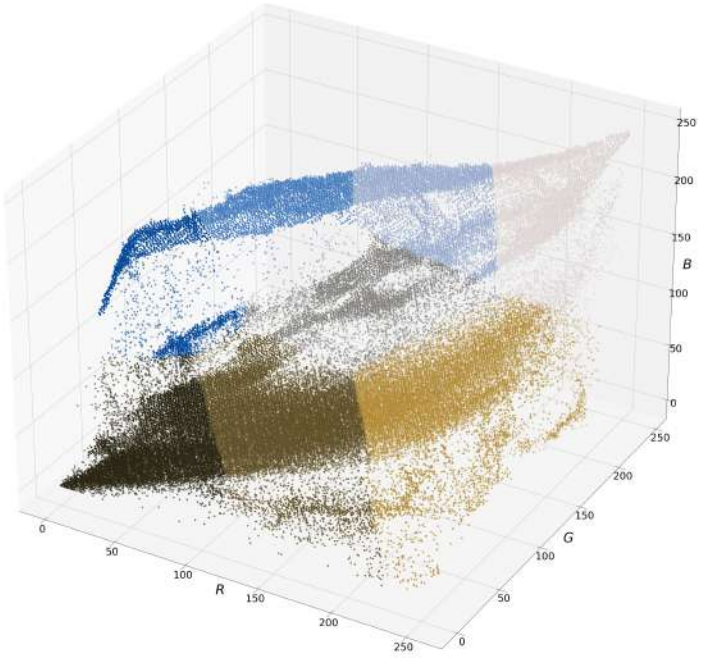
Приметимо да сликом доминирају плава, бела и пар нијанси зелене, док се на самим балонима налази неколико јаркијих боја од чега је црвена најдоминантнија. Са слике 6.11 можемо приметити да групације боја имају сложене облике, при чему постоје сегменти који су јасно одвојени, али такође присутне су и целине које се састоје од различитих боја.

Због претпоставке о сферном облику кластера, као и тежње да кластери буду сличних величина, метода K средина не успева да препозна балоне као посебан кластер. Овакав закључак можемо уочити и посматрањем слика 6.13 и 6.14, на којима се налази приказ кластера у простору за оба начина кластеровања. Јасно је да мешавине нормалних расподела превазилазе сложеније облике и формирају кластере у складу са њима, док су, са друге стране, за метод K средина границе између кластера видно оштрије и осликавајуману ове методе у раду са кластерима који нису сферног облика.

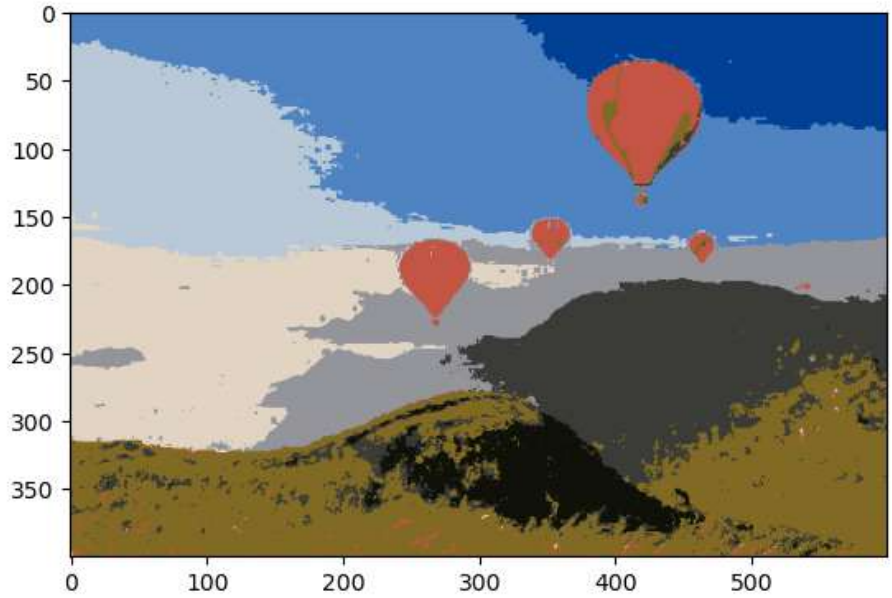
На сликама 6.15 и 6.16 налазе се крајње фотографије добијене након кластеровања мешавинама и методом K средина, редом.



