Tartu Ülikool
Loodus- ja täppisteaduste valdkond
Matemaatika ja statistika instituut

Eric Jakobson

Statistiline mudeli valik Akaike informatsioonikriteeriumi ja suurimate vahemike meetodi abil

Matemaatiline statistika

Bakalaureusetöö (9 EAP)

Juhendaja: PhD Kristi Kuljus

Tartu 2024

STATISTILINE MUDELI VALIK
AKAIKE INFORMATSIOONIKRITEERIUMI JA
SUURIMATE VAHEMIKE MEETODI ABIL

Bakalaureusetöö

Eric Jakobson

Lühikokkuvõte

Bakalaureusetöö eesmärk on tutvuda Kullback-Leibleri informatsioonimõõdu lähendamisel põhineva Akaike informatsioonikriteeriumiga ning uurida, millist täiendavat informatsiooni annab mudelite valideerimise kontekstis suurimate vahemike meetod. Töös vaadeldakse Kullback-Leibleri informatsiooni ja selle seost suurima tõepära meetodiga, käsitletakse Akaike informatsioonikriteeriumit ning sobitatakse andmetele erineva komponentide arvuga normaaljaotuste segujaotusi eesmärgiga tuvastada sobiv komponentide arv. Andmestikule segujaotuste sobitamisel selgus, et suurimate vahemike meetod andis selgema tulemuse sobivaima komponentide arvu valikul kui Akaike informatsioonikriteerium.

CERCS teaduseriala: P160 Statistika, operatsioonianalüüs, programmeerimine, finants- ja kindlustusmatemaatika.

Märksõnad: Kullback-Leibleri informatsioon, Akaike informatsioonikriteerium, suurimate vahemike meetod, normaaljaotuste segud, mudelite valideerimine.

STATISTICAL MODEL SELECTION USING THE AKAIKE INFORMATION CRITERION AND THE MAXIMUM SPACING METHOD

Bachelor thesis

Eric Jakobson

Abstract

The aim of this bachelor thesis is to learn about the Akaike information criterion which is based on the approximation of the Kullback-Leibler information measure, and to study what additional information the maximum spacing method provides in the context of model validation. The thesis examines the Kullback-Leibler information and its relation to the maximum likelihood method, discusses the Akaike information criterion, and fits Gaussian mixture models with different number of components to the data in order to identify a suitable number of components. Fitting mixture distributions to the data revealed that the maximum spacing method provided more clear results regarding the selection of a suitable number of components than the Akaike information criterion.

CERCS research specialisation: P160 Statistics, operations research, programming, financial and actuarial mathematics.

Key Words: Kullback-Leibler information, Akaike information criterion, maximum spacing method, Gaussian mixture models, model validation.

Sisukord

Sissejuhatus

Statistilise modelleerimise ülesande oluliseks osaks on mudeli valik. Sageli kasutatakse parima mudeli valimiseks erinevaid informatsioonikriteeriume ja nende modifikatsioone. Informatsioonikriteeriumi abil saab hinnata vaatluse all olevaid mudeleid, järjestades need väärtuste suuruse järgi. Probleem seisneb aga selles, et informatsioonikriteeriumite väärtused on suhtelised, mistõttu võib ainult nende suuruse põhjal olla raske aru saada, kui lähedal me oleme heale mudelile. On võimalik, et kui eeldatav mudelite klass ei sobi, siis on kõik meie vaadeldavad mudelid halvad ja parima valimine nende seast tähendab tegelikult parima halva mudeli valimist. Suurimate vahemike meetod võimaldab saada täiendavat infot vaatluse all olevate mudelite kohta, sest teoreetiliselt on teada suurimate vahemike funktsiooni käitumine piisavalt suure arvu vaatluste korral. Seega teame ka, milline võiks olla ligikaudu suurimate vahemike funktsiooni väärtus andmetega sobiva mudeli korral.

Käesoleva bakalaureusetöö eesmärk on tutvuda Kullback-Leibleri informatsioonimõõdu lähendamisel põhineva Akaike informatsioonikriteeriumiga ning uurida andmetele mudelite sobitamise abil, millist täiendavat informatsiooni on lisaks Akaike informatsioonikriteeriumile uuritavate mudelite kohta võimalik saada suurimate vahemike meetodi abil.

