# Model Gaussove mješavine i algoritam maksimizacije očekivanja

**Vuknić, Andrea**





PRIRODOSLOVNO–MATEMATIČKI FAKULTET
SVEUČILIŠTA U SPLITU

ANDREA VUKNIĆ

# MODEL GAUSSOVE MJEŠAVINE I ALGORITAM MAKSIMIZACIJE OČEKIVANJA

DIPLOMSKI RAD

Split, rujan 2023.

PRIRODOSLOVNO–MATEMATIČKI FAKULTET
SVEUČILIŠTA U SPLITU

ODJEL ZA MATEMATIKU

# MODEL GAUSSOVE MJEŠAVINE I ALGORITAM MAKSIMIZACIJE OČEKIVANJA

DIPLOMSKI RAD

Student(ica):
Andrea Vuknić

Mentor(ica):
doc. dr. sc. Ivo Ugrina

Split, rujan 2023.

TEMELJNA DOKUMENTACIJSKA KARTICA

PRIRODOSLOVNO–MATEMATIČKI FAKULTET
SVEUČILIŠTA U SPLITU
ODJEL ZA MATEMATIKU

DIPLOMSKI RAD

# MODEL GAUSSOVE MJEŠAVINE I ALGORITAM MAKSIMIZACIJE OČEKIVANJA

Andrea Vuknić

**Sažetak**:

*Definiran je model konačne mješavine s naglaskom na model Gaussove mješavine te predstavljen problem procjene parametara takvih modela kao i rješavanje istog putem algoritma maksimizacije očekivanja. U tu svrhu, prvo su definirani potrebni pojmovi iz teorije vjerojatnosti. Na kraju je dan opis općeg algoritma maksimizacije očekivanja te analiza njegove monotonosti i konvergencije.*

**Ključne riječi**:

*normalna distribucija, funkcija izglednosti, procjenitelj najveće izglednosti, model konačne mješavine, latentne varijable, EM algoritam*

**Podatci o radu**:

*55 stranica, 2 slike, 19 literaturnih navoda, jezik izvornika: hrvatski*

**Mentor(ica)**: *doc. dr. sc. Ivo Ugrina*

**Članovi povjerenstva**:

*prof. dr. sc. Milica Klaričić Bakula*

*doc. dr. sc. Goran Erceg*

TEMELJNA DOKUMENTACIJSKA KARTICA

Povjerenstvo za diplomski rad je prihvatilo ovaj rad *29. rujna 2023.*

TEMELJNA DOKUMENTACIJSKA KARTICA

FACULTY OF SCIENCE, UNIVERSITY OF SPLIT
DEPARTMENT OF MATHEMATICS

MASTER'S THESIS

# GAUSSIAN MIXTURE MODEL AND EXPECTATION - MAXIMIZATION ALGORITHM

Andrea Vuknić

**Abstract**:

*The finite mixture model with a focus on the Gaussian mixture model is defined, and the problem of estimating the parameters of such models is presented, along with its solution through the Expectation-Maximization (EM) algorithm. To achieve this, the necessary concepts from probability theory are defined first. In the end, a description of the general EM algorithm is provided, along with an analysis of its monotonicity and convergence.*

**Key words**:

*normal distribution, likelihood function, maximum likelihood estimator, finite mixture model, latent variables, EM algorithm*

**Specifications**:

*55 pages, 2 figures, 19 references, written in Croatian*

**Mentor**: *doc. dr. sc. Ivo Ugrina*

**Committee**:

*prof. dr. sc. Milica Klaričić Bakula*

*doc. dr. sc. Goran Erceg*

This thesis was approved by a Thesis commettee on *September 29, 2023.*

# Uvod

Modeli konačne mješavine svoju vrijednost u statističkoj analizi pridaju svojstvu fleksibilnosti modeliranja podataka. Često se koriste prilikom opisivanja neke heterogene populacije unutar koje se nalazi više homogenih podpopulacija od kojih svaka prati određenu parametarsku distribuciju. U ovom radu će od interesa biti populacije čije podpopulacije prate normalnu ili Gaussovu razdiobu za koju se pokazalo da opisuje mnoge prirodne fenomene. Upravo je model mješavine Gaussovih distribucija prvi zabilježen model konačne mješavine koji je iskoristio statističar Karl Pearson 1894. godine za modeliranje dviju podpopulacija unutar populacije rakova u Napuljskom zaljevu. Danas se model Gaussove mješavine često koristi u raznim društvenim i prirodnim znanostima. Nadalje, predstavljen je problem procjene parametara modela Gaussove mješavine i algoritam koji nudi rješenje istog - algoritam maksimizacije očekivanja, formalno predstavljen 1977. U radu je, osim algoritma maksimizacije očekivanja primjenjenog na modelu Gaussove mješavine, dan i opći opis rada algoritma te njegova analiza.

# Sadržaj

TEMELJNA DOKUMENTACIJSKA KARTICA

# Poglavlje 1

# Teorija vjerojatnosti

Prvo definiramo osnovne pojmove iz teorije vjerojatnosti koji će nam poslužiti kao gradivni blokovi u daljnjoj konstrukciji teorije modela miješane gustoće te konačno modela Gaussove mješavine kao i algoritma maksimizacije očekivanja. Definicije u ovom dijelu rada preuzete su iz [18], [8] i [5].

Osnovni polazni objekt u teoriji vjerojatnosti jest neprazan skup $\Omega$ koji zovemo **prostor elementarnih događaja** i koji reprezentira skup svih ishoda slučajnog pokusa, odnosno pokusa čiji ishodi nisu jednoznačno određeni uvjetima u kojima se pokus izvodi. Točke $\omega \in \Omega$ nazivamo **elementarnim događajima**.

**Definicija 1.1** *Neka je $\Omega$ neprazan skup i $\mathcal{P}(\Omega)$ partitivni skup od $\Omega$. Skup $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ za kojeg vrijedi:*

*(A1)* $\emptyset \in \mathcal{A}$

*(A2)* *ako je $A \in \mathcal{A}$, onda je $A^c \in \mathcal{A}$*

*(A3)* *ako su $A_1, ..., A_n \in \mathcal{A}, n \in \mathbb{N}$, onda je $\bigcup_{i=1}^{n} A_i \in \mathcal{A}$*

*nazivamo* ***algebrom skupova*** *na $\Omega$.*

Dakle, algebra $\mathcal{A}$ je zatvorena na komplementiranje i konačne unije. Štoviše, algebra je zatvorena na konačne presjeke i skupovne razlike.

Naime, iz svojstava (A1) - (A3) slijedi:

$$\Omega = \emptyset^c \in \mathcal{A}$$
$$A_1, ..., A_n \in \mathcal{A} \Rightarrow \bigcap_{i=1}^n A_i = (\bigcup_{i=1}^n A_i^c)^c \in \mathcal{A}$$
$$A, B \in \mathcal{A} \Rightarrow A \setminus B = A \cap B^c \in \mathcal{A}$$

**Definicija 1.2** *Familiju $\mathcal{F}$ podskupova od $\Omega$ za koju vrijede svojstva:*

*(F1)* $\emptyset \in \mathcal{F}$

*(F2)* $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$

*(F3)* $A_i \in \mathcal{F}, i \in \mathbb{N} \Rightarrow \bigcup_i^\infty A_i \in \mathcal{F}$

*nazivamo* ***$\sigma$-algebra skupova*** *na $\Omega$, a uređeni par $(\Omega, \mathcal{F})$* ***izmjerivi prostor****.*

Slično se pokaže da za $\sigma$-algebru $\mathcal{F}$ vrijedi $\Omega \in \mathcal{F}$ te zatvorenost na prebrojive presjeke i skupovne razlike. Svaka $\sigma$-algebra je ujedno i algebra, a ukoliko je skup $\Omega$ konačan, pojmovi algebre i $\sigma$-algebre se podudaraju.

**Definicija 1.3** *Neka je $(\Omega, \mathcal{F})$ izmjerivi prostor. Funkciju $P\colon \mathcal{F} \to \mathbb{R}$ nazivamo* ***funkcijom vjerojatnosti*** *na $\mathcal{F}$ (ili $\Omega$) ako je*

*(P1)* $P(A) \geqslant 0,\ \forall A \in \mathcal{F}$ *(svojstvo nenegativnosti)*

*(P2)* $P(\Omega) = 1$ *(svojstvo normiranosti)*

*(P3) ako su $A_i \in \mathcal{F}, i \in \mathbb{N}$, te $A_i \cap A_j = \emptyset$ za sve $i \neq j$, onda je $P(\bigcup_{n=1}^\infty A_n) = \sum_{n=1}^\infty P(A_n)$ ($\sigma$-aditivnost)*

*Uređenu trojku $(\Omega, \mathcal{F}, P)$ nazivamo* ***vjerojatnosnim prostorom****.*

Neka je $(\Omega, \mathcal{F}, P)$ vjerojatnosni prostor. Tada elemente $\sigma$-algebre $\mathcal{F}$ nazivamo **događaji**, a broj $P(A), A \in \mathcal{F}$, **vjerojatnost događaja** $A$.

**Definicija 1.4** *Neka je $(\Omega, \mathcal{F}, P)$ proizvoljni vjerojatnosni prostor i $A \in \mathcal{F}$ takav da je $P(A) > 0$. Tada je funkcija $P_A \colon \mathcal{F} \to [0,1]$ definirana sa*

$$P_A(B) = P(B|A) = \frac{P(A \cap B)}{P(A)}$$

*vjerojatnost i nazivamo je* ***uvjetna vjerojatnost*** *uz uvjet $A$. Broj $P(B|A)$ nazivamo* ***vjerojatnost događaja $B$ uz uvjet da se dogodio događaj $A$****.*

Dakle, svakom događaju A za koji vrijedi $P(A) > 0$ pridružujemo vjerojatnosni prostor $(\Omega, \mathcal{F}, P_A)$.

**Definicija 1.5** *Neka je $(\Omega, \mathcal{F}, P)$ proizvoljni vjerojatnosni prostor i $A, B \in \mathcal{F}$. Kažemo da su događaji $A$ i $B$* ***nezavisni*** *ako je $P(A \cap B) = P(A) \cdot P(B)$.*

Definicija nezavisnosti se može proširiti i na familije događaja.

**Definicija 1.6** *Neka je $(\Omega, \mathcal{F}, P)$ vjerojatnosni prostor i $\mathcal{S} = \{A_i \in \mathcal{F} : i \in I\}$ proizvoljna familija događaja. Kažemo da je $\mathcal{S}$ famislija nezavisnih događaja ako za svaku njenu konačnu podfamiliju $\{A_{i_1}, ..., A_{i_n}\} \subseteq \mathcal{S}$ vrijedi*

$$P(A_{i_1} \cap \cdots \cap A_{i_n}) = P(A_{i_1}) \cdot ... \cdot P(A_{i_n}).$$

**Definicija 1.7** *Konačnu ili prebrojivu familiju $\{H_i \in \mathcal{F} : i \in I \subseteq \mathbb{N}\}$ neprazznih disjunktnih događaja u vjerojatnosnom prostoru $(\Omega, \mathcal{F}, P)$ za koje vrijedi $\bigcup_{i \in I} H_i = \Omega$ zovemo* ***potpuni sistem događaja*** *na $\Omega$.*

**Propozicija 1.8** *Neka je $\{H_i \in \mathcal{F} : i \in I \subseteq \mathbb{N}\}$ potpun sistem događaja u vjerojatnosnom prostoru $(\Omega, \mathcal{F}, P)$. Tada za svaki $A \in \mathcal{F}$ vrijedi*

1. ***formula totalne vjerojatnosti***

$$P(A) = \sum_{i \in I} P(H_i) \cdot P(A|H_i)$$

2. ***Bayesova formula**: ako je $P(A) \neq 0$, tada za svaki $i \in I$ vrijedi*

$$P(H_i|A) = \frac{P(H_i) \cdot P(A|H_i)}{P(A)}$$

**Dokaz.**

1. $P(A) = P(A \cap \Omega) = P(A \cap (\bigcup_{i \in I} H_i)) \overset{distrib.}{=} P(\bigcup_{i \in I}(A \cap H_i))$
   $\overset{\sigma-aditiv.}{=} \sum_{i \in I} P(A \cap H_i) = \sum_{i \in I} P(H_i) \cdot P(A|H_i)$
2. $P(H_i|A) = \frac{P(H_i \cap A)}{P(A)} = \frac{P(H_i) \cdot P(A|H_i)}{P(A)}$

■

# 1.1 Slučajne varijable

## 1.1.1 Diskretne slučajne varijable

U slučaju kada je $\Omega$ konačan ili prebrojiv skup, vjerojatnosni prostor $(\Omega, \mathcal{F}, P)$ nazivamo diskretnim vjerojatnosnim prostorom.

**Definicija 1.9** *Neka je $(\Omega, \mathcal{F}, P)$ diskretni vjerojatnosni prostor. Funkciju $X : \Omega \to \mathbb{R}$ nazivamo **diskretnom slučajnom varijablom**.*

Primijetimo da diskretna slučajna varijabla može poprimiti najviše prebrojivo mnogo vrijednosti jer je $X(\Omega)$ najviše prebrojiv skup (jer je $\Omega$ najviše prebrojiv). Vrijednosti slučajnih varijabli rezultati su određenih mjerenja čije su vrijednosti realni brojevi. Slika diskretne slučajne varijable $X(\Omega)$ predstavlja mogući skup vrijednosti za takva mjerenja.

Zanima nas vjerojatnost da slučajna varijabla X poprimi određenu vrijednost $a \in \mathbb{R}$, odnosno broj $P(X^{-1}\{a\})$, gdje je $X^{-1}\{a\} = \{\omega \in \Omega | X(\omega) = a\} \subseteq \Omega$. Jednako tako, možemo tražiti vjerojatnost da slučajna varijabla X

poprimi vrijednosti iz podskupa $E \subseteq \mathbb{R}$ te tada tražimo broj $P(X \in E) = P(X^{-1}(E))$, gdje je $X^{-1}(E) = \{\omega \in \Omega | X(\omega) \in E\} \subseteq \Omega$.

Ako je $X \colon \Omega \to \mathbb{R}$ slučajna varijabla u $\mathbb{R}$ i $a, b \in \mathbb{R}$, tada je

$$P(a \leqslant X \leqslant b) = P(X^{-1}([a,b])) = P(\{\omega \in \Omega | a \leqslant X(\omega) \leqslant b\})$$
$$P(X \leqslant a) = P(X^{-1}(\langle -\infty, a])) = P(\{\omega \in \Omega | X(\omega) \leqslant a\})$$
$$P(X \geqslant a) = P(X^{-1}([a, \infty\rangle)) = P(\{\omega \in \Omega | X(\omega) \geqslant a\})$$

Analogno se definira $P(X < a), P(X > a), P(a < X < b), P(a \leqslant X < b)$ te $P(a < X \leqslant b)$.