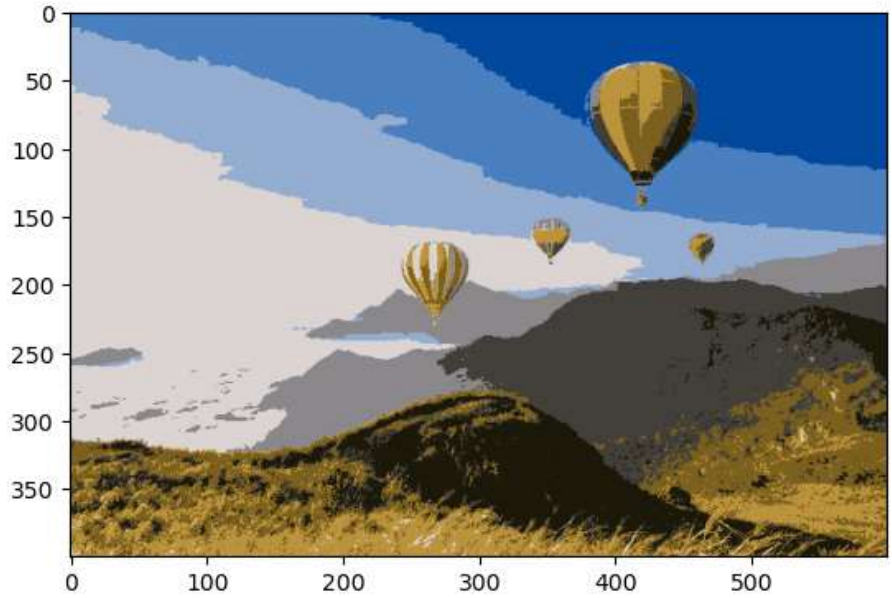
Слика 6.18: Слика приказана у простору основних боја након кластеровања методом мешавине нормалних расподела



Слика 6.19: Слика приказана у простору основних боја након кластеровања методом K средина



Слика 6.20: Слика након кластеровања методом мешавине нормалних расподела



Слика 6.21: Слика након кластеровања методом K средина

Глава 7

Закључак

У овом раду бавили смо се мешавинама вишедимензионих нормалних расподела и њиховом применом у кластеровању података. Значајан део рада усмерен је на EM алгоритам и његову теоријску позадину, без чијег разумевања се не може ући у суштину оваквог начина кластеровања. Упознавање са општим случајем EM алгоритма омогућило нам је да сагледамо његове примене и ван концепта мешавина вишедимензионих нормалних расподела.

Посебна пажња посвећена је алгоритму K средина као специјалном случају кластеровања мешавинама и једном од најпознатијих алгоритама за кластеровање. Циљ је био упоредити два различита приступа истом проблему и сагледати њихове предности и мане. Као главна предност мешавина нормалних расподела издвојила се чињеница да кластери не морају бити сферног облика. Сходно томе, приликом кластеровања мешавине превазилазе проблем комплексних података и успевају да конструишу кластере различитих облика и величина. Такође, ова особина пружа и могућност отпорности на одударајуће податке, будући да се облик кластера може прилагодити и удаљенијим тачкама без великог утицаја на центроиде и крајње резултате.

Иако мешавине нормалних расподела имају доста предности, у случајевима где се сусрећемо са подацима великих димензија кластеровање може трајати доста дуго. Насупрот томе, алгоритам K средина, због своје једноставности, успева да изврши исти задатак у доста краћем временском периоду.