Töö koosneb neljast peatükist. Esimeses peatükis esitatakse Kullback-Leibleri informatsiooni definitsioon, omadused ning näitlikustatakse selle käitumist normaaljaotuste korral. Teises peatükis näidatakse, kuidas jõutakse Kullback-Leibleri informatsioonimõõdu minimeerimisel suurima tõepära meetodini. Kolmandas peatükis esitatakse Akaike informatsioonikriteeriumi definitsioon ning näidatakse, kuidas selleni jõutakse Kullback-Leibleri informatsiooni kaudu, siin on eriline rõhk Akaike informatsioonikriteeriumi parandusliikmel. Viimases peatükis sobitatakse andmetele erineva komponentide arvuga normaaljaotuste segujaotusi eesmärgiga tuvastada sobiv komponentide arv.

1 Kullback-Leibleri informatsioon

Peatükk põhineb raamatul *Information Criteria and Statistical Modeling* (Konishi ja Kitagawa, 2008, lk. 29–32).

Erinevad informatsioonikriteeriumid, mida kasutatakse sobitatud statistiliste mudelite võrdlemiseks ja nende headuse hindamiseks, põhinevad Kullback-Leibleri informatsioonimõõdul. Käesolevas peatükis esitame Kullback-Leibleri informatsiooni definitsiooni, omadused ning vaatleme selle käitumist normaaljaotuste korral.

1.1 Definitsioon

Olgu $\mathbf{x} = \{x_1, ..., x_n\}$ sõltumatute vaatluste realisatsioonid, mis on pärit tundmatust jaotusest jaotusfunktsiooniga $G(x)$. Tähistagu $G(x)$ järgnevalt tegelikule jaotusele vastavat jaotusfunktsiooni ning $F(x)$ mistahes muud jaotusfunktsiooni. Pidevate jaotuste korral on $g(x)$ ning $f(x)$ neile vastavad tihedusfunktsioonid. Diskreetsete jaotuste puhul on $g(x_i)$ ning $f(x_i)$ neile vastavad tõenäosusfunktsioonid:

$$g(x_i) = P_g(X = x_i),$$
$$f(x_i) = P_f(X = x_i),$$

kus $\{x_i : i = 1, 2, \dots\}$ on juhusliku suuruse X väärtuste hulk.

Kullback-Leibleri informatsioon (K-L informatsioon) võimaldab hinnata jaotuste vahelist kaugust ja on defineeritud järgnevalt:

$$I(g; f) = E_g \left[\ln \left(\frac{g(X)}{f(X)} \right) \right],$$

kus E_g tähistab keskväärtust jaotuse g suhtes.

Seega mõõdab K-L informatsioon, kui lähedal on jaotus $f(x)$ tegelikule jaotusele $g(x)$. Et K-L informatsiooni puhul on tegemist keskväärtusega, saab pidevate ja

diskreetsete jaotuste korral selle kirja panna järgmisel kujul:

$$I(g;f) = \begin{cases} \displaystyle\int_{-\infty}^{\infty} \ln\left(\frac{g(x)}{f(x)}\right) g(x)dx, & \text{pidevate jaotuste puhul,} \\ \displaystyle\sum_{i=1}^{\infty} g(x_i) \ln\left(\frac{g(x_i)}{f(x_i)}\right), & \text{diskreetsete jaotuste puhul.} \end{cases}$$

Siin $g(x)$ ja $f(x)$ tähistavad pidevatele jaotusfunktsioonidele vastavaid tihedusfunktsioone, diskreetsete jaotuste korral on arvestatud, et jaotusfunktsioonidele vastavad tõenäosusfunktsioonid on antud kujul $\{g(x_i);\ i = 1, 2, \dots\}$ ja $\{f(x_i);\ i = 1, 2, \dots\}$.

1.2 Kullback-Leibleri informatsiooni omadused

Kullback-Leibleri informatsioonil on järgmised omadused:

1) $I(g;f) \geq 0$,

2) $I(g;f) = 0 \quad \Leftrightarrow \quad g(x) = f(x)$.