**Propozicija 1.10** *Neka je $X \colon \Omega \to \mathbb{R}$ diskretna slučajna varijabla. Neka je $\Omega' = X(\Omega)$ i $P' \colon \mathcal{P}(\Omega') \to [0,1]$ definirana sa $P'(E) = P(X \in E) = P(X^{-1}(E))$. Tada je $P'$ vjerojatnost i nazivamo je **distribucijom** ili **razdiobom** slučajne varijable $X$ te označavamo sa $P' = P_X$.*

Neka je $X \colon \Omega \to \mathbb{R}$ diskretna slučajna varijabla, $X(\Omega) = \{a_1, a_2, ...\}$. Distribucija slučajne varijable X određena je područjem vrijednosti koje slučajna varijabla X poprima, tj. elementima $a_1, a_2, ...$, i njima pripadnim vrijednostima $p_i = P(X = a_i)$, $i = 1, 2, ...$, što zapisujemo kao

$$X \sim \begin{pmatrix} \mathrm{a}_1 & \mathrm{a}_2 & \cdots & \mathrm{a}_n & \cdots \\ \mathrm{p}_1 & \mathrm{p}_2 & \cdots & \mathrm{p}_n & \cdots \end{pmatrix}$$

**Definicija 1.11** *Kažemo da su slučajne varijable $X_i \colon \Omega \to \mathbb{R}$, $i = 1, ..., n$, **nezavisne** ako su događaji $(X_1 \in E_1), ..., (X_n \in E_n)$ nezavisni, za svaki $E_i \subseteq \Omega$, $i = 1, ..., n$, tj. ako vrijedi*

$$P((X_1 \in E_1) \cap \cdots \cap (X_n \in E_n)) = P(X_1 \in E_1) \cdot ... \cdot P(X_n \in E_n).$$

**Definicija 1.12** *Neka je $X \colon \Omega \to \mathbb{R}$ diskretna slučajna varijabla zadana distribucijom*

$$X \sim \begin{pmatrix} a_1 & a_2 & \cdots & a_n & \cdots \\ p_1 & p_2 & \cdots & p_n & \cdots \end{pmatrix}.$$

***Funkcija gustoće vjerojatnosti slučajne varijable*** *$X$ (kraće,* ***gustoća*** *od $X$) je funkcija $f_X \colon \mathbb{R} \to \mathbb{R}$ definirana sa*

$$f_X(x) = P(X = x) = \begin{cases} 0, & \text{ako je } x \neq a_i \text{ za svaki } i \\ p_i, & \text{ako je } x = a_i \text{ za neki } i \end{cases}$$

Neka je $E \subseteq \mathbb{R}$. Tada je

$$P(X \in E) = P(X^{-1}(E)) = P(X^{-1}(E \cap \{a_1, a_2, ...\})) = \sum_{a_i \in E} P(X = a_i)$$
$$= \sum_{a_i \in E} p_i = \sum_{x \in E} f_X(x).$$

**Definicija 1.13** ***Funkcija distribucije*** *slučajne varijable $X \colon \Omega \to \mathbb{R}$ je funkcija $F_X \colon \mathbb{R} \to [0, 1]$ definirana sa*

$$F_X(x) = P(X \leqslant x) = P(\{\omega \in \Omega \colon X(\omega) \leqslant x\}).$$

Primijetimo da vrijedi

$$F_X(x) = \sum_{a_i \leqslant x} p_i = \sum_{y \leqslant x} f_X(y),\ x \in \mathbb{R}$$

pa je funkcija distribucije definirana za sve realne brojeve.

**Definicija 1.14** *Neka je $(\Omega, \mathcal{F}, P)$ diskretni vjerojatnosni prostor. Funkciju $X \colon \Omega \to \mathbb{R}^n$ nazivamo* ***n-dimenzionalnim diskretnim slučajnim vektorom****.*

Koordinate slučajnog vektora $X = (X_1, ..., X_n) \colon \Omega \to \mathbb{R}^n$ su slučajne varijable $X_i \colon \Omega \to \mathbb{R}$, $i = 1, ..., n$.

### 1.1.2 Neprekidne slučajne varijable

U diskretnom vjerojatnosnom prostoru slučajna varijabla može poprimiti najviše prebrojivo mnogo različitih vrijednosti. Međutim, ako se pretpostavi da slučajna varijabla može poprimiti sve realne vrijednosti iz nekog intervala, diskretni vjerojatnosni prostor u tom slučaju nije dostatan te je potrebno promatrati opći vjerojatnosni prostor.

**Definicija 1.15** *$\sigma$-algebru generiranu familijom svih otvorenih skupova na $\mathbb{R}$ nazivamo **$\sigma$-algebra Borelovih skupova** na $\mathbb{R}$ i označavamo s $\mathcal{B}$. Elemente $\sigma$-algebre $\mathcal{B}$ zovemo **Borelovi skupovi**.*

Slijedi da je svaki otvoreni interval $\langle a, b\rangle, a, b \in \mathbb{R}$ Borelov skup. Također, iz svojstava $\sigma$-algebre slijedi da je i svaki zatvoreni interval $[a, b], a, b \in \mathbb{R}$ Borelov skup kao komplement otvorenog skupa. Nadalje, iz

$$\langle a, b] = \bigcap_{n=1}^{\infty} (a, b + \frac{1}{n})$$
$$[a, b\rangle = \bigcap_{n=1}^{\infty} (a - \frac{1}{n}, b)$$

za $a < b, a, b \in \mathbb{R}$ slijedi da su intervali $\langle a, b]$ i $[a, b\rangle$ Borelovi skupovi.

Neka je $(\Omega, \mathcal{F}, P)$ vjerojatnosni prostor i X realna funkcija na $\Omega$. Svakom događaju $\omega \in \Omega$ je pridužen realan broj $X(\omega)$. Kada tražimo vjerojatnost od $a < X(\omega) < b, a, b \in \mathbb{R}, a < b$, računamo $P(X \in \langle a, b\rangle) = P(X^{-1}\langle a, b\rangle)$. Kako bi prethodno mogli izračunati, mora vrijediti $X^{-1}\langle a, b\rangle \in \mathcal{F}$.

**Definicija 1.16** *Neka je $(\Omega, \mathcal{F}, P)$ vjerojatnosni prostor. Funkciju $X : \Omega \to \mathbb{R}$ nazivamo **slučajna varijabla** na $\Omega$ ako je $X^{-1}(B) \in \mathcal{F}$, za proizvoljni $B \in \mathcal{B}$, tj. $X^{-1}(\mathcal{B}) \subseteq \mathcal{F}$.*

**Definicija 1.17** *$\sigma$-algebru generiranu familijom svih otvorenih podskupova od $\mathbb{R}^n$ nazivamo* ***$\sigma$-algebra Borelovih skupova na*** *$\mathbb{R}^n$ i označavamo s $\mathcal{B}^n$. Elemente od $\mathcal{B}^n$ zovemo* ***Borelovi skupovi na*** *$\mathbb{R}^n$.*

Neka je $(\Omega, \mathcal{F}, P)$ vjerojatnosni prostor. Pretpostavimo da izvodimo slučajni pokus u kojem mjerimo dvije veličine, $X_1$ i $X_2$, što znači da svakom događaju $\omega \in \Omega$ pridružujemo točku $X(\omega) = (X_1(\omega), X_2(\omega)) \in \mathbb{R}^2$. Najčešće tražimo vjerojatnost za $a < X_1(\omega) < b, c < X_2(\omega) < d$, odnosno vjerojatnost da $X$ upadne u pravokutnik $(a, b) \times (c, d)$ u $\mathbb{R}^2$. Dakle, želimo izračunati vjerojatnost $P(\{\omega \in \Omega | X(\omega) \in (a, b) \times (c, d)\}) = P(X^{-1}((a, b) \times (c, d))) = P(a < X_1 < b, c < X_2 < d)$. Da bi to bilo moguće, mora vrijediti $X^{-1}(\mathcal{B}) \in \mathcal{F}$ za svaki pravokutnik $\mathcal{B} = (a, b) \times (c, d)$ u $\mathbb{R}^2$.

**Definicija 1.18** *Neka je $(\Omega, \mathcal{F}, P)$ vjerojatnosni prostor i $X \colon \Omega \to \mathbb{R}^n$. Kažemo da je $X$* ***n-dimenzionalni slučajni vektor*** *(ili kraće* ***slučajni vektor****) na $\Omega$ ako je $X^{-1}(B) \in \mathcal{F}$ za svaki $B \in \mathcal{B}^n$, tj. $X^{-1}(\mathcal{B}^n) \subseteq \mathcal{F}$.*

Neka je $(\Omega, \mathcal{F}, P)$ vjerojatnosni prostor i $X$ slučajna varijabla na $\Omega$.
Za $B \in \mathcal{B}$ definiramo funkciju $P_X \colon \mathcal{B} \to [0, 1]$ sa

$$P_X(B) = P(X^{-1}(B)) = P(\{\omega \in \Omega : X(\omega) \in B\}) = P(X \in B).$$

Lako se provjeri da je $P_X$ vjerojatnost. Vjerojatnosni prostor $(\mathbb{R}, \mathcal{B}, P_X)$ zovemo **vjerojatnosni prostor induciran sa** $X$. Na ovaj način, svakoj slučajnoj varijabli $X$ se na prirodan način pridružuje vjerojatnosni prostor $(\mathbb{R}, \mathcal{B}, P_X)$. $P_X$ često zovemo i **zakon razdiobe** od $X$. Vrijedi

$$\begin{aligned} P_X(\langle -\infty, x]) &= P(X^{-1}\langle -\infty, x]) = P(\{\omega \in \Omega \colon X(\omega) \leqslant x\}) \\ &= P(X \leqslant x) = F_X(x), \quad x \in \mathbb{R}, \end{aligned}$$

gdje je $F_X$ funkcija distribucije slučajne varijable X.

**Definicija 1.19** *Neka je $X$ slučajna varijabla na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, P)$ i neka je $F_X$ njena funkcija distribucije. Kažemo da je $X$* ***neprekidna slučajna varijabla*** *ako postoji funkcija $f_X \colon \mathbb{R} \to [0, +\infty\rangle$ takva da je $F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt$.*
*Za funkciju distribucije $F_X$ neprekidne slučajne varijable $X$ kažemo da je* ***apsolutno neprekidna funkcija distribucije****, a nenegativnu realnu funkciju $f_X$ nazivamo* ***funkcija gustoće vjerojatnosti od*** *$X$ ili jednostavno* ***gustoćom od*** *$X$.*

**Propozicija 1.20** *Neka je $f \colon \mathbb{R} \to \mathbb{R}$ neprekidna funkcija. Da bi $f$ bila gustoća vjerojatnosti neke neprekidne slučajne varijable $X$, nužno je i dovoljno da vrijedi*

1. $f(x) \geqslant 0, \quad x \in \mathbb{R}$
2. $\int_{-\infty}^{+\infty} f(x)\, dx = 1.$

Za sve $a, b \in \mathbb{R}, a < b$, vrijedi

$$P(a < X \leqslant b) = P((X \leqslant b) \backslash (X \leqslant a)) = P(X \leqslant b) - P(X \leqslant a)$$
$$= F_X(b) - F_X(a) = \int_{-\infty}^{b} f(t)\, dt - \int_{-\infty}^{a} f(t)\, dt = \int_{a}^{b} f(t)\, dt.$$

Također, kod neprekidnih slučajnih varijabli imamo $P(X = a) = 0$, $\forall a \in \mathbb{R}$, jer $P(X = a) = \int_a^a f(t)\, dt = 0$ pa slijedi da je $P(a \leqslant X \leqslant b) = P(a \leqslant X < b) = P(a < X \leqslant b) = P(a < X < b)$, $\forall a, b \in \mathbb{R}, a < b$.

Neka je $(\Omega, \mathcal{F}, P)$ vjerojatnosni prostor i $X \colon \Omega \to \mathbb{R}^n$ slučajni vektor, tj. $X = (X_1, ..., X_n)$, gdje su $X_i \colon \Omega \to \mathbb{R}, i = 1, ..., n$, slučajne varijable. Za $B \in \mathcal{B}^n$ definiramo funkciju $P_X \colon \mathcal{B}^n \to [0, 1]$ sa $P_X(B) = P(X^{-1}(B))$.
$P_X$ je vjerojatnost na $\mathcal{B}^n$ i zovemo je **zakon razdiobe slučajnog vektora** $X$. Dakle, svakom n-dimenzionalnom slučajnom vektoru $X$ se na prirodan

način pridružuje vjerojatnosni prostor $(\mathbb{R}^n, \mathcal{B}^n, P_X)$ koji zovemo **vjerojatnosni prostor induciran slučajnim vektorom** $X$.

**Definicija 1.21** *Neka je $X\colon \Omega \to \mathbb{R}^n$ n-dimenzionalni slučajni vektor, $X = (X_1, ..., X_n)$.* ***Funkcija distribucije slučajnog vektora*** *$X$ je funkcija $F_X = F\colon \mathbb{R}^n \to [0,1]$ definirana sa*

$$\begin{aligned} F(x) = F(x_1, ..., x_n) &= P_x(\langle -\infty, x]) = P(X \leqslant x) \\ &= P(X_1 \leqslant x_1, ..., X_n \leqslant x_n), \quad x = (x_1, ..., x_n) \in \mathbb{R}^n. \end{aligned}$$

**Definicija 1.22** *Neka su $n, m \in \mathbb{N}$ te $\mathcal{B}^n, \mathcal{B}^m$ $\sigma$-algebre Borelovih skupova na $\mathbb{R}^n$ i $\mathbb{R}^m$ respektivno. Za funkciju $g\colon \mathbb{R}^n \to \mathbb{R}^m$ kažemo da je* ***Borelova funkcija*** *ako je $g^{-1}(B) \in \mathcal{B}^n$ za svaki $B \in \mathcal{B}^m$, tj. ako je $g^{-1}(\mathcal{B}^m) \subseteq \mathcal{B}^n$.*

Svaka neprekidna funkcija $g\colon \mathbb{R}^n \to \mathbb{R}^m$ je Borelova funkcija.

**Definicija 1.23** *Neka je $X = (X_1, ..., X_n)$ n-dimenzionalni slučajni vektor na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, P)$ i $F$ njegova funkcija distribucije. $X$ je* ***neprekidan slučajni vektor*** *ako postoji nenegativna Borelova funkcija $f\colon \mathbb{R}^n \to \mathbb{R}$ takva da je*

$$F(x) = \int_{\langle -\infty, x]} f(t)\, dt = \int_{-\infty}^{x_1} ... \int_{-\infty}^{x_n} f(t_1, ..., t_n)\, dt_1 ... dt_n,$$
$$x = (x_1, ..., x_n) \in \mathbb{R}^n.$$

*Funkciju $f$ nazivamo* ***funkcijom gustoće slučajnog vektora*** *$X$.*

### 1.1.3 Karakteristične vrijednosti slučajnih varijabli

**Definicija 1.24** ***Matematičko očekivanje*** *slučajne varijable $X\colon \Omega \to \mathbb{R}$ definiramo kao broj*

1. $\mathbb{E}X = \sum_{\omega\in\Omega} X(\omega)P(\{\omega\})$
   *u slučaju kada je X diskretna slučajna varijabla i ako taj red apsolutno konvergira,*

2. $\mathbb{E}X = \int_{-\infty}^{+\infty} x f_X(x)\,dx$
   *u slučaju kada je X neprekidna slučajna varijabla s gustoćom $f_X$ i ako taj integral apsolutno konvergira.*

Matematičko očekivanje slučajne varijable X interpretiramo kao srednju vrijednost od X. Ako je $g\colon \mathbb{R} \to \mathbb{R}$ Borelova funkcija i $X$ neprekidna slučajna varijabla s gustoćom $f_X$, onda je funkcija $g(X)\colon \Omega \to \mathbb{R}$ također neprekidna slučajna varijabla i vrijedi

$$\mathbb{E}(g(X)) = \int_{-\infty}^{+\infty} g(x) f_X(x)\,dx.$$

**Definicija 1.25** *Neka je $X\colon \Omega \to \mathbb{R}$ slučajna varijabla. Realni broj $VarX = \mathbb{E}(X - \mathbb{E}X)^2$ nazivamo **varijancom** slučajne varijable X.*

Varijanca je, dakle, srednje kvadratno odstupanje varijable X od njenog očekivanja $\mathbb{E}X$ i kao takva predstavlja mjeru raspršenja vrijednosti slučajne varijable od njenog očekivanja. Vrijedi

$$\begin{aligned} VarX &= \mathbb{E}(X - \mathbb{E}X)^2 \\ &= \mathbb{E}(X^2 - 2X\mathbb{E}X + (\mathbb{E}X)^2) \\ &= \mathbb{E}X^2 - 2\mathbb{E}X \cdot \mathbb{E}X + (\mathbb{E}X)^2 \\ &= \mathbb{E}X^2 - 2(\mathbb{E}X)^2 + (\mathbb{E}X)^2 \\ &= \mathbb{E}X^2 - (\mathbb{E}X)^2. \end{aligned}$$

**Definicija 1.26** *Broj $\sigma_X = \sqrt{VarX}$ nazivamo **standardna devijacija** od X gdje je X slučajna varijabla.*

**Definicija 1.27** ***Kovarijanca*** *opisuje odnos između dviju slučajnih varijabli i za slučajne varijable X i Y je definirana kao*

$$Cov(X,Y) = \sigma_{X,Y} = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)].$$

Kovarijancom mjerimo koliko je promjena jedne varijable povezana s promjenom druge varijable, odnosno mjeri stupanj linearne povezanosti dviju varijabli. Primijetimo da je $Cov(X,Y) = Cov(Y,X)$ i $Cov(X,X) = VarX$. Ako su X i Y nezavisne slučajne varijable, onda je nužno $Cov(X,Y) = 0$.