Са циљем да се уверимо у велику моћ коришћења мешавина и истакнемо поменуте предности, изабрани су стварни примери кластеровања купаца у маркетинг индустрији као и примена кластеровања у сегментацији слика. Кроз ове примере илустрована је разноврсност задатка кластеровања, а током рада исти-

цане су и његове различите употребне вредности. Доказ да кластеровање може бити коришћено и у случајевима кад нам груписање података није крајњи циљ представља и чињеница да га можемо користити и за упознавање са подацима, као и да се може употребљавати као помоћ при означавању података након чега се датом задатку може приступити коришћењем неких од метода надгледаног учења.

Све у свему, кластеровање представља добар приступ решавању различитих проблема, а због својих бројних предности у које смо се уверили, мешавине нормалних расподела представљају добар избор за решавање задатка кластеровања и могу помоћи у добијању што бољих резултата.

Библиографија

[1] Arthur Dempster, Laird Nan, Rubin Donald, *Maximum likelihood from incomplete data via the EM algorithm*, J. R. Stat. Soc, 1977.

[2] Barun Kumar, *Customer Churn*, URL: `https://www.kaggle.com/datasets/barun2104/telecom-churn/data`, децембар 2023.

[3] Chloe Bi, *The EM Algorithm Explained*, URL: `https://medium.com/@chloebee/the-em-algorithm-explained-52182dbb19d9`, јануар 2024.

[4] Christopher Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[5] Cindy Rasmussen, *10 Stunning Yellow Colored Fish*, URL: `https://a-z-animals.com/blog/stunning-yellow-colored-fish-freshwater-and-saltwater/`, децембар 2024.

[6] Geoffrey McLachlan, Thriyambakam Krishnan, See Ket Ng, *The EM Algorithm*, Humboldt-Universität zu Berlin, Center for Applied Statistics and Economics (CASE), 2004.

[7] Geoffrey McLachlan, Sharon Lee, Suren Rathnayake, *Finite Mixture Models*, URL: `https://snunnari.github.io/SBE/mclachlan.pdf`, децембар 2024.

[8] Karl Pearson *Contributions to the mathematical theory of evolution*, Phil. Trans. R. Soc. Lond, 1894.

[9] Marco Taboga, *EM Algorithm*, URL: `https://www.statlect.com/fundamentals-of-statistics/EM-algorithm`, децембар 2023.

[10] Marco Taboga, *Multivariate normal distribution*, URL: `https://www.statlect.com/probability-distributions/multivariate-normal-distribution`, децембар 2023.

[11] Maya Gupta, Yihua Chen, *Theory and Use of the EM Algorithm*, Now Publishers Inc, 2011.

[12] Mladen Nikolić, Anđelka Zečević, *Mašinsko učenje*, URL: `https://ml.matf.bg.ac.rs/readings/ml.pdf`, децембар 2023.

[13] Sean Borman, *The Expectation Maximization Algorithm A short tutorial*, URL: `https://www.lri.fr/~sebag/COURS/EM_algorithm.pdf`, децембар 2023.

[14] Theodore Anderson, *An Introduction to Multivariate Statistical Analysis*, John Wiley Sons, Inc, 2003.

[15] Trevor Hastie, Robert Tibshirani, Jerome Friedman, *The Elements of Statistical Learning, Data Mining, Inference, and Prediction*, Second edition, Springer, 2009.

[16] *3D and contour plots of the bivariate normal distribution*, URL: `https://datasciencegenie.com/3d-contour-plots-of-bivariate-normal-distribution/`, децембар 2023.

[17] URL: `https://gr.pinterest.com/pin/692780355168236263/`

[18] URL:`https://www.istockphoto.com/photo/colourful-hot-air-balloons-flying-over-the-mountain-gm543839546-97676771`

Биографија аутора

Бојана Ђука рођена је 20. марта 1997. године у Београду. Завршила је Основну школу „Светозар Марковић” у Београду, а потом Прву београдску гимназију. Године 2016. уписује Математички факултет универзитета у Београду на смеру статистика, актуарска и финансијска математика. Основне студије завршава 2021. године, након чега исте године уписује мастер студије. Области интересовања су јој вероватноћа, статистика и машинско учење.