Teeme läbi tõestuse pidevate jaotuste korral. Defineerime funktsiooni $K(t) := \ln t - t + 1$. Antud funktsioon on defineeritud $\forall\ t > 0$ korral. Funktsiooni tuletis on $K'(t) = t^{-1} - 1$. Kuna $K'(1) = 0$, siis funktsioon $K(t)$ saavutab oma maksimumi punktis $t = 1$. Seega $\forall\ t > 0$ korral kehtib võrratus $K(t) \leq 0$ ehk $\ln t \leq t - 1$. Asendades $t = \frac{f(x)}{g(x)}$ võrratusse saame omakorda, et

$$\ln \frac{f(x)}{g(x)} \leq \frac{f(x)}{g(x)} - 1.$$

Korrutades võrratuse mõlemat poolt funktsiooniga $g(x)$ ja seejärel integreerides üle x väärtuste saame

$$\int \ln\left(\frac{f(x)}{g(x)}\right) g(x)dx \leq \int f(x)dx - \int g(x)dx = 0.$$

Sellest järeldub, et

$$\int \ln\left(\frac{g(x)}{f(x)}\right) g(x)dx = -\int \ln\left(\frac{f(x)}{g(x)}\right) g(x)dx \geq 0.$$

Seega oleme näidanud, et kehtib 1). Omadus 2) kehtib juhul, kui $g(x) = f(x)$ peaaegu kindlasti. Diskreetsete jaotuste puhul piisab tõestuses kasutada $g(x)$ ja $f(x)$ asemel vastavalt tõenäosusfunktsioone $g(x_i)$ ja $f(x_i)$ ning integreerimise asemel summeerida üle $i = 1, 2, \ldots$. Tõestatud omadustest 1) ja 2) lähtuvalt on näha, et mida väiksem on K-L informatsioonimõõt, seda lähemal on jaotus $f(x)$ tegelikule jaotusele $g(x)$.

On oluline omadusena ka märkida, et Kullback-Leibleri informatsioon on asümmeetriline:

3) $I(g; f) \neq I(f; g)$.

Suurust $I(g; f)$ nimetatakse vahel kirjanduses ka Kullback-Leibleri kauguseks, kuid tegemist ei ole siiski kaugusmõõduga, sest antud suurus ei vasta kõigile kauguse aksioomidele. Kuigi $I(g; f) = 0$ parajasti siis, kui $g(x) = f(x)$, siis ei ole I sümmeetriline ning samuti ei kehti ka kolmnurgavõrratus. K-L informatsiooni puhul on tegemist statistilise kaugusmõõduga, mis mõõdab, kui erinevad kaks jaotust on.

1.3 Näited Kullback-Leibleri informatsioonist normaaljaotuste korral

Illustreerimaks Kullback-Leibleri informatsioonimõõdu käitumist, vaatleme selle käitumist normaaljaotuste korral. Valime tegelikuks jaotuseks $N(\nu, \kappa^2)$ tihedusega $g(x)$ ning vaatleme mingit teist normaaljaotust $N(\mu, \sigma^2)$ tihedusega $f(x)$. Normaaljaotuse $N(\mu, \sigma^2)$ tihedusfunktsiooniks on $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. Leiame suuruse $I(g; f)$. Et

$$I(g; f) = E_g\left[\ln g(X)\right] - E_g\left[\ln f(X)\right],$$

leiame alguses suuruse $E_g[\ln f(X)]$:

$$E_g[\ln f(X)] = E_g\left[-\frac{1}{2}\ln(2\pi\sigma^2) - \frac{(X-\mu)^2}{2\sigma^2}\right] = -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}E_g\left[(X-\mu)^2\right].$$

Eelneva suuruse leidmiseks on vaja teada suurust $E_g\left[(X-\mu)^2\right]$. Selle saame arvutada järgmiselt:

$$\begin{aligned} E_g\left[(X-\mu)^2\right] &= E_g\left[(X-\nu+\nu-\mu)^2\right] \\ &= E_g\left[(X-\nu)^2 + 2(X-\nu)(\nu-\mu) + (\nu-\mu)^2\right] \\ &= \kappa^2 + (\nu-\mu)E_g\left[2(X-\nu)\right] + (\nu-\mu)^2 = \kappa^2 + (\nu-\mu)^2. \end{aligned}$$

Seega

$$E_g[\ln f(X)] = -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{\kappa^2 + (\nu-\mu)^2}{2\sigma^2}. \tag{1}$$