**Definicija 1.28** *Neka je $X = (X_1, ..., X_n)$ n-dimenzionalni slučajni vektor na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, P)$.* ***Matematičko očekivanje od*** *X je vektor*

$$\mathbb{E}X = (\mathbb{E}X_1, ..., \mathbb{E}X_n) \in \mathbb{R}^n$$

*uz pretpostavku da je $\mathbb{E}X_i \in \mathbb{R}$ za svaki $i = 1, ..., n$. Ako $\mathbb{E}X$ postoji kažemo da je X integrabilni slučajni vektor.*

Za slučajni vektor $X = (X_1, ..., X_n)$ može nas zanimati i međusobni odnos njegovih koordinata, odnosno slučajnih varijabli $X_i, i = 1, ..., n$.
**Kovarijacijska matrica** $\Sigma$ definirana je kao

$$\Sigma = \mathbb{E}[(X - \mathbb{E}X)(X - \mathbb{E}X)^T].$$

To je kvadratna matrica koja na glavnoj dijagonali ima varijance varijabli (jer je $Cov(X,X) = VarX$), a izvan dijagonale kovarijance svih parova varijabli. Štoviše, ona je i simetrična jer vrijedi

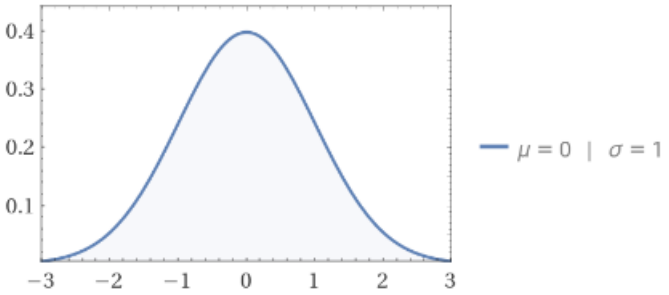
$$\Sigma_{ij} = Cov(X_i, X_j) = Cov(X_j, X_i) = \Sigma_{ji}.$$

## 1.2 Normalna (Gaussova) razdioba

**Definicija 1.29** *Kažemo da neprekidna slučajna varijabla $X$ ima normalnu (Gaussovu) razdiobu s parametrima $\mu \in \mathbb{R}, \sigma^2 > 0$ i pišemo $X \sim \mathcal{N}(\mu, \sigma^2)$ ako joj je funkcija gustoće dana sa*

$$\mathcal{N}(x|\mu, \sigma^2) \equiv f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}.$$

Graf ove funkcije gustoće je zvonolika krivulja koju nazivamo **Gaussova krivulja**.




Slika 1.1: Gaussova krivulja

Vidimo kako funkcija gustoće normalne razdiobe zadovoljava svojstva funkcije gustoće:

$$\mathcal{N}(x|\mu, \sigma^2) > 0$$
$$\int_{-\infty}^{+\infty} \mathcal{N}(x|\mu, \sigma^2)\, dx = 1.$$

Također, vrijedi:

$$\mathbb{E}X = \int_{-\infty}^{+\infty} \mathcal{N}(x|\mu, \sigma^2) x\, dx = \mu$$
$$\mathbb{E}(X^2) = \int_{-\infty}^{+\infty} \mathcal{N}(x|\mu, \sigma^2) x^2\, dx = \mu^2 + \sigma^2$$
$$VarX = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \sigma^2$$

Parametar $\mu$ nazivamo srednja vrijednost, a $\sigma^2$ varijancom normalne razdiobe. Kada stavimo $\mu = 0$ i $\sigma = 1$ dobijemo slučajnu varijablu $\mathcal{N}(0,1)$ koju nazivamo **jedinična normalna razdioba**. U [9] se može vidjeti kako se jedinična i opća normalna razdioba mogu dobiti jedna iz druge linearnom transformacijom:

$$X \sim \mathcal{N}(0,1) \implies \mu + \sigma X \sim \mathcal{N}(\mu, \sigma^2)$$
$$X \sim \mathcal{N}(\mu, \sigma^2) \implies \tfrac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$$

Od interesa nam je često promatrati normalnu razdiobu u višedimenzionalnom prostoru.

**Definicija 1.30** *Neka je $n \in \mathbb{N}$ i $X$ n-dimenzionalni neprekidni slučajni vektor. Kažemo da $X$ ima normalnu (Gaussovu) razdiobu ako mu je funkcija gustoće dana sa*

$$\mathcal{N}(x|\mu, \Sigma) = \tfrac{1}{(2\pi)^{n/2}} \tfrac{1}{|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)},$$

*gdje je $\mu$ n-dimenzionalni vektor kojeg zovemo srednja vrijednost te $\Sigma$ $n \times n$ matrica koju zovemo kovarijacijska matrica, a $|\Sigma|$ determinanta kovarijacijske matrice.*

Kako bi Gaussova razdioba bila dobro definirana, vidimo da nam je potreban inverz $\Sigma^{-1}$ te $|\Sigma|$ determinanta kovarijacijske matrice.

**Napomena 1.31** *Za matricu $A \in M_n(\mathbb{R})$ kažemo da je **pozitivno definitna** ako je $x^T Ax > 0$, za svaki $x \in \mathbb{R}^n, x \neq 0$.*
*Ako vrijedi $x^T Ax \geqslant 0$ kažemo da je $A$ **pozitivno semidefinitna**.*
*Svojstvene vrijednosti pozitivno definitne matrice su strogo pozitivne te svaka pozitivno definitna matrica ima inverz.*
*Svojstvene vrijednosti pozitivne semidefinitne matrice su nenegativne i pozitivno semidefinitna matrica ne mora nužno imati inverz.*

### 1.2. Normalna (Gaussova) razdioba

Neka je $v \in \mathbb{R}^n$, $v \neq 0$, proizvoljan. Tada je

$$\begin{aligned} v^T \Sigma v &= v^T \mathbb{E}[(X - \mathbb{E}X)(X - \mathbb{E}X)^T] v \\ &= \mathbb{E}[v^T (X - \mathbb{E}X)(X - \mathbb{E}X)^T v] \end{aligned}$$

Stavimo $\mu = \mathbb{E}X$ te $Y = v^T(X - \mu)$. Tada slijedi

$$v^T \Sigma v = \mathbb{E}[v^T (X - \mu)(v^T (X - \mu))^T] = \mathbb{E}[(v^T (X - \mu))^2] = \mathbb{E}Y^2 \geqslant 0.$$

Vidimo da je kovarijacijska matrica $\Sigma$ pozitivno semidefinitna što znači da ne mora nužno imati inverz. $\Sigma$ će imati inverz i tada determinantu različitu od nule ako je pozitivno definitna. Dakle, ako je neki od elemenata na dijagonali jednak nuli ili ako postoji linearna zavisnost između njenih redaka, odnosno stupaca, kovarijacijska matrica neće imati inverz. Međutim, znamo da se na dijagonali od $\Sigma$ nalaze varijance za svako obilježje opažanja. Kada bi neka od tih varijanci bila jednaka nuli, to bi značilo da je pripadno obilježje konstantno. Takvo obilježje u opažanjima tada nije od koristi zbog toga što je očito jednako za svako opažanje. Linearnu zavisnost redaka, odnosno stupaca, kovarijacijske matrice $\Sigma$ imati ćemo ako postoji savršena multikolinearnost obilježja što bi značilo da se neko obilježje može predvidjeti iz ostalih.

Primijetimo da je ovisnost funkcije $\mathcal{N}(x|\mu, \Sigma)$ o $x$ prikazana u kvadratnoj formi

$$\Delta^2 = (x - \mu)^T \Sigma^{-1} (x - \mu)$$

koju pronalazimo u eksponentu. Korijen od $\Delta^2$ nazivamo **Mahalanobisova udaljenost** između $x$ i $\mu$ te se ona reducira na euklidsku udaljenost u slučaju kada je $\Sigma$ jedinična matrica ([4]).

**Napomena 1.32** *Spomenimo još i važnost normalne distribucije kroz prizmu centralnog graničnog teorema. Centralni granični teorem govori o tome da se zbroj slučajnih varijabli, uz neke uvjete na njihove distribucije, asimptotski ponaša kao normalna (Gaussova) razdioba ([9]). Naime, ako je broj uzoraka iz neke populacije dovoljno velik, distribucija srednjih vrijednosti tih uzoraka se može aproksimirati normalnom distribucijom čak i kada sama populacija nije normalno distribuirana. Više o centralnim graničnim teoremima može se pronaći u [18] i [9].*

***Teorem 1.33 (Levy)*** *Neka je* $(X_n, n \in \mathbb{N})$ *niz nezavisnih, jednako distribuiranih slučajnih varijabli s očekivanjem* $m$ *i varijancom* $\sigma^2$, $0 < \sigma^2 < \infty$, *i neka je* $S_n = \sum_{k=1}^{n} X_k, n \in \mathbb{N}$. *Tada vrijedi*

$$\frac{S_n - \mathbb{E}S_n}{\sigma\sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1) \text{ za } n \to \infty.$$

*Odnosno, iz teorema slijedi*

$$\frac{\sum_{k=1}^{n}(X_k - m)}{\sigma\sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1) \text{ za } n \to \infty$$

*gdje* $\xrightarrow{\mathcal{D}}$ *označava konvergenciju po distribuciji.*

## 1.3 Procjena parametara

U praksi, najčešće na raspolaganju imamo nekakav uzorak iz određene populacije, tj. na uvidu nam je konačan podskup populacije. Cilj nam je na temelju tog uzorka opisati cijelu populaciju. Dakle, pitamo se kako na temelju danog uzorka za kojeg se pretpostavlja da dolazi iz neke distribucije s parametrom $\theta$, procijeniti upravo taj parametar $\theta$ i tako dati opis cijele populacije kojoj pripada dani uzorak.

### 1.3. Procjena parametara

Pretpostavimo da imamo skup podataka $X = (x_1, ..., x_N), N \in \mathbb{N}$, gdje je $x_i = (x_{i_1}, ..., x_{i_D})$ za neki $D \in \mathbb{N}$. Odnosno, dostupan nam je skup podataka od $N$ opažanja gdje svako opažanje ima $D$ obilježja. Za opažanja koja su nezavisna (jedno opažanje ne utječe na drugo) i dolaze iz iste distribucije kažemo da su **nezavisna i jednako distribuirana** za što se u literaturi često koristi oznaka **i.i.d.**[1].

**Definicija 1.34** *Neka je $X$ slučajna varijalba s razdiobom $F_X$. Za slučajne varijable $X_1, ..., X_n$ kažemo da su nezavisne kopije slučajne varijable $X$ ako su međusobno nezavisne i imaju razdiobu identičnu razdiobi slučajne varijable $X$. Tada $n$-torku slučajnih varijabli $(X_1, ..., X_n)$ nazivamo **uzorak**.*
*Ako je $x_i$ realizacija varijable $X_i, i = 1, ..., n$, tada se $(x_1, ..., x_n)$ naziva **vrijednost** ili **realizacija** uzorka $(X_1, ..., X_n)$. Broj $n$ označava **dimenziju** uzorka.*

Pretpostavimo da je u razdiobi slučajne varijable X nepoznat jedan parametar $\theta$. Vrijednost parametra $\theta$ trebamo procijeniti na temelju realizacija $(x_1, ..., x_n)$ uzorka $(X_1, ..., X_n)$. Ta će procjena, u oznaci $\hat{\theta}$, naravno, ovisiti o realizacijama $x_1, ..., x_n$ pa je definiramo kao funkciju tih realizacija

$$\hat{\theta} := g(x_1, ..., x_n).$$

Kako su $x_1, ..., x_n$ realizacije slučajnih varijabli $X_1, ..., X_n$, tako je $\hat{\theta}$ realizacija slučajne varijable

$$\Theta := g(X_1, ..., X_n),$$

tj. slučajne varijable čija je vrijednost izračunata na temelju slučajnog uzorka ($g$ je funkcija koja izračunava nešto na temelju uzorka). Takvu slučajnu varijablu nazivamo **statistika**.

[1] engl. independent and identically distributed

**Definicija 1.35** *Neka je $\theta$ nepoznati parametar u populaciji X. Za statistiku $\Theta$ kažemo da je* ***procjenitelj*** *parametra $\theta$, a vrijednost te statistike $\hat{\theta} := g(x_1, ..., x_n)$ nazivamo* ***procjenom*** *parametra $\theta$.*

Budući da je procjenitelj slučajna varijabla, kao i svaka slučajna varijabla ima svoje očekivanje.

**Definicija 1.36** *Za statistiku $\Theta$ kažemo da je* ***nepristrani procjenitelj*** *parametra $\theta$ ako vrijedi $\mathbb{E}(\Theta) = \theta$.*

### 1.3.1 Procjenitelj najveće izglednosti

**Napomena 1.37** *Do sada smo funkcije gustoće označavali sa $f$, no nadalje će za funkciju gustoće vrijediti oznaka $p$ kako bi pratili oznake u literaturi.*

Pretpostavimo da imamo na raspolaganju uzorak od N opažanja, odnosno skup $\{x_1, ..., x_N\}$, $N \in \mathbb{N}$, za koje pretpostavljamo da dolaze iz populacije opisane distribucijom $p(x; \theta)$ koja ovisi o nekom parametru ili skupu parametara $\theta$. Tada je

$$p(x_1, ..., x_N; \theta) \overset{i.i.d.}{=} \prod_{i=1}^{N} p(x_i; \theta).$$

Uočimo kako je ovdje jedino varijabilan parametar $\theta$ pa gornji izraz možemo shvatiti kao funkciju od $\theta$.

**Definicija 1.38** *Neka je $x_1, ..., x_n$ realizacija uzorka iz populacije X čija funkcija gustoće $p(x; \theta)$ ovisi o nepoznatom parametru $\theta$.* ***Funkcija izglednosti***[2] *definira se kao umnožak*

$$\mathcal{L}(\theta) \equiv \mathcal{L}(\theta | x_1, ..., x_n) := p(x_1; \theta) \cdot ... \cdot p(x_n; \theta).$$

[2]engl. likelikood function

### 1.3. Procjena parametara

Funkcija izglednosti nam za zadani parametar kaže koliko je vjerojatan uzorak koji imamo. Jedan od često korištenih kriterija za određivanje parametara vjerojatnosne distribucije putem danog uzorka jest **kriterij najveće izglednosti** gdje tražimo vrijednost parametara distribucije koji maksimiziraju funkciju izglednosti. Intuitivno možemo pretpostaviti da je naš uzorak vjerojatniji od ostalih iz razloga što imamo upravo njega, a ostali se nisu realizirali. Dakle, za procjenu parametra $\theta$ uzimamo onu vrijednost za koju funkcija izglednosti poprima globalni maksimum,

$$\theta_{MLE} = \underset{\theta}{argmax}\, \mathcal{L}(\theta|x_1, ..., x_n).$$

Takvu procjenu nazivamo **procjenitelj najveće izglednosti**[3].
Primijetimo da je funkcija izglednosti $\mathcal{L}(\theta)$ uvijek pozitivna pa je dobro definiran i logaritam izglednosti, tzv. funkcija log-izglednosti

$$\ln \mathcal{L}(\theta) \equiv \ln \mathcal{L}(\theta|x_1, ..., x_n) = \ln(\prod_{i=1}^{n} p(x_i; \theta)) = \sum_{i=1}^{n} \ln p(x_i; \theta).$$

Kako je ln rastuća funkcija, log-izglednost poprima maksimum u istim točkama kao i funkcija izglednosti te je zbog jednostavnosti često korištena i u tom slučaju tražimo

$$\theta_{MLE} = \underset{\theta}{argmax}\, \ln \mathcal{L}(\theta|x_1, ..., x_n).$$

U slučaju višeparametarske distribucije, funkcija izglednosti ima oblik

$$\mathcal{L}(\theta_1, ..., \theta_m) \equiv \mathcal{L}(\theta_1, ..., \theta_m|x_1, ..., x_n) = p(x_1; \theta_1, ..., \theta_m) \cdot ... \cdot p(x_n; \theta_1, ..., \theta_m)$$

i tada nepoznate parametre $\theta_1, ..., \theta_m$ dobivamo iz uvjeta

$$\frac{\partial \mathcal{L}(\theta_1, ..., \theta_m)}{\partial \theta_i} = 0, \quad i = 1, ..., m.$$

[3]engl. maximum likelihood estimator (MLE)

### 1.3.2 Procjena parametara normalne razdiobe

Neka je $(x_1, .., x_N)$, $N \in \mathbb{N}$, realizacija nekog uzorka za kojeg pretpostavljamo da dolazi iz normalne razdiobe $\mathcal{N}(\mu, \sigma^2)$. Pripadna funkcija log-izglednosti je

$$\begin{aligned}
\ln \mathcal{L}(\mu, \sigma^2 | x_1, ..., x_N) &= \ln \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}} \\
&= \sum_{i=1}^{N} \ln(\frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}) \\
&= \sum_{i=1}^{N} \ln \frac{1}{\sigma\sqrt{2\pi}} - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2} \\
&= \sum_{i=1}^{N} (\ln 1 - \ln(\sigma\sqrt{2\pi})) - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2} \\
&= -\sum_{i=1}^{N} \ln(\sigma\sqrt{2\pi}) - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2} \\
&= -\sum_{i=1}^{N} (\ln \sigma + \frac{1}{2}\ln(2\pi)) - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2} \\
&= -N \cdot \ln \sigma - \frac{N}{2}\ln(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2.
\end{aligned}$$

Sada maksimiziramo log-izglednost:

$$\nabla \ln \mathcal{L}(\mu, \sigma^2) = 0$$

$$\frac{\partial}{\partial \mu} \ln \mathcal{L}(\mu, \sigma^2) = \frac{1}{2\sigma^2} \cdot 2 \sum_{i=1}^{N}(x_i - \mu) = 0 \Rightarrow \mu_{MLE} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\frac{\partial}{\partial \sigma} \ln \mathcal{L}(\mu, \sigma^2) = \frac{-N}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{N}(x_i-\mu)^2 = 0 \Rightarrow \sigma^2_{MLE} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_{MLE})^2$$

Dakle, kriterijem najveće izglednosti za parametre normalne distribucije dobijemo $\mu_{MLE} = \frac{1}{N}\sum_{i=1}^{N} x_i$ što je srednja vrijednost uzorka te $\sigma^2_{MLE} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_{MLE})^2$ što je varijanca uzorka, tj. srednje kvadratno odstupanje od sredine uzorka. Primijetimo kako su $\mu_{MLE}$ i $\sigma^2_{MLE}$ funkcije od $x_1, ..., x_N$ te se može pokazati da vrijedi

$$\mathbb{E}[\mu_{MLE}] = \mu$$
$$\mathbb{E}[\sigma^2_{MLE}] = \tfrac{N-1}{N}\sigma^2$$

što ukazuje na to da je $\mu_{MLE}$ nepristran procjenitelj srednje vrijednosti populacije, dok $\sigma^2_{MLE}$ nije nepristran procjenitelj varijance jer $\mathbb{E}[\sigma^2_{MLE}] \neq \sigma^2$. Kako je $\frac{N-1}{N} < 1, \forall N \in \mathbb{N}$, slijedi da procjenitelj $\sigma^2_{MLE}$ uvijek podcjenjuje pravu varijancu populacije $\sigma^2$, tj. uvijek ćemo dobiti manju varijancu. Međutim, ovo se može ispraviti tako što $\sigma^2_{MLE}$ pomnožimo s $\frac{N}{N-1}$ i tako dobijemo nepristran procjenitelj varijance, tj.

$$\sigma^2_{nepr.} = \tfrac{N}{N-1}\sigma^2_{MLE} = \tfrac{1}{N-1}\textstyle\sum_{i=1}^{N}(x_i - \mu_{MLE})^2.$$

Ovakvu korekciju je potrebno raditi za mali uzorak, odnosno mali broj opažanja $N$. Kada $N \to \infty$ imamo $\lim\limits_{N\to\infty} \sigma^2_{MLE} = \sigma^2$ te stoga korekciju nije potrebno raditi. Može se pokazati kako primjenom maksimizacije funkcije log-izglednosti za normalnu razdiobu u $D$-dimenzionalnom prostoru vrijedi

$$\mu_{MLE} = \frac{1}{N}\sum_{i=1}^{N} x_i$$
$$\Sigma_{MLE} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_{MLE})(x_i - \mu_{MLE})^T$$

gdje je $x_i = (x_{i_1}, ..., x_{i_D})$, te za očekivanja dobijemo

$$\mathbb{E}[\mu_{MLE}] = \mu$$
$$\mathbb{E}[\Sigma_{MLE}] = \tfrac{N-1}{N}\Sigma$$

iz čega možemo izvući jednak zaključak o nepristranosti srednje vrijednosti i varijance kao i prije te za procjenu varijance napraviti korekciju kako bi dobili nepristranog procjenitelja:

$$\Sigma_{nepr.} = \frac{1}{N-1}(x_i - \mu_{MLE})(x_i - \mu_{MLE})^T.$$

# Poglavlje 2

# Model Gaussove mješavine

## 2.1 Model konačne mješavine

Zbog svoje fleksibilnosti prilikom modeliranja podataka, modeli miješane gustoće često su korišteni u raznim društvenim i prirodnim znanostima u svrhu bolje statističke analize, grupiranja i klasifikacije podataka te procjene gustoće istih. To su modeli koji su, dakle, mješavine više (ne nužno) različitih distribucija. Ako se model sastoji od mješavine konačnog broja distribucija kažemo da se radi o modelu konačne mješavine. Među prvima koji je u statističkoj analizi prije više od 120 godina koristio miješani model bio je slavni matematičar i statističar Karl Pearson. Na skupu podataka koji se sastojao od omjera širine frontalnog dijela glave i dužine tijela 1000 rakova uzorkovanih u Napuljskom zaljevu, Pearson je u [17] uz korištenje miješanog modela od dvije normalne distribucije pokazao pristutnost dviju podvrsta rakova.

Promotrimo populaciju koja se sastoji od K homogenih podpopulacija koje nazivamo **komponentama**. Pretpostavimo da uzimamo slučajni uzorak iz takve populacije i zapisujemo ga kao $(X_i, J_i)$ za $i = 1, ..., N$, gdje je

$X_i = x_i$ rezultat mjerenja i-te jedinice uzorka, a $J_i \in \{1, ..., K\}$ ukazuje kojoj komponenti i-ta jedinica uzorka pripada. Nadalje, ako bi uzimali uzorak samo iz k-te komponente, zbog homogenosti bi imali prikladan vjerojatnosni model za distribuciju uzorka

$$P(X = x|J = k) = p(x; \theta_k)$$

gdje je $\theta_k$ nepoznati parametar k-te komponente.
Udio cijele populacije koji se nalazi u k-toj komponenti označavamo s $\pi_k$ i nazivamo **koeficijentom mješavine** k-te komponente. Očito vrijedi $\pi_k \geqslant 0$ i $\sum_{k=1}^{K} \pi_k = 1$. Koeficijenti mješavine su obično nepoznati. Kako pretpostavljamo da smo uzeli slučajan uzorak, vjerojatnost da realizacija dolazi iz k-te komponente iznosi $P(J = k) = \pi_k$. Zajednička vjerojatnost je dana s

$$P(X = x, J = k) = P(X = x|J = k)P(J = k) = p(x; \theta_k)\pi_k.$$

Kada nam oznaka pripadnosti komponentama $J_1, ..., J_N$ nije dostupna i promatramo uzorak $X_1, ..., X_N$, zajedničku vjerojatnost dobijemo marginalizacijom po varijabli J. Stoga model miješane gustoće možemo zapisati kao

$$\begin{aligned} p(x; \pi, \theta) &= \sum_{k=1}^{K} P(X = x, J = k) \\ &= \sum_{k=1}^{K} P(J = k)P(X = x|J = k) = \sum_{k=1}^{K} \pi_k p(x; \theta_k). \end{aligned}$$

gdje je $\pi = (\pi_1, ..., \pi_K)$ za koji vrijedi $\pi_k \geqslant 0$ i $\sum_{k=1}^{K} \pi_k = 1$, $\theta = (\theta_1, ..., \theta_K)$ te $p(x; \theta_k)$ gustoća k-te komponente ([11]).

**Definicija 2.1** *Neka je $X = (X_1, ..., X_n)$ n-dimenzionalni slučajni vektor za neki $n \in \mathbb{N}$ koji poprima vrijednosti iz prostora uzoraka $\mathcal{X}$. Kažemo da $X$ dolazi iz distribucije konačne mješavine ako je funkcija gustoće od $X$ dana sa*

$$p(x; \theta) = \textstyle\sum_{k=1}^{K} \pi_k p(x; \theta_k), \quad \textit{za svaki } x \in \mathcal{X}$$

*gdje je* $\theta = (\pi_1, ..., \pi_K, \theta_1, ..., \theta_K)$ *skup parametara modela,* $p(x; \theta_k)$ *funkcija gustoće k-te komponente,* $\pi_k$ *težina k-te komponente te vrijedi* $\pi_k \geqslant 0$ *i* $\sum_{k=1}^{K} \pi_k = 1$*, za svaki* $i = 1, ..., K$*.*

## 2.2 Model Gaussove mješavine

U primjeni se većinom pretpostavlja da funkcije gustoće svih komponenti dolaze iz iste familije parametarske distribucije, tj. prate jednaku distribuciju uz različite parametre. Kada je u modelu konačne mješavine svaka od komponenti normalno distribuirana, govorimo o modelu Gaussove mješavine.

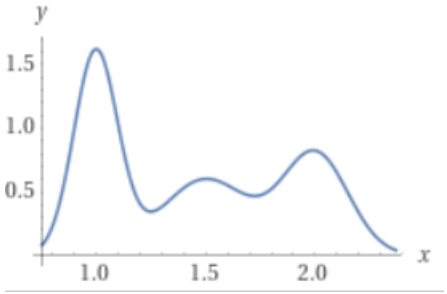
**Definicija 2.2** *Neka su* $(x_1, ..., x_N)$*,* $N \in \mathbb{N}$*, gdje je* $x_i = (x_{i_1}, ..., x_{i_D})$*, nezavisne i jednako distribuirane realizacije slučajnog vektora X dimenzije D,* $D \in \mathbb{N}$*. Kažemo da X dolazi iz distribucije modela Gaussove mješavine s K komponenti ako je funkcija gustoće od X dana sa*

$$p(x|\pi, \mu, \Sigma) = \pi_1 \mathcal{N}(x|\mu_1, \Sigma_1) + ... + \pi_K \mathcal{N}(x|\mu_K, \Sigma_K)$$

*gdje je* $\pi = (\pi_1, ..., \pi_K)$ *takav da je* $\pi_k \geqslant 0$*,* $\forall k \in \{1, ..., K\}$*,* $\sum_{k=1}^{K} \pi_k = 1$*,* $\mu = (\mu_1, ..., \mu_K), \Sigma = (\Sigma_1, ..., \Sigma_K)$ *te* $\mathcal{N}(x|\mu_k, \Sigma_k)$ *funkcija gustoće normalne razdiobe s očekivanjem* $\mu_k$ *te kovarijacijskom matricom* $\Sigma_k$ *k-te komponente, za svaki* $k = 1, ..., K$*.*

Za $D = 1$ imamo $\Sigma_k = \sigma_k^2$ pa je model Gaussove mješavine definiran sa

$$p(x|\pi, \mu, \sigma^2) = \pi_1 \mathcal{N}(x|\mu_1, \sigma_1^2) + ... + \pi_K \mathcal{N}(x|\mu_K, \sigma_K^2).$$




Slika 2.1: Gaussova mješavina dana sa

$$0.4\frac{1}{0.1\sqrt{2\pi}}e^{\frac{-(x-1)^2}{2\cdot 0.1^2}}+0.3\frac{1}{0.2\sqrt{2\pi}}e^{\frac{-(x-1.5)^2}{2\cdot 0.2^2}}+0.3\frac{1}{0.15\sqrt{2\pi}}e^{\frac{-(x-2)^2}{2\cdot 0.15^2}}$$

Postavlja se pitanje kako za dane podatke, za koje pretpostavljamo da dolaze iz modela Gaussove mješavine, pronaći optimalne parametre takvog modela koji bi najbolje naše podatke opisali. Prisjetimo se, jedan od efikasnijih načina procjene parametara neke distribucije bio je putem kriterija najveće izglednosti. Tražili smo parametre koji maksimiziraju funkciju izglednosti za dani uzorak, odnosno log-izglednost zbog jednostavnosti postupka.
Neka je $x = (x_1, ..., x_N)$ realizacija uzorka iz modela Gaussove mješavine. Funkcija log-izglednosti glasi

$$\ln \mathcal{L}(\pi, \mu, \Sigma) \equiv \ln \mathcal{L}(\pi, \mu, \Sigma|x) = \ln \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$$
$$= \sum_{i=1}^{N} \ln(\sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)).$$

Sljedeći korak bio bi izračun $\nabla \ln \mathcal{L}(\pi, \mu, \Sigma) = 0$, međutim zbog logaritma sume račun ne možemo provesti analitički te ne postoji rješenje u zatvorenoj formi. Iz tog razloga, rješenje pronalazimo putem iterativne metode - algoritmom maksimizacije očekivanja.

Prije opisa algoritma spomenimo dva problema koja se mogu javiti pri prona-

lasku rješenja najveće izglednosti. Jedan od njih su točke singularnosti. Jednostavnosti radi, pretpostavimo da promatramo Gaussov model mješavine u kojem sve komponente imaju kovarijacijsku matricu oblika $\Sigma_k = \sigma_k^2 I$, gdje je $I$ jedinična matrica. Također, pretpostavimo da jedna od komponenti modela, npr. j-ta komponenta, ima srednju vrijednost $\mu_j$ jednaku jednoj od točaka podataka, $\mu_j = x_i$ za neki $i$. Tada je $\mathcal{N}(x_i|x_i, \sigma_j^2 I) = \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sigma_j}$. Ako $\sigma_j \to 0$, onda $\mathcal{N}(x_i|x_i, \sigma_j^2 I) \to \infty$ pa i log-izglednost ide u beskonačnost. Stoga maksimizacija funkcije log-izglednosti nije dobro postavljen problem zbog prisutnosti ovakvih singularnosti kada se jedna od komponenti svede na specifičnu točku iz skupa podataka. Sljedeći problem je poznat kao **raspoznatljivost** *(engl. identifiability)* i javlja se pri interpretaciji optimalnih parametarskih vrijednosti. Za svako rješenje maksimizacije funkcije izglednosti, mješavina s K komponenti će imati ukupno $K!$ ekvivalentnih rješenja zbog $K!$ načina pridjeljivanja K skupova parametara na K komponenti. Drugim riječima, za svaku točku u prostoru parametarskih vrijednosti biti će još $K! - 1$ dodatnih točaka za koje imamo istu distribuciju ([4]). Dakle, postoji više kombinacija parametara koje daju istu funkciju izglednosti.
Sljedeće definicije mogu se pronaći u [12].