Suuruse $E_g[\ln g(X)]$ leidmiseks piisab asendada $\mu = \nu$ ja $\sigma^2 = \kappa^2$ eelnevasse avaldisse ehk

$$E_g[\ln g(X)] = -\frac{1}{2}\ln(2\pi\kappa^2) - \frac{1}{2}.$$

Kokkuvõttes avaldub K-L informatsioon tegeliku jaotuse $g(x)$ ja mingi muu normaaljaotuse $f(x)$ vahel järgmiselt:

$$I(g;f) = \frac{1}{2}\left(\ln\frac{\sigma^2}{\kappa^2} + \frac{\kappa^2 + (\nu-\mu)^2}{\sigma^2} - 1\right). \tag{2}$$

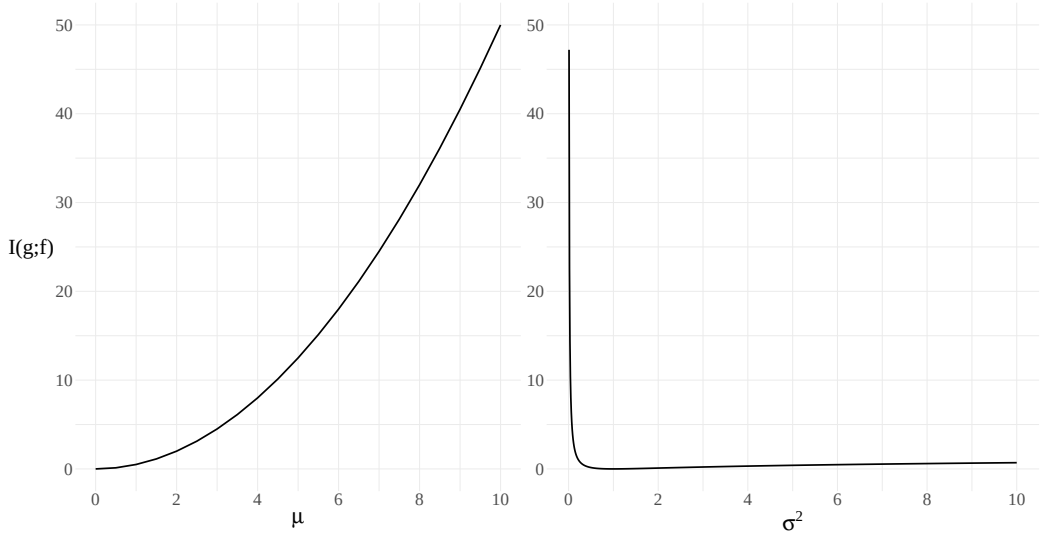
Järgnevalt vaatleme K-L informatsioonimõõtu $I(g;f)$, kui fikseerime jaotuse $g(x)$ ning varieerime jaotust $f(x)$ muutes vastavalt parameetreid μ või σ^2. Samuti leiame asümmeetrilisuse näitlikustamiseks suuruse $I(f;g)$, st vahetame f ja g rollid. Seda teeme ainult teisel juhul, kui muudame parameetrit σ^2, sest μ muutmisel tulevad informatsioonimõõdud $I(g;f)$ ja $I(f;g)$ võrdsed.

Fikseerime jaotuseks $g(x)$ normaaljaotuse $N(0,1)$. Kuna $\nu = 0$, siis $\sigma^2 = \kappa^2$ korral $I(g;f)$ on kujul $\frac{\mu^2}{2\sigma^2}$ ning seega näiteks $\mu = 1$ ja $\mu = -1$ puhul on K-L informatsioonimõõt sama, sest mõlema väärtuse korral oleme tegelikust keskväärtusest

sama kaugel.

Tabel 1: Näiteid K-L informatsioonimõõdust erinevate normaaljaotuse parameetrite korral, kui g on $N(0,1)$ tihedus ja f vastab tabeli veergudes toodud normaaljaotuste tihedustele.

$f : N(\mu, 1)$		$f : N(0, \sigma^2)$		
μ	$I(g;f)$	σ^2	$I(g;f)$	$I(f;g)$
0	0	0,01	47,197	1,808
0,5	0,125	0,1	3,349	0,701
1	0,5	0,4	0,292	0,158
1,5	1,125	0,7	0,036	0,028
2	2	1	0	0
2,5	3,125	1,3	0,016	0,019
3	4,5	1,6	0,048	0,065
3,5	6,125	1,9	0,084	0,129



Joonis 1: Kullback-Leibleri informatsioonimõõt erinevate normaaljaotuse parameetrite korral, vasakpoolsel joonisel tähistab f jaotuse $N(\mu, 1)$ tihedust, parempoolsel jaotuse $N(0, \sigma^2)$ tihedust, g on $N(0, 1)$ tihedus.

Tabelist 1 on näha, et fikseerides $\sigma^2 = 1$ ja suurendades μ väärtuseid, suureneb K-L informatsioonimõõt $I(g; f)$ samuti. Jooniselt 1 näeme, et kasvamine ei ole lineaarne. Samast tabelist on ka näha, et fikseerides $\mu = 0$ ja suurendades σ^2 väärtuseid, siis väga väikese σ^2 väärtuse korral on tulemuseks suure väärtusega K-L informatsioonimõõt $I(g; f)$. Selline seos tuleneb avaldisest (2), sest fikseeritud normaaljaotuse $N(0, 1)$ ja parameetri $\mu = 0$ korral on Kullback-Leibleri informatsioon kujul $\frac{1}{2}\left(\ln \sigma^2 + \frac{1}{\sigma^2} - 1\right)$, millest on näha, et väga väikeste σ^2 väärtuste korral on tulemuseks suur K-L informatsioonimõõt. Parameetri σ^2 suurendamine kuni väärtuseni 1 põhjustab K-L informatsioonimõõdu kahanemise ning pärast seda hakkab mõõt aeglaselt suurenema. Samuti on näha, et Kullback-Leibleri informatsioonimõõt ei ole sümmeetriline.

Toodud näited illustreerivad seda, et K-L informatsioonimõõdu suuruse kohta on raske anda mingit konkreetset hinnangut. Teame vaid, et mida väiksem on Kullback-Leibleri informatsioonimõõt, seda lähemal on valitud mudel $f(x)$ tegelikule mudelile $g(x)$.

2 Seos Kullback-Leibleri informatsioonimõõdu ja suurima tõepära meetodi vahel

Peatükk põhineb raamatul *Information Criteria and Statistical Modeling* (Konishi ja Kitagawa, 2008, lk. 35–37).

Käesolevas peatükis näitame, milline seos on Kullback-Leibleri informatsioonimõõdu ja suurima tõepära meetodi vahel ehk teisisõnu, kuidas jõuame K-L informatsioonimõõdu minimeerimisel suurima tõepära meetodini.

Olgu $\mathbf{X} = \{X_1, ..., X_n\}$ sõltumatud juhuslikud suurused tundmatust jaotusest tihedusega $g(x)$ ning olgu vaatlused $\mathbf{x} = \{x_1, ..., x_n\}$ nende realisatsioonid. Olgu X samuti juhuslik suurus jaotusest tihedusega $g(x)$. Tahame vaatlustele sobitada sobiva jaotuse. Vaatleme parameetrilist mudelit $\{f_\theta : \theta \in \Theta\}$, kus θ tähistab kõiki mudeli parameetreid ja Θ tähistab parameetrite ruumi, st kõikvõimalike parameetrite väärtuste hulka. Näiteks normaaljaotuse $N(\mu, \sigma^2)$ tiheduse f_θ parameetrite $\theta = (\mu, \sigma^2)$ puhul on parameetrite hulgaks $\Theta = \mathbb{R} \times (0, \infty)$. Sobitame vaatlustele jaotuste klassi, kust tahame leida sobivaima esindaja. Selleks hindame parameetrit θ selliselt, et $I(g; f_\theta)$ oleks võimalikult väike.

Kuna K-L informatsioon sisaldab tundmatut jaotust g, siis ei saa praktikas seda otse välja arvutada. Kullback-Leibleri informatsioon esitub kujul

$$I(g; f_\theta) = E_g \left[\ln \left\{ \frac{g(X)}{f_\theta(X)} \right\} \right] = E_g \left[\ln g(X) \right] - E_g \left[\ln f_\theta(X) \right],$$

kus võrduse parema poole esimene liige on konstant, mis sõltub ainult tegelikust jaotusest g. Seetõttu piisab $I(g; f_\theta)$ minimeerimiseks vaadelda vaid teist liiget $E_g \left[\ln f_\theta(X) \right]$. Seega, mida suurem on $E_g \left[\ln f_\theta(X) \right]$, seda väiksem on K-L informatsioonimõõt ja seda parem valitud tihedus $f_\theta(x)$ on.