**Definicija 2.3** *Za parametarsku familiju gustoća* $\{p(x;\theta) \colon \theta \in \Theta\}$*, gdje je* $\Theta$ *parametarski prostor, kažemo da je raspoznatljiva ako različite vrijednosti parametara* $\theta$ *definiraju različite gustoće* $p(x;\theta)$*.*

Raspoznatljivost se, međutim, za modele miješane gustoće definira ponešto drugačije.

**Definicija 2.4** *Neka su* $p(x;\theta) = \sum_{k=1}^{K} \pi_k p_k(x;\theta_k)$ *i* $p(x;\theta^*) = \sum_{l=1}^{L} \pi_l^* p_l(x;\theta_l^*)$ *bilo koja dva člana parametarske familije miješanih gustoća. Za klasu modela konačne mješavine kažemo da je raspoznatljiva za* $\theta \in \Theta$ *ako* $p(x;\theta) \equiv$

*$p(x;\theta^*)$ te ako i samo ako vrijedi $K = L$ te možemo permutirati oznake komponenti tako da je $\pi_k = \pi_k^*$ i $p_k(x;\theta_k) = p_k(x;\theta_k^*)$ za svaki $k = 1, ..., K$.*

U gornjoj definiciji oznaka $\equiv$ implicira jednakost gustoća za skoro svaki $x_i$. Manjak raspoznatljivosti zbog moguće zamjene oznaka komponenti ipak nije od velike relevantnosti u praksi jer se vrlo lako može zaobići uvođenjem određenih uvjeta za parametre modela.

### 2.2.1 EM algoritam na modelu Gaussove mješavine

Vratimo se na problem pronalaska procjenitelja najveće izglednosti za model Gaussove mješavine. Za dani model Gaussove mješavine, cilj je maksimizirati funkciju izglednosti s obzirom na parametre koji se sastoje od srednjih vrijednosti, kovarijacijskih matrica i koeficijenata mješavine. Tražimo parametre koji maksimiziraju funkciju log-izglednosti

$$\ln \mathcal{L}(\pi, \mu, \Sigma) = \sum_{i=1}^{N} \ln(\sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)).$$

Suma unutar logaritma stvara problem te ne postoji rješenje u zatvorenoj formi. Međutim, procjenitelja najveće izglednosti možemo pronaći putem iterativne metode poznate kao algoritam maksimizacije očekivanja. Prema [7] gdje je prvi put formalno definiran **algoritam maksimizacije očekivanja** ili **EM algoritam**[1] model mješavine se uvijek može proširiti skrivenim, odnosno latentnim varijablama - varijablama koje ne opažamo u podacima, ali imaju utjecaj na opažene varijable. Primjer latentne varijable je upravo oznaka kojoj komponenti određeno opažanje pripada. Ako definiramo zajedničku distribuciju opaženih i latentnih varijabli, odnosno potpunih podataka, distribuciju samo opaženih varijabli možemo dobiti marginalizacijom.

[1] engl. Expectation-Maximization (EM) algorithm

Ovo omogućuje relativno kompleksnim marginalnim distribucijama da budu izražene preko zajedničkih distribucija, koje je lakše pratiti, na proširenom prostoru varijabli s latentnima ([4]).

Opis EM algoritma na modelu Gaussove mješavine preuzet je iz [4]. Pretpostavimo da imamo skup nezavisnih opažanja $\{x_1, ..., x_N\}$, $N \in \mathbb{N}$, za koje pretpostavljamo da dolaze iz distribucije modela Gaussove mješavine. Svako opažanje ima $D$ obilježja, tj. $x_i = (x_{i_1}, ..., x_{i_D})$ za neki $D \in \mathbb{N}$. Također, neka je $z_i = (z_{i_1}, ...., z_{i_K})$, $i \in \{1, ..., N\}$, K-dimenzionalni slučajni vektor takav da je $z_{i_k} \in \{0, 1\}$, $\forall k = 1, ..., K$ i $\sum_{k=1}^{K} z_{i_k} = 1$, tj. $z_i$ je latentna varijabla koja ukazuje kojoj komponenti pripada opažanje $x_i$:

$$z_{i_k} = \begin{cases} 1, & \text{ako je opažanje } x_i \text{ u k-toj komponenti} \\ 0, & \text{inače} \end{cases}$$

Zbog jednostavnijeg zapisa ovdje ćemo latentne varijable označiti sa $z$ umjesto $z_i$, ali i dalje vrijedi činjenica da za svako opažanje $x_i$ imamo po jednu latentnu varijablu $z$. Primijetimo da vrijedi $P(z_k = 1) = \pi_k$, odnosno vjerojatnost k-te komponente je upravo koeficijent mješavine k-te komponente $\pi_k$.

$$P(z) = \prod_{k=1}^{K} \pi_k^{z_k}$$
$$p(x|z_k = 1) = \mathcal{N}(x|\mu_k, \Sigma_k)$$
$$p(x|z) = \prod_{k=1}^{K} \mathcal{N}(x|\mu_k, \Sigma_k)^{z_k}.$$

Konačno, imamo

$$p(x, z) = P(z)p(x|z) = \prod_{k=1}^{K} \pi_k^{z_k} \mathcal{N}(x|\mu_k, \Sigma_k)^{z_k}$$

što je funkcija gustoće modela Gaussove mješavine koji uključuje latentne varijable.

Pogledajmo uvjetnu vjerojatnost $P(z_{i_k} = 1|x_i)$ i označimo je s $\gamma_k(x_i)$. Po Bayesovom teoremu slijedi

$$\gamma_k(x_i) = P(z_{i_k} = 1|x_i) = \frac{P(z_{i_k} = 1)p(x_i|z_{i_k} = 1)}{\sum_{j=1}^{K} P(z_{i_j} = 1)p(x_i|z_{i_j} = 1)} = \frac{\pi_k \mathcal{N}(x_i|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_i|\mu_j, \Sigma_j)}.$$

$\gamma_k(x_i)$ je vjerojatnost da je k-ta komponenta odgovorna za generiranje opažanja $x_i$ te je zovemo **odgovornost** ([4]).

Označimo sa X matricu tipa $N \times D$ čije redove čine opažanja $x_i$, $i \in \{1, ..., N\}$ i sa Z matricu tipa $N \times K$ čije redove čine latentne varijable $z_i$, $i \in \{1, ..., K\}$. Napišimo uvjete koji moraju biti zadovoljeni u maksimumu funkcije log-izglednosti:

1. $\frac{\partial}{\partial \mu_k} \ln \mathcal{L}(\pi, \mu, \Sigma|X) = 0$

   Iz ovog uvjeta slijedi

   $$0 = -\sum_{i=1}^{N} \underbrace{\frac{\pi_k \mathcal{N}(x_i|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_i|\mu_j, \Sigma_j)}}_{\gamma_k(x_i)} \Sigma_k(x_i - \mu_k)$$

   Množimo s $\Sigma_k^{-1}$ i tako dođemo do

   $$\mu_k = \frac{1}{N_k} \sum_{i=1}^{N} \gamma_k(x_i) x_i \tag{2.1}$$

   gdje je

   $$N_k = \sum_{i=1}^{N} \gamma_k(x_i) \tag{2.2}$$

   Vidimo da je srednja vrijednost $\mu_k$ za k-tu Gaussovu komponentu dobivena putem težinske srednje vrijednosti svih točaka iz skupa podataka u kojoj je težinski faktor za točku $x_i$ dan sa posteriornom vjerojatnošću $\gamma_k(x_i)$ da je k-ta komponenta odgovorna za generiranje $x_i$.

2. $\frac{\partial}{\partial \Sigma_k} \ln \mathcal{L}(\pi, \mu, \Sigma|X) = 0$

   Na isti način, uz korištenje rješenja najveće izglednosti za kovarijacijsku matricu jedne normalne razdiobe, dobijemo

$$\Sigma_k = \frac{1}{N_k} \sum_{i=1}^{N} \gamma_k(x_i)(x_i - \mu_k)(x_i - \mu_k)^T \tag{2.3}$$

   gdje ponovno za svaku točku iz skupa podataka imamo težinu po pripadnoj posteriornoj vjerojatnosti $\gamma_k(x_i)$ i $N_k = \sum_{i=1}^{N} \gamma_k(x_i)$.

3. $\max \ln \mathcal{L}(\pi, \mu, \Sigma|X)$

   $\pi_k \geqslant 0$, $\sum_{k=1}^{K} \pi_k = 1$

   Konačno, maksimiziramo $\ln \mathcal{L}(\pi, \mu, \Sigma|X)$ s obzirom na koeficijente mješavine $\pi_k$ s tim da moramo uzeti u obzir uvjete koeficijenata mješavine. Ovaj problem predstavlja optimizacijski problem i može se riješiti korištenjem Lagrangeovih multiplikatora i maksimiziranjem sljedećeg:

$$\ln \mathcal{L}(\pi, \mu, \Sigma|X) + \lambda(\sum_{k=1}^{K} \pi_k - 1),$$

   gdje je $\lambda$ Lagrangeov multiplikator, čime dolazimo do

$$0 = \sum_{i=1}^{N} \frac{\mathcal{N}(x_i|\mu_k, \Sigma_k)}{\sum_j \pi_j \mathcal{N}(x_i|\mu_j, \Sigma_j)} + \lambda$$

   gdje opet možemo uočiti pojavu odgovornosti. Sada množimo obje strane sa $\pi_k$ i sumiramo po k koristeći uvjet $\sum_{k=1}^{K} \pi_k = 1$ čime dolazimo do $\lambda = -N$. Ponovnim uvrštavanjem $\lambda$ dolazimo do $\pi_k = \frac{N_k}{N}$. Težinski koeficijent za k-tu komponentu dan je sa prosječnom odgovornošću koju ta komponenta preuzima za generiranje točaka podataka.

Gornji rezultati ne predstavlju zatvorenu formu rješenja za parametre modela mješavine jer odgovornosti $\gamma_k(x_i)$ ovise o njima. Međutim, rezultati sugestiraju jednostavnu iterativnu shemu za pronalazak rješenja problema

najveće izglednosti za što ćemo kasnije pokazati da je instanca EM algoritma za specijalan slučaj modela Gaussove mješavine.

Prvo odaberemo neke inicijalne vrijednosti za srednje vrijednosti, kovarijacijske matrice i koeficijente mješavine. Zatim alterniramo između dva koraka koja nazivamo E-korak i M-korak. U E-koraku *(engl. expectation step)* koristimo trenutne vrijednosti parametara za izračun posteriornih vjerojatnosti, tj. odgovornosti. Nakon toga, koristimo te vjerojatnosti u M-koraku *(engl. maximization step)* za novu procjenu srednjih vrijednosti, kovarijacijskih matrica te koeficijenata mješavine koristeći gore izvedene formule. Pritom se prvo računaju srednje vrijednosti te zatim koristimo njihove nove vrijednosti za izračun kovarijanci.

**Algoritam 1** EM algoritam na modelu Gaussove mješavine

1. Inicijaliziraj $\mu_k, \Sigma_k, \pi_k$ za svaku od K komponenti modela te izračunaj početnu vrijednost log-izglednosti.

2. **E-korak**

   Izračunaj odgovornosti koristeći trenutne vrijednosti parametara.

   $\gamma_k(x_i) = \frac{\pi_k \mathcal{N}(x_i|\mu_k,\Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_i|\mu_j,\Sigma_j)}$

3. **M-korak**

   Izračunaj novu procjenu parametara koristeći izračunate odgovornosti.

   $\mu_k = \frac{1}{N_k} \sum_{i=1}^{N} \gamma_k(x_i)$

   $\Sigma_k = \frac{1}{N_k} \sum_{i=1}^{N} \gamma_k(x_i)(x_i - \mu_k)(x_i - \mu_k)^T$

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

   gdje je $N_k = \sum_{i=1}^{N} \gamma_k(x_i)$.

4. Izračunaj funkciju log-izglednosti

   $\ln \mathcal{L}(\pi, \mu, \Sigma|X) = \sum_{i=1}^{N} \ln(\sum_{k=1}^{K} \mathcal{N}(x_i|\mu_k, \Sigma_k))$

   te provjeri konvergenciju za log-izglednost ili za parametre. Ako uvjet zaustavljanja nije zadovoljen vrati se na korak 2.

### 2.2.2 Algoritam K-sredina

Algoritam K-sredina (*engl. K-means*) jedan je od najčešće korištenih algoritama grupiranja te je specijalni slučaj EM algoritma na modelu Gaussove mješavine. Dok se kod izvođenja EM algoritma na modelu Gaussove mješavine računa posteriorna vjerojatnost, odnosno odgovornost komponenti za generiranje opažanja i time za neko opažanje imamo vjerojatnosti pripadnosti za više komponenti (ovakvo grupiranje se još naziva i **meko grupira-**

**nje**), kod algoritma K-sredina jedno opažanje pripada točno jednoj komponenti (tzv. **čvrsto grupiranje**).

Pretpostavimo da imamo model Gaussove mješavine u kojem su kovarijacijske matrice komponenti dan sa $\epsilon I$, gdje je $\epsilon$ varijanca koja je jednaka za sve komponente te $I$ jedinična matrica. Dakle, k-ta komponenta modela je dana sa

$$\mathcal{N}(x|\mu_k, \Sigma_k) = \frac{1}{(2\pi\epsilon)^{1/2}} e^{-\frac{1}{2\epsilon}\|x-\mu_k\|^2}.$$

Sada promatramo EM algoritam na ovako danom modelu mješavine od K normalnih distribucija gdje je $\epsilon$ fiksirana konstanta, a ne parametar kojeg treba procijeniti. Posteriorna vjerojatnost, odnosno odgovornost, za opažanje $x_i$ dano je sa

$$\gamma_k(x_i) = \frac{\pi_k e^{-\|x_i-\mu_k\|^2/2\epsilon}}{\sum_{j=1}^{K} \pi_j e^{-\|x_i-\mu_j\|^2/2\epsilon}}.$$

Kada $\epsilon \to 0$, vidimo da će član u nazivniku za koji je $\|x_i - \mu_j\|^2$ najmanji ići najsporije prema nuli, a time i odgovornosti $\gamma_k(x_i)$ ta opažanje $x_i$ idu u nulu, osim j-tog člana za kojeg odgovornost $\gamma_j(x_i)$ koji ide u jedinicu. Primijetimo da ovo vrijedi neovisno o vrijednostima koeficijenata mješavine $\pi_k$ sve dok nijedan od njih nije nula. Stoga, kada je $\epsilon \to 0$, imamo

$$\gamma_k(x_i) \to \begin{cases} 1, & \text{ako je } k = \underset{j}{argmin}\, \|x_i - \mu_j\|^2 \\ 0, & \text{inače} \end{cases}$$

Dakle, svako opažanje je dodijeljeno komponenti s najbližom srednjom vrijednosti. Kako sada imamo čvrsto grupiranje, vrijedi $\gamma_k(x_i) = 1$ ako je $x_i$ dodijeljen k-toj komponenti te $\gamma_j(x_i) = 0$ za sve $j \neq k$.

Neka je

$$J = \sum_{i=1}^{N} \sum_{k=1}^{K} \gamma_k(x_i) \|x_i - \mu_k\|^2,$$

tj. $J$ je suma kvadrata udaljenosti svakog opažanja od srednje vrijednosti komponente kojoj je dodijeljen. Cilj je pronaći $\{\gamma_k(x_i)\}$ te $\{\mu_k\}$ tako da minimiziramo $J$. Prvo se izabiru inicijalne vrijednosti za $\mu_k$ te nakon toga iteriraju dva koraka do konvergencije:

1. Fiksiramo $\mu_k$ i minimiziramo $J$ obzirom na $\gamma_k(x_i)$.

   Pridjeljujemo opažanje $x_i$ komponenti s najbližom srednjom vrijednosti

$$\gamma_k(x_i) = \begin{cases} 1, & \text{ako je } k = \underset{j}{argmin}\, \|x_i - \mu_j\|^2 \\ 0, & \text{inače} \end{cases}$$

2. Fiksiramo $\gamma_k(x_i)$ i minimiziramo $J$ obzirom na $\mu_k$.

$$\frac{\partial J}{\partial \mu_k} = 0 \implies 2\sum_{i=1}^{N} \gamma_k(x_i)(x_i - \mu_k) = 0 \implies \mu_k = \frac{\sum_{i=1}^{N} \gamma_k(x_i) x_i}{\sum_{i=1}^{N} \gamma_k(x_i)}$$

   Primijetimo da u nazivniku $\sum_{i=1}^{N} \gamma_k(x_i)$ označava broj opažanja u k-toj komponenti. Dakle, $\mu_k$ se interpretira kao srednja vrijednost svih opažanja $x_i$ u k-toj komponenti.

Ova dva koraka korespondiraju s E i M korakom EM algoritma te se iteriraju do konvergencije - dok nema promjene u srednjim vrijednostima komponenti $\mu_k$. Dakle, algoritam K-sredina procjenjuje samo srednje vrijednosti komponenti, tj. njihove centre, a ne procjenjuje kovarijacijske matrice. Verzija modela Gaussove mješavine s čvrstim grupiranjem poput ovdje opisanog te s općim kovarijacijskim matricama naziva se eliptični algoritam K-sredina (*engl. elliptical K-means*) ([4]).

**Algoritam 2** Algoritam K-sredina

Inicijaliziraj $\mu_k$, $k = 1, ..., K$.

**ponavljaj**

1. za svaki $x_i$, $i = 1, ..., N$ izračunaj
$$\gamma_k(x_i) = \begin{cases} 1, & \text{ako je } k = \underset{j}{argmin} \|x_i - \mu_j\|^2 \\ 0, & \text{inače} \end{cases}$$
2. za svaki $\mu_k$, $k = 1, ..., K$ izračunaj
$$\mu_k = \frac{\sum_{i=1}^{N} \gamma_k(x_i) x_i}{\sum_{i=1}^{N} \gamma_k(x_i)}$$

**dok** $\mu_k$ ne konvergira

## 2.3 Optimalan broj komponenti

Broj komponenti K modela najčešće nije unaprijed poznat, a njegov odabir utječe na točnost modela te rješenje problema najveće izglednosti. Kako pronaći optimalnu vrijednost za K, tj. broj komponenti modela koje će na najbolji mogući način modelirati dane podatke?

**Definicija 2.5** *Neka je $p(x)$ neka distribucija i $q(x)$ njena aproksimacija. Mjera udaljenosti, odnosno odstupanja $q(x)$ od $p(x)$ definira se kao*

$$KL(p\|q) = \int p(x) \ln \frac{p(x)}{q(x)} \, dx$$

*i naziva **Kullback-Leiblerova divergencija** između distribucija $p(x)$ i $q(x)$.*

Primijetimo kako ova mjera nije simetrična, tj. $KL(p(\|q) \neq KL(q\|p)$. Kullback-Leiblerova divergencija može se shvatiti kao mjera gubitka informacija prilikom procjene prave funkcije gustoće $p(x)$ s njenom aproksimacijom $q(x)$ ([6]). Očito je cilj umanjiti gubitak informacija, odnosno pronaći

aproksimaciju koja minimizira Kullback-Leiblerovu divergenciju.

Pretpostavimo da modeliramo neku nepoznatu distribuciju $p(x)$ modelom Gaussove mješavine $q(x;\theta)$ sa skupom parametara $\theta$. Jedan od načina za određivanje $\theta$ jest minimiziranje Kullback-Leibler divergencije između $p(x)$ i $q(x;\theta)$ obzirom na $\theta$. Međutim, ovo se ne može direktno izvesti jer nam je $p(x)$ nepoznata. Pretpostavimo sada da imamo konačan skup opažanja $x_1, ..., x_n$ za neki $n \in \mathbb{N}$ iz distribucije $p(x)$. Tada se očekivanje obzirom na $p(x)$ može aproksimirati konačnom sumom tih opažanja te imamo $KL(p\|q) \approx \sum_{i=1}^{n}[-\ln q(x_i;\theta) + \ln p(x_i)]$. Kako $\ln p(x_i)$ ne ovisi o $\theta$ tako je minimiziranje Kullback-Leiblerove divergencije ekvivalentno maksimiziranju funkcije izglednosti ([4]).

U [1] je uočena veza između Kullback-Leibler divergencije i najveće izglednosti te dana definicija Akaikeovog informacijskog kriterija (AIC) koji se često koristi za određivanje broja komponenti modela:

$$\mathrm{AIC} = -2\ln\mathcal{L}(\hat{\theta}) + 2R,$$

gdje je $\mathcal{L}(\hat{\theta})$ maksimalna vrijednost funkcije izglednosti s nepoznatim parametrom $\theta$ za dane podatke i model te $R$ ukupni broj parametara ([6]). Upravo se Akaikeov informacijski kriterij često koristi za određivanje broja komponenti K modela. Za model mješavine s K komponenti jasno je da ukupni broj parametara modela ovisi o samom broju komponenti te ćemo ga iz tog razloga označiti s $R(K)$. Dakle, imamo $AIC = -2\ln\mathcal{L}(\hat{\theta}) + 2R(K)$ te broj K komponenti modela mješavine pronalazimo minimiziranjem tog izraza:

$$K_{AIC} = \underset{K}{argmin}\,(-2\ln\mathcal{L}(\hat{\theta}) + 2R(K)).$$

### 2.3. Optimalan broj komponenti

Osim Akaikeovog informacijskog kriterija, u primjeni se često koristi i Bayesov informacijski kriterij:

$$\text{BIC} = -2 \ln \mathcal{L}(\hat{\theta}) + R(K) \ln n,$$

gdje su $\mathcal{L}(\hat{\theta})$ i $R(K)$ definirani kao i prije, a $n$ je broj opažanja u uzorku. Bayesov informacijski kriterij također uzima najmanju vrijednost za optimalan broj komponenti modela mješavine:

$$K_{BIC} = \underset{K}{argmin} \, (-2 \ln \mathcal{L}(\hat{\theta}) + R(K) \ln n).$$

U [14] se osim spomenutih informacijskih kriterija može pronaći još načina za određivanje broja komponenti u modelu Gaussove mješavine.

# Poglavlje 3

# Algoritam maksimizacije očekivanja (EM algoritam)

U ovom dijelu rada opisat ćemo opći EM algoritam, iterativnu metodu kojom se pronalazi procjenitelj najveće izglednosti parametra $\theta$ neke parametarske vjerojatnosne distribucije, te dati njegov opis za opći model konačne mješavine. Prema [7], u kojem je prvi put definiran EM algoritam, model mješavine se uvijek može proširiti skrivenim, odnosno latentnim varijablama - varijablama koje ne opažamo u podacima, ali imaju utjecaja na opažene varijable. Neka je $x = (x_1, ..., x_n)$, $x_i \in \mathbb{R}^d$, skup opaženih podataka za koje pretpostavljamo da dolaze iz distribucije $p(x; \theta)$ gdje je $\theta$ parametar modela. Potpuni podaci $y = (y_1, ..., y_n)$ dani su sa $y_i = (x_i, z_i)$, $i = 1, ..., n$, gdje su $z_i$ latentne varijable. Vrijedi $p(x; \theta) = \int p(x, z; \theta)\, dz$. Ideja EM algoritma je u svakoj iteraciji maksimizirati uvjetno očekivanje log-izglednosti modela za potpune podatke kako bi pronašli procjenitelja najveće izglednosti. Kao što je već ranije u radu naglašeno, tijek rada EM algoritma se sastoji od dva koraka (E i M). U [10] je međutim dan detaljan opis algoritma u pet koraka:

1. Neka je $m = 0$ i $\theta^{(m)}$ inicijalne vrijednosti od $\theta$.

2. Za dana opažanja $x$ i uz pretpostavku da su procjene $\theta^{(m)}$ točne, izvedi uvjetnu funkciju gustoće $p(y|x;\theta^{(m)})$ za potpune podatke $y$.

3. Koristeći uvjetnu funkciju gustoće $p(y|x;\theta^{(m)})$ izvedi uvjetno očekivanje log-izglednosti

$$Q(\theta|\theta^{(m)}) = \int_{\mathcal{Y}(x)} \ln p(y;\theta) p(y|x;\theta^{(m)})\, dy = \mathbb{E}_{Y|x,\theta^{(m)}}[\ln p(Y;\theta)],$$

gdje je Y slučajna varijabla koja predstavlja potpune podatke, a $\mathcal{Y}(x) = Cl\{y\colon p(y|x;\theta) > 0\}$ zatvorenje skupa $\{y\colon p(y|x;\theta) > 0\}$ i pretpostavljamo da $\mathcal{Y}(x)$ ne ovisi o $\theta$. Primijetimo kako je $Q$-funkcija funkcija od $\theta$, ali ovisi o trenutnim procjenama parametara $\theta^{(m)}$.

4. Pronađi $\theta$ takav da maksimizira $Q$-funkciju i rezultat je nova procjena $\theta^{(m)}$.

5. Neka je $m := m + 1$ i vrati se na korak 2.

Standardni uvjeti zaustavljanja EM algoritma su iteriranje sve dok se procjene parametara ne prestanu mijenjati, tj. $\|\theta^{(m+1)} - \theta^{(m)}\| < \epsilon$ za neki unaprijed određen prag $\epsilon > 0$, ili iteriranje dok se vrijednosti funkcije log-izglednosti ne prestanu mijenjati, odnosno dok ne vrijedi $|\ln \mathcal{L}(\theta^{(m+1)}) - \ln \mathcal{L}(\theta^{(m)})| < \epsilon$ za neki $\epsilon > 0$.

Sada ćemo dati standardni opis EM algoritma koji se sastoji od dva koraka. Na ulazu u algoritam dajemo početne vrijednosti parametara $\theta^{(0)}$.

**Algoritam 3** EM algoritam

1. **E-korak**

   Za danu procjenu iz prethodne iteracije $\theta^{(m)}$, izračunaj uvjetno očekivanje $Q(\theta|\theta^{(m)})$ dano sa

   $Q(\theta|\theta^{(m)}) = \int_{\mathcal{Y}(x)} \ln p(y;\theta) p(y|x;\theta^{(m)})\, dy = \mathbb{E}_{Y|x,\theta^{(m)}}[\ln p(Y;\theta)]$

2. **M-korak**

   $\theta^{(m+1)} = \underset{\theta\in\Theta}{argmax}\, Q(\theta|\theta^{(m)})$.

Općenito, možemo zapisati $Q$-funkciju kao integral na domeni od $Z$, označenoj sa $\mathcal{Z}$, radije nego po domeni od $Y$ iz razloga što je jedini slučajni dio, tj. varirajući dio potpunih podataka $Y$ upravo skup podataka koji nedostaju, odnosno $Z$. Dakle, za problem podataka koji nedostaju gdje je $y = (x, z)$ imamo

$$\begin{aligned} Q(\theta|\theta^{(m)}) &= \int_{\mathcal{Y}} \ln p(y;\theta) p(y|x;\theta^{(m)})\, dy \\ &= \int_{\mathcal{Z}} \ln p(x,z;\theta) p(x,z|x;\theta^{(m)})\, dz \\ &= \int_{\mathcal{Z}} \ln p(x,z;\theta) p(z|x;\theta^{(m)})\, dz \\ &= \mathbb{E}_{Z|x,\theta^{(m)}}[\ln p(x,Z;\theta)]. \end{aligned}$$

### Inicijalizacija algoritma

Nedostatak EM algoritma je taj što njegovo rješenje ovisi o početnim vrijednostima na ulazu algoritma. U [3] uspoređene su neke od metoda inicijalizacije EM algoritma na modelu Gaussove mješavine uz unaprijed definiranim brojem iteracija algoritma. Preciznije, uspoređene su sljedeće metode inicijalizacije: CEM (*engl. classification EM algoritam*), SEM (*engl. stochastic*

*EM algoritam*), kratko izvođenje EM algoritma (*engl. short runs of EM*) te, u primjeni najčešće korištena, metoda slučajne inicijalizacije (*engl. random initialization*) čiji se opisi mogu pronaći u [3]. Usporedba je napravljena kroz numeričke eksperimente putem kojih je zaključeno da je slučajna inicijalizacija često lošija metoda od CEM, SEM ili metode kratkog izvođenja EM algoritma. Također, za niti jednu od spomenutih strategija ne može se reći da je najbolja te je teško okarakterizirati situaciju u kojoj bi mogli očekivati da će pojedina strategija biti bolja od preostalih. Autori su u [3] preporučili metodu inicijalizacije putem kratkog izvođenja EM algoritma, u [2] nazvanu mali EM (*engl. small EM*), tj. metode koja se sastoji od većeg broja kratkog izvođenja EM algoritma što znači da se ne čeka konvergencija algoritma već se algoritam zaustavlja nakon određenog broja iteracija. Nakon toga, EM algoritam se pokreće uz početne vrijednosti parametara uz koje dobijemo najveću vrijednost funkcije izglednosti tijekom prethodno spomenutih kratkih izvođenja EM algoritma.

Kako je već spomenuto, u praksi se često koristi metoda slučajne inicijalizacije za pronalazak početnih vrijednosti parametara na ulazu u EM algoritam. Spomenuta metoda se sastoji od pokretanja EM algoritma više puta do njegove konvergencije uz više različitih slučajno odabranih početnih vrijednosti. Zatim se konačno pokreće EM algoritam uz one početne vrijednosti parametara koji su dali najveću vrijednost funkcije izglednosti. Osim metode slučajne inicijalizacije, znaju se koristiti i neki algoritmi grupiranja poput algoritma K-sredina (*engl. K-means*), ili neke hijerarhijske metode ako broj opažanja nije "prevelik", u svrhu pronalaska optimalnih inicijalnih vrijednosti EM algoritma ([13]).

## 3.1 MAP EM algoritam

Može se dogoditi da EM algoritam podbaci zbog postojanja singularnosti funkcije izglednosti, odnosno funkcije log-izglednosti, pogotovo u slučaju kada je broj komponenti modela prevelik u odnosu na broj opažanja ([10]). Problem singularnosti se može pojaviti u modelu Gaussove mješavine obzirom da funkcija izglednosti nije omeđena odozgo. Prisjetimo se, u dijelu rada gdje smo definirali model Gaussove mješavine spomenut je problem singularnosti koji se javlja kada je najvjerojatnije rješenje to da je nekoj komponenti modela pridijeljena samo jedna točka opažanja. Međutim, taj problem može se zaobići uvođenjem nekih *a priori* informacija o rješenju za $\theta$. Jedan od pristupa jest ograničiti skup mogućih vrijednosti za $\theta$. Tada modificiramo EM algoritam tako što ga proširujemo na *maksimum a posteriori* (MAP) procjenu,

$$\theta_{MAP} = \underset{\theta}{argmax}\, \ln p(\theta|y) = \underset{\theta}{argmax}\, (\ln p(y|\theta) + \ln p(\theta)).$$

**Napomena 3.1** *MAP procjenitelj kombinira znanje koje imamo o mogućim vrijednostima parametara s informacijama koje dobivamo iz opažanja. Takvo znanje se definira putem apriorne distribucije parametara, $p(\theta)$, koja nam ukazuje na to koje su vrijednosti parametara više te koje manje vjerojatne.*

**Algoritam 4** MAP EM algoritam

1. **E-korak**

   Za danu procjenu iz prethodne iteracije $\theta^{(m)}$, izračunaj

   $Q(\theta|\theta^{(m)}) = \mathbb{E}_{Y|x,\theta^{(m)}}[\ln p(Y;\theta)]$

2. **M-korak**

   $\theta^{(m+1)} = \underset{\theta}{argmax}\, [Q(\theta|\theta^{(m)}) + \ln p(\theta)].$

## 3.2 Generalizirani EM algoritam (GEM)

Rješenje M-koraka EM algoritma najčešće postoji u zatvorenoj formi. U iznimnim slučajevima gdje takvo rješenje ne postoji, definira se generalizirani EM algoritam gdje se u M-koraku vrijednost za $\theta^{(m+1)}$ odabire tako da vrijedi

$$Q(\theta^{(m+1)}|\theta^{(m)}) \geqslant Q(\theta^{(m)}|\theta^{(m)}).$$

Odnosno, biramo $\theta^{(m+1)}$ na način da uvećamo vrijednost $Q$-funkcije, $Q(\theta|\theta^{(m)})$, u odnosu na vrijednost u $\theta = \theta^{(m)}$ (za razliku od maksimiziranja po svim $\theta \in \Theta$, gdje je $\Theta$ parametarski prostor). Ovako opisan algoritam definira funkciju mapiranja $M\colon \Theta \to \Theta$, $\theta \mapsto M(\theta)$, takvu da vrijedi

$$\theta^{(m+1)} = M(\theta^{(m)}),\ m = 0, 1, 2, ....$$

Ako niz $\theta^{(m)}$ konvergira ka nekoj vrijednosti $\theta^*$ i $M$ je neprekidna funkcija, tada $\theta^*$ zadovoljava jednakost

$$\theta^* = M(\theta^*),$$

tj. $\theta^*$ je fiksna točka funkcije mapiranja $M$ ([16]).

## 3.3 Monotonost EM algoritma

U [7] je pokazano kako funkcija izglednosti (za nepotpune podatke) $\mathcal{L}(\theta)$ ne opada nakon iteracije EM algoritma, tj. da vrijedi

$$\mathcal{L}(\theta^{(m+1)}) \geqslant \mathcal{L}(\theta^{(m)}) \tag{3.1}$$

za $m = 0, 1, 2, ....$ U [16] se može pronaći sljedeći izvod monotonosti funkcije izglednosti tijekom EM iteracija. Neka je sa $Y$ označen skup potpunih podataka, tj. $Y = (X, Z)$, gdje je $X$ skup opaženih, a $Z$ skup latentnih

varijabli. Označimo sa $p(y|x;\theta) = \frac{p(y,x;\theta)}{p(x;\theta)}$ uvjetnu gustoću od Y za dani X = x. Možemo pisati i $p(y|x;\theta) = p(x,z|x;\theta) = p(z|x;\theta)$ pa gledamo uvjetnu gustoću $p(z|x;\theta) = \frac{p(x,z;\theta)}{p(x;\theta)}$. Tada je log-izglednost dana sa

$$\ln \mathcal{L}(\theta|X) = \ln p(x;\theta) = \ln p(z,x;\theta) - \ln p(z|x;\theta) = \ln \mathcal{L}(\theta|Y) - \ln p(z|x;\theta)$$

Računanjem očekivanja obje strane gornje jednakosti obzirom na uvjetnu distribuciju od Z za dani X = x te vrijednosti parametra $\theta = \theta^{(m)}$ imamo

$$\begin{aligned}\ln \mathcal{L}(\theta|X) &= \mathbb{E}_{Z|x,\theta^{(m)}}[\ln \mathcal{L}(\theta|Y)] - \mathbb{E}_{Z|x,\theta^{(m)}}[\ln p(Z|x;\theta)] \\ &= Q(\theta|\theta^{(m)}) - H(\theta|\theta^{(m)}),\end{aligned}$$

gdje je $H(\theta|\theta^{(m)}) = \mathbb{E}_{Z|x,\theta^{(m)}}[\ln p(Z|x;\theta)]$.
Slijedi,

$$\begin{aligned}&\ln \mathcal{L}(\theta^{(m+1)}|X) - \ln \mathcal{L}(\theta^{(m)}|X) = \\ &= [Q(\theta^{(m+1)}|\theta^{(m)}) - Q(\theta^{(m)}|\theta^{(m)})] - [H(\theta^{(m+1)}|\theta^{(m)}) - H(\theta^{(m)}|\theta^{(m)})].\end{aligned}$$

Vrijedi $Q(\theta^{(m+1)}|\theta^{(m)}) - Q(\theta^{(m)}|\theta^{(m)}) \geqslant 0$ jer se vrijednost $\theta^{(m+1)}$ izabire tako da vrijedi $Q(\theta^{(m+1)}|\theta^{(m)}) \geqslant Q(\theta|\theta^{(m)})$, za svaki $\theta \in \Theta$.
Pokažimo da vrijedi $H(\theta^{(m+1)}|\theta^{(m)}) - H(\theta^{(m)}|\theta^{(m)}) \leqslant 0$.

**Napomena 3.2** *Kažemo da je funkcija $f\colon \mathbb{R}^n \to \mathbb{R}$ konveksna ako vrijedi*

$$f(\lambda x + (1-\lambda)y) \leqslant \lambda f(x) + (1-\lambda) f(y),$$

*za svaki $x, y \in \mathbb{R}^n$ i $\lambda \in [0,1]$.*
*Za funkciju $f$ kažemo da je konkavna ako je $-f$ konveksna. Uočimo da je logaritam konkavna funkcija.*

**Napomena 3.3 (Jensenova nejednakost)** *Neka je $f$ konveksna funkcija i $X$ slučajna varijabla. Tada vrijedi $\mathbb{E}[f(X)] \geqslant f(\mathbb{E}X)$.*

Za proizvoljni $\theta$, imamo

$$\begin{aligned} H(\theta|\theta^{(m)}) - H(\theta^{(m)}|\theta^{(m)}) &= \mathbb{E}_{Z|x,\theta^{(m)}}\left[\ln \frac{p(Z|x;\theta)}{p(Z|x;\theta^{(m)})}\right] \\ &\leqslant \ln \mathbb{E}_{Z|x,\theta^{(m)}}\left[\frac{p(Z|x;\theta)}{p(Z|x;\theta^{(m)})}\right] \\ &= \ln \int \frac{p(z|x;\theta)}{p(z|x;\theta^{(m)})} p(z|x;\theta^{(m)})\, dz \\ &= \ln \int p(z|x;\theta)\, dz \\ &= \ln 1 \\ &= 0, \end{aligned}$$

gdje je nejednakost posljedica Jensenove nejednakosti i konkavnosti logaritamske funkcije. Dakle, za $\theta = \theta^{(m+1)}$ vrijedi nejednakost

$$H(\theta^{(m+1)}|\theta^{(m)}) - H(\theta^{(m)}|\theta^{(m)}) \leqslant 0$$

te imamo

$$\begin{gathered} \ln \mathcal{L}(\theta^{(m+1)}|X) - \ln \mathcal{L}(\theta^{(m)}|X) = \\ [\underbrace{Q(\theta^{(m+1)}|\theta^{(m)}) - Q(\theta^{(m)}|\theta^{(m)})}_{\geqslant 0}] - [\underbrace{H(\theta^{(m+1)}|\theta^{(m)}) - H(\theta^{(m)}|\theta^{(m)})}_{\leqslant 0}] \geqslant 0, \end{gathered}$$

tj. $\ln \mathcal{L}(\theta^{(m+1)}|X) \geqslant \ln \mathcal{L}(\theta^{(m)}|X)$, stoga konačno $\mathcal{L}(\theta^{(m+1)}|X) \geqslant \mathcal{L}(\theta^{(m)}|X)$, tj. vrijedi 3.1.

Na ovaj način smo pokazali kako funkcija izglednosti ne opada tijekom iteracije EM algoritma te takvo svojstvo nazivamo monotonost EM algoritma. Funkcija izglednosti raste ukoliko vrijedi stroga nejednakost za $Q$-funkciju, tj. $Q(\theta^{(m+1)}|\theta^{(m)}) > Q(\theta|\theta^{(m)})$. Stoga, omeđeni niz vrijednosti funkcije izglednosti $(\mathcal{L}(\theta^{(m)}))_{m\in\mathbb{N}_0}$ konvergira monotono prema nekom $L^*$.

U [10] je također izvedena monotonost EM algoritma te pokazano kako $Q$-funkcija proizvodi donju granicu za funkciju izglednosti. Naime, pokazano

je da vrijedi $\ln \mathcal{L}(\theta) \geqslant \ln \mathcal{L}(\theta^{(m)})$ ako i samo ako $Q(\theta|\theta^{(m)}) \geqslant Q(\theta^{(m)}|\theta^{(m)})$. Prvo je izvedena donja granica log-izglednosti

$$\ln \mathcal{L}(\theta) \geqslant \ln \mathcal{L}(\theta^{(m)}) + Q(\theta|\theta^{(m)}) - Q(\theta^{(m)}|\theta^{(m)}).$$

Primijetimo kako je u donjoj granici $Q(\theta|\theta^{(m)})$ jedini izraz koji ovisi o $\theta$. Sada ako vrijedi $Q(\theta|\theta^{(m)}) \geqslant Q(\theta^{(m)}|\theta^{(m)})$ imamo

$$\ln \mathcal{L}(\theta) \geqslant \ln \mathcal{L}(\theta^{(m)}) + [Q(\theta|\theta^{(m)}) - Q(\theta^{(m)}|\theta^{(m)})] \geqslant \ln \mathcal{L}(\theta^{(m)}).$$

Monotonost EM algoritma garantira da se s iteracijama procjene parametara neće pogoršavati u smislu izglednosti. Međutim, sama monotonost ne garantira konvergenciju niza procjena $\theta^{(m)}$. Štoviše, ne postoji općeniti teorem o konvergenciji za EM algoritam - konvergencija niza $\theta^{(m)}$ ovisi o karakteristikama log-izglednosti i $Q$-funkcije te jednako tako o početnim vrijednostima $\theta^{(0)}$. Pod određenim uvjetima regularnosti, može se pokazati da niz procjena $\theta^{(m)}$ konvergira prema stacionarnoj točki od $\ln \mathcal{L}(\theta)$, a takva konvergencija je linearna ([10]).

**Napomena 3.4** *Linearna konvergencija podrazumijeva postojanje $M > 0$ i $0 < C < 1$ takvih da je $\|\theta^{(m+1)} - \theta^*\| \leqslant C\|\theta^{(m)} - \theta^*\|$ za svaki $m \geqslant M$, gdje je $\theta^*$ optimalna vrijednost za $\theta$.*

## 3.4 Konvergencija EM algoritma

Ako je niz monoton i omeđen, onda je konvergentan. Dakle, niz odozgo omeđenih vrijednosti funkcije izglednosti $(\mathcal{L}(\theta^{(m)}))_{m\in\mathbb{N}_0}$, zbog svojstva monotonosti pri iteracijama EM algoritma, konvergira prema nekoj vrijednosti $L^*$. Gotovo uvijek je $L^*$ stacionarna vrijednost, tj. vrijedi $L^* = \mathcal{L}(\theta^*)$ za neku vrijednost parametra $\theta^*$ za koju je $\frac{\partial \mathcal{L}(\theta)}{\partial \theta} = 0$ ili ekvivalentno $\frac{\partial \ln \mathcal{L}(\theta)}{\partial \theta} = 0$

($\theta^*$ je stacionarna točka). Štoviše, u mnogim praktičnim primjenama, $L^*$ će biti lokalni maksimum. U slučaju da niz vrijednosti parametara $(\theta^{(m)})_{m\in\mathbb{N}_0}$ proizveden iteracijama EM algoritma postane "zarobljen" u nekoj stacionarnoj točki $\theta^*$ koja nije ekstrem (u sedlastoj točki), malena perturbacija od $\theta$ dalje od sedlaste točke $\theta^*$ biti će dovoljna da EM algoritam ne konvergira prema njoj. Općenito, ako funkcija izglednosti $\mathcal{L}(\theta)$ ima više stacionarnih točaka, konvergencija niza proizvedenog iteracijama EM algoritma ovisi o izboru početne vrijednosti $\theta^{(0)}$ ([16]).

Definiramo $\mathcal{M}(\theta^{(m)}) = \underset{\theta\in\Theta}{argmax}\, Q(\theta|\theta^{(m)})$. $\mathcal{M}(\theta^{(m)})$ je skup svih vrijednosti $\theta$ koje maksimiziraju $Q(\theta|\theta^{(m)})$. Za GEM algoritam, skup $\mathcal{M}(\theta^{(m)})$ je definiran izborom $\theta^{(m)}$ za koje u M-koraku algoritma vrijedi $Q(\theta^{(m+1)}|\theta^{(m)}) \geqslant Q(\theta^{(m)}|\theta^{(m)})$. Prema [16] te izvorno [19] neka vrijede sljedeće pretpostavke:

1. $\Theta \subseteq \mathbb{R}^d$, gdje je $\Theta$ parametarski prostor
2. $\Theta_{\theta_0} = \{\theta \in \Theta\colon \mathcal{L}(\theta) \geqslant \mathcal{L}(\theta_0)\}$ je kompaktan skup za svaki $\mathcal{L}(\theta_0) > -\infty$

   **Napomena 3.5** *Podskup euklidskog prostora je kompaktan ako i samo ako je zatvoren i omeđen.*

3. $\mathcal{L}(\theta)$ je neprekidna na $\Theta$ i diferencijabilna na $Int\Theta$.

Ove uvjete nazivamo uvjetima regularnosti. Kao posljedicu uvjeta regularnosti imamo da je niz $(\mathcal{L}(\theta^{(m)}))_{m\in\mathbb{N}_0}$ omeđen odozgo za svaki $\theta_0 \in \Theta$, gdje pretpostavljamo da vrijedi $\mathcal{L}(\theta_0) > -\infty$. Primjer gdje pretpostavka 2. o kompaktnosti nije zadovoljena je mješavina normalnih distribucija kada se neka od komponenti modela svede na jednu točku iz skupa podataka, tj. $\mu_j = x_i$ za neku j-tu komponentu i opažanje $x_i$. Prisjetimo se, tada dolazimo do pojave singularnosti, odnosno funkcija izglednosti teži ka beskonačnosti

kada $\sigma_j^2 \to 0$. Tako u ovom primjeru, ako je $\theta_0 \in \Theta$ bilo koja točka takva da je $\mu_j = x_i$, onda $\Theta_{\theta_0}$ nije kompaktan.

Sada ćemo dati iskaz teorema o konvergenciji za GEM algoritam koji se može pronaći u [16] te izvorno u [19]. Teorem vrijedi i za EM algoritam kako je on sam poseban slučaj GEM algoritma.

**Teorem 3.6 (O konvergenciji generaliziranog EM algoritma)** *Neka je* $(\theta^{(m)})_{m\in\mathbb{N}_0}$ *niz parametara generiran GEM algoritmom tako da je* $\theta^{(m+1)} \in \mathcal{M}(\theta^{(m)})$ *i neka je skup stacionarnih točaka u* $Int\Theta$ *označen sa* $S$*. Pretpostavimo da vrijedi sljedeće:*

1. $\mathcal{M}(\theta^{(m)})$ *je zatvoren na komplementu od* $S$
2. $\mathcal{L}(\theta^{(m+1)}) > \mathcal{L}(\theta^{(m)})$ *za svaki* $\theta^{(m)} \notin S$

*Tada je svako gomilište od* $(\theta^{(m)})_{m\in\mathbb{N}_0}$ *stacionarna točka i* $\mathcal{L}(\theta^{(m)})$ *konvergira monotono prema* $L^* = \mathcal{L}(\theta^*)$ *za neku stacionarnu točku* $\theta^* \in S$*.*

Uvjet 2. iz Teorema 3.6 vrijedi i za EM algoritam. Uzmimo $\theta^{(m)} \notin S$. Tada je $\frac{\partial Q(\theta|\theta^{(m)})}{\partial\theta}\big|_{\theta=\theta^{(m)}} = \frac{\partial \ln \mathcal{L}(\theta^{(m)})}{\partial\theta} \neq 0$ jer $\theta^{(m)} \notin S$. Stoga $Q(\theta|\theta^{(m)})$ nije maksimizirana u $\theta = \theta^{(m)}$ pa je $Q(\theta^{(m+1)}|\theta^{(m)}) > Q(\theta^{(m)}|\theta^{(m)})$ što implicira $\mathcal{L}(\theta^{(m+1)}) > \mathcal{L}(\theta^{(m)})$. Za EM algoritam, u [19] je uočena dovoljnost uvjeta za zatvorenost od $\mathcal{M}(\theta^{(m)})$ taj da je $Q(\theta|\vartheta)$ neprekidna u $\theta$ i u $\vartheta$.

**Teorem 3.7 (O konvergenciji EM algoritma)** *Pretpostavimo da je* $Q(\theta|\vartheta)$ *neprekidna u* $\theta$ *i u* $\vartheta$*. Tada su sve točke gomilišta niza* $(\theta^{(m)})_{m\in\mathbb{N}_0}$ *generiranog EM algoritmom stacionarne točke od* $\mathcal{L}(\theta)$ *te* $(\mathcal{L}(\theta^{(m)}))_{m\in\mathbb{N}_0}$ *konvergira monotono prema* $L^* = \mathcal{L}(\theta^*)$ *za neku stacionarnu točku* $\theta^* \in S$*.*

Teorem 3.7 slijedi direktno iz Teorema 3.6.

Konvergencija niza vrijednosti funkcije izglednosti $(\mathcal{L}(\theta^{(m)}))_{m\in\mathbb{N}_0}$ prema nekoj vrijednosti $L^*$ ne implicira konvergenciju pripadnog niza parametara $(\theta^{(m)})_{m\in\mathbb{N}_0}$ prema točki $\theta^*$. No, prema [19], konvergencija od $(\theta^{(m)})_{m\in\mathbb{N}_0}$ nije važna koliko i konvergencija od $(\mathcal{L}(\theta^{(m)}))_{m\in\mathbb{N}_0}$ prema stacionarnoj vrijednosti, posebno u lokalni maksimum.

Definiramo $S(a) = \{\theta \in S\colon \mathcal{L}(\theta) = a\}$ kao podskup skupa stacionarnih točaka u $Int\Theta$ u kojima $\mathcal{L}(\theta)$ poprima vrijednost $a$.

**Teorem 3.8** *Neka je $(\theta^{(m)})_{m\in\mathbb{N}_0}$ niz parametara generiran GEM algoritmom koji zadovoljava uvjete 1. i 2. iz Teorema 3.6. Neka je $S(L^*) = \{\theta^*\}$ (tj. ne postoje dvije različite stacionarne točke s istom vrijednosti $L^*$), gdje je $L^*$ gomilište od $\mathcal{L}(\theta^{(m)})$. Tada $(\theta^{(m)})_{m\in\mathbb{N}_0}$ konvergira prema $\theta^*$.*

Teorem 3.8 proizlazi iz Teorema 3.6 jer je $S(L^*)$ jednotočkovni skup.

**Teorem 3.9** *Neka je $(\theta^{(m)})_{m\in\mathbb{N}_0}$ niz parametara generiran GEM algoritmom koji zadovoljava uvjete 1. i 2. iz Teorema 3.6. Ako vrijedi $\|\theta^{(m+1)}-\theta^{(m)}\| \to 0$ kada $m \to \infty$, tada se sve točke gomilišta niza $(\theta^{(m)})_{m\in\mathbb{N}_0}$ nalaze u povezanom i kompaktnom podskupu od $S(L^*)$. Posebno, u slučaju kada je $S(L^*)$ diskretan, tada $(\theta^{(m)})_{m\in\mathbb{N}_0}$ konvergira prema nekoj točki $\theta^*$ iz $S(L^*)$.*

Dokaz za Teorem 3.9 se može pronaći u [16] i [19].

Konvergencija niza $(\theta^{(m)})_{m\in\mathbb{N}_0}$ prema stacionarnoj točki se može dokazati i bez korištenja Teorema 3.6. Neka je $\mathcal{K}(L_0) = \{\theta \in \Theta\colon \mathcal{L}(\theta) = L_0\}$ za neku vrijednost $L_0$.

**Teorem 3.10** *Neka je* $(\theta^{(m)})_{m\in\mathbb{N}_0}$ *niz parametara generiran GEM algoritmom sa svojstvom* $\frac{\partial Q(\theta|\theta^{(m)})}{\partial\theta}\big|_{\theta=\theta^{(m+1)}} = 0$. *Pretpostavimo da je* $\frac{\partial Q(\theta|\vartheta)}{\partial\theta}$ *neprekidna u* $\theta$ *i u* $\vartheta$. *Tada* $(\theta^{(m)})_{m\in\mathbb{N}_0}$ *konvergira prema stacionarnoj točki* $\theta^*$ *takvoj da je* $\mathcal{L}(\theta^*) = L^*$ *gdje je* $L^*$ *gomilište niza* $(\mathcal{L}(\theta^{(m)}))_{m\in\mathbb{N}_0}$ *ako je ili* $\mathcal{K}(L^*) = \{\theta^*\}$ *ili* $\|\theta^{(m+1)} - \theta^{(m)}\| \to 0$ *kada* $m \to \infty$ *i* $\mathcal{K}(L^*)$ *je diskretan.*

**Dokaz.** Uvjeti regularnosti impliciraju da je $(\mathcal{L}(\theta^{(m)}))_{m\in\mathbb{N}_0}$ omeđen odozgo pa konvergira prema nekom $L^*$. U slučaju kada je $\mathcal{K}(L^*) = \{\theta^*\}$, $(\theta^{(m)})_{m\in\mathbb{N}_0}$ očito konvergira prema $\theta^*$. U slučaju kada je $\mathcal{K}(L^*)$ diskretan skup, ali nije jednotočkovan, uvjet $\|\theta^{(m+1)} - \theta^{(m)}\| \to 0$ kada $m \to \infty$ je dovoljan za konvergenciju od $\theta^{(m)}$ prema $\theta^*$ u $\mathcal{K}(L^*)$. Vrijedi $\frac{\partial \ln\mathcal{L}(\theta^*)}{\partial\theta} = \frac{\partial Q(\theta|\theta^*)}{\partial\theta}\big|_{\theta=\theta^*}$. Naime, prisjetimo se da smo imali $\ln\mathcal{L}(\theta) = Q(\theta|\theta^{(m)}) - H(\theta|\theta^{(m)})$ pa je $\frac{\partial \ln\mathcal{L}(\theta)}{\partial\theta} = \frac{Q(\theta|\theta^{(m)})}{\partial\theta} - \frac{H(\theta|\theta^{(m)})}{\partial\theta}$. Također, pokazali smo da je $H(\theta|\theta^{(m)}) \leqslant H(\theta^{(m)}|\theta^{(m)})$ za svaki $\theta \in \Theta$ što povlači $\frac{\partial H(\theta|\theta^{(m)})}{\partial\theta}\big|_{\theta=\theta^{(m)}} = 0$. Sada za $\theta^{(m)} = \theta^*$ slijedi $\frac{\partial \ln\mathcal{L}(\theta^*)}{\partial\theta} = \frac{\partial Q(\theta|\theta^*)}{\partial\theta}\big|_{\theta=\theta^*}$. Dalje,

$$\frac{\partial \ln\mathcal{L}(\theta^*)}{\partial\theta} = \frac{\partial Q(\theta;\theta^*)}{\partial\theta}_{\theta=\theta^*} \overset{neprekidnost}{=} \lim_{m\to\infty} \frac{\partial Q(\theta|\theta^{(m)})}{\partial\theta}\Big|_{\theta=\theta^{(m+1)}} \overset{pretpostavka}{=} 0,$$

čime je pokazano da je točka gomilišta $\theta^*$ stacionarna točka od $\mathcal{L}(\theta)$. ■

Treba naglasiti uvjet $\frac{\partial Q(\theta,\theta^{(m)})}{\partial\theta}\big|_{\theta=\theta^{(m+1)}} = 0$ iz Teorema 3.10 zadovoljava svaki niz $(\theta^{(m)})_{m\in\mathbb{N}_0}$ generiran EM algoritmom pod pretpostavkom da vrijede uvjeti regularnosti.

**Korolar 3.11** *Neka je* $\mathcal{L}(\theta)$ *unimodalna (ima točno jedan maksimum) na* $\Theta$ *te* $\theta^*$ *jedina stacionarna točka. Neka je* $\frac{\partial Q(\theta|\vartheta)}{\partial\theta}$ *neprekidna u* $\theta$ *i u* $\vartheta$. *Tada svaki niz* $(\theta^{(m)})_{m\in\mathbb{N}_0}$ *generiran EM algoritmom konvergira prema jedinstvenom* $\theta^*$ *koji maksimizira* $\mathcal{L}(\theta)$, *tj. konvergira prema jedinstvenom procjenitelju najveće izglednosti od* $\theta$.

Rezultati konvergencije EM algoritma koji su izvedeni u [19] vrijede za situacije gdje se niz procjena parametara dobiven EM iteracijama nalazi u

potpunosti u unutrašnjosti prostora parametara, tj. u $Int\Theta$. Međutim, kod nekih problema procjene parametara sa ograničenim prostorom parametara, parametarska vrijednost koja maksimizira log-izglednost može biti na granici prostora parametara. Tako neki elementi niza generiranog EM iteracijama mogu ležati na granici i ne ispunjavati uvjete za konvergenciju postavljene u ovom dijelu rada.

### 3.4.1 Stopa konvergencije

Vidjeli smo kako EM algoritam definira funkciju mapiranja $\theta \mapsto M(\theta)$ iz prostora parametara $\Theta$ u njega samoga tako da je svaka iteracija $\theta^{(m)} \to \theta^{(m+1)}$ definirana sa $\theta^{(m+1)} = M(\theta^{(m)})$ za $m = 0, 1, 2, ....$ Ako $(\theta^{(m)})_{m\in\mathbb{N}_0}$ konvergira prema nekom $\theta^*$ i $M(\theta)$ je neprekidna, tada je $\theta^*$ fiksna točka algoritma, tj. vrijedi $\theta^* = M(\theta^*)$. Taylorovim razvojem $\theta^{(m+1)} = M(\theta^{(m)})$ oko točke $\theta^{(m)} = \theta^*$ dobijemo

$$\theta^{(m+1)} - \theta^* \approx J(\theta^*)(\theta^{(m)} - \theta^*),$$

gdje je $J(\theta^*)$ $d\times d$ Jakobijeva matrica za $M(\theta) = (M_1(\theta), ..., M_d(\theta))$. Element $J_{ij}(\theta)$ Jakobijeve matrice na mjestu $(i, j)$ jednak je $J_{ij}(\theta) = \frac{\partial M_i(\theta)}{\partial \theta_j}$, gdje je $M_i(\theta)$ i-ta koordinata od $M(\theta)$ i $\theta_j$ j-ta koordinata od $\theta$. Dakle, u blizini od $\theta^*$ EM algoritam je linearna iteracija sa matričnom stopom $J(\theta^*)$ kako $J(\theta^*)$ obično nije nul-matrica. Iz tog razloga, za $J(\theta^*)$ se obično kaže da je matrična stopa konvergencije, ili jednostavnije, stopa konvergencije ([16]).
Za vektor $\theta$, mjera stope konvergencije je definirana sa

$$r = \lim_{m\to\infty} \frac{\|\theta^{(m+1)}-\theta^*\|}{\|\theta^{(m)}-\theta^*\|},$$

gdje je $\|\cdot\|$ euklidska norma u prostoru $\mathbb{R}^d$. Poznato je da pod određenim uvjetima regularnosti vrijedi

$$r = \lambda_{max} \equiv \text{najveća svojstvena vrijednost od } J(\theta^*).$$

U praksi, $r$ se najčešće računa kao

$$r = \lim_{m\to\infty} \frac{\|\theta^{(m+1)} - \theta^{(m)}\|}{\|\theta^{(m)} - \theta^{(m-1)}\|}.$$

Primijetimo da veća vrijednost od $r$ implicira sporiju konvergenciju. Definira se $s = 1 - r$ kao globalna stopa konvergencije. Tako je $s$ najmanja svojstvena vrijednost od $S = I_d - J(\theta^*)$, koja se može zvati i (matričnom) brzinom konvergencije ([15]).

# Poglavlje 4

# Zaključak

Gaussov model mješavine i EM algoritam pokazali su se kao vrijedan alat pri statističkom modeliranju. Sam Gaussov model mješavine omogućuje fleksibilnu reprezentaciju kompleksnih distribucija, dok EM algoritam efikasno daje procjenu parametara mješavine. Također, u radu smo vidjeli i određena poželjna svojstva EM algoritma poput monotonosti i konvergencije, ali i istaknuli njegovu osjetljivost na inicijalizaciju parametara te se osvrnuli na problem određivanja komponenti mješavine. Unatoč određenim ograničenjima, Gaussov model mješavine i EM algoritam mogu se pronaći u širokoj primjeni u raznim područjima znanosti.

# Literatura


[1] Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In Selected papers of hirotugu akaike (pp. 199-213). New York, NY: Springer New York.

[2] Baudry, J. P., Celeux, G. (2015). EM for mixtures: Initialization requires special care. Statistics and computing, 25, 713-726.

[3] Biernacki, C., Celeux, G., Govaert, G. (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. Computational Statistics & Data Analysis, 41(3-4), 561-575.

[4] Bishop, C. M., Nasrabadi, N. M. (2006). Pattern recognition and machine learning (Vol. 4, No. 4, p. 738). New York: springer.

[5] Braić, S., Uvod u vjerojatnost i statistiku, 2011.

[6] Burnham, K. P., Anderson, D. R. (2001). Kullback-Leibler information as a basis for strong inference in ecological studies. Wildlife research, 28(2), 111-119.

[7] Dempster, A. P., Laird, N. M., Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society: series B (methodological), 39(1), 1-22.

**Literatura**


[8] Elezović, N., Matematička statistika, Stohastički procesi, Element, Zagreb, 2007.

[9] Elezović, N., Slučajne varijable, Element, Zagreb, 2007.

[10] Gupta, M. R., Chen, Y. (2011). Theory and use of the EM algorithm. Foundations and Trends in Signal Processing, 4(3), 223-296.

[11] Lindsay, B. G. (1995). Mixture models: theory, geometry, and applications. Ims.

[12] McLachlan, G. J., Lee, S. X., Rathnayake, S. I. (2019). Finite mixture models. Annual review of statistics and its application, 6, 355-378.

[13] Peel, D., MacLahlan, G. (2000). Finite mixture models. John & Sons.

[14] McLachlan, G. J., Rathnayake, S. (2014). On the number of components in a Gaussian mixture model. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 4(5), 341-355.

[15] Meng, X. L. (1994). On the rate of convergence of the ECM algorithm. The Annals of Statistics, 326-339.

[16] Ng, S. K., Krishnan, T., McLachlan, G. J. (2012). The EM algorithm. Handbook of computational statistics: concepts and methods, 139-172.

[17] Pearson, K. (1894). Contributions to the mathematical theory of evolution. Philosophical Transactions of the Royal Society of London. A, 185, 71-110.

[18] Sarapa, N., Teorija vjerojatnosti, Školska knjiga, Zagreb, 2002.

[19] Wu, C. J. (1983). On the convergence properties of the EM algorithm. The Annals of statistics, 95-103.