Juhusliku suuruse keskväärtust on loomulik lähendada valimi keskmisega, sest suurte arvude seaduse põhjal koondub valimi keskmine tõenäosuse järgi juhusliku suu-

ruse keskväärtuseks:

$$E_g\left[\ln f_\theta(X)\right] \approx \frac{1}{n}\sum_{i=1}^{n} \ln f_\theta(x_i) \quad \text{ehk} \quad nE_g\left[\ln f_\theta(X)\right] \approx \sum_{i=1}^{n} \ln f_\theta(x_i).$$

Suurte arvude seaduse kohaselt kehtib koondumine

$$\frac{1}{n}\sum_{i=1}^{n} \ln f_\theta(X_i) \xrightarrow{p} E_g\left[\ln f_\theta(X)\right], \quad \text{kui } n \to \infty.$$

Lähendasime $E_g\left[\ln f_\theta(X)\right]$ valimi keskmisega. Seega parameeter θ, mis minimeerib meie Kullback-Leibleri informatsiooni lähendi, maksimeerib avaldise $\frac{1}{n}\sum_{i=1}^{n} \ln f_\theta(x_i)$, mis tähendab ühtlasi $\sum_{i=1}^{n} \ln f_\theta(x_i)$ ehk vaatluste $x_1, \ldots x_n$ log-tõepära maksimeerimist mudeli $\{f_\theta : \theta \in \Theta\}$ korral. Log-tõepära maksimeerimine vastab aga parameetri θ hinnangu leidmisele suurima tõepära meetodil. Tähistades

$$l(\theta; \mathbf{x}) = \sum_{i=1}^{n} \ln f_\theta(x_i),$$

saame suurima tõepära meetodi abil leida parameetri $\hat{\theta}$, mis annab suurima log-tõepära:

$$l(\hat{\theta}; \mathbf{x}) = \sum_{i=1}^{n} \ln f_{\hat{\theta}}(x_i), \text{ kus } \quad l(\hat{\theta}; \mathbf{x}) = \max_{\theta \in \Theta} l(\theta; \mathbf{x}).$$

Seeläbi saamegi leida vähima Kullback-Leibleri informatsioonimõõdu ning selle abil valida sobitatud jaotuste klassist sobivaima esindaja.

3 Akaike informatsioonikriteerium

Peatükk põhineb raamatul *Information Criteria and Statistical Modeling* (Konishi ja Kitagawa, 2008, lk. 51–64).

Järgnevalt kirjeldame, kuidas on defineeritud Akaike informatsioonikriteerium ja kuidas selleni jõutakse Kullback-Leibleri informatsiooni kaudu.

Akaike informatsioonikriteerium (AIC) on maksimaalse logaritmilise tõepära $l(\hat{\theta}; \mathbf{x})$ kaudu defineeritud kui

$$\text{AIC} = -2\,(\text{maksimaalne log-tõepära}) + 2\,(\text{hinnatavate parameetrite arv}).$$

Hinnatavate parameetrite arv tähistab hinnatava mudeli $\{f_\theta : \theta \in \Theta\}$ hinnatavate parameetrite arvu.

Olgu meil vaatluse all mitu erinevat mudelite klassi: $\{f_{1,\theta_1} : \theta_1 \in \Theta_1\}, \{f_{2,\theta_2} : \theta_2 \in \Theta_2\}, \ldots, \{f_{m,\theta_m} : \theta_m \in \Theta_m\}$. Eesmärgiks on sobitatud mudeleid omavahel võrrelda. Hindame suurima tõepära meetodi abil iga mudeli jaoks parameetrid $\theta_j,\ j = 1, \ldots, m$, olgu need hinnangud $\hat{\theta}_1, \ldots, \hat{\theta}_m$. Seega meil on m erinevat mudelit, mille seast tahame valida parima. Selle jaoks leiame iga mudeli korral Kullback-Leibleri informatsiooni ehk arvutame välja $I(g; f_{1,\hat{\theta}_1}), \ldots, I(g; f_{m,\hat{\theta}_m})$ ning valime nendest kõige väiksema. Tähistame suurima tõepära meetodi abil leitud mudelite tiheduste hinnanguid $f_{1,\hat{\theta}_1}, \ldots, f_{m,\hat{\theta}_m}$. Tahame leida nende seast sobivaima mudeli.

Tähistame praegu üldiselt mingi mudeli $\{f_\theta\}$ suurima tõepära meetodi abil leitud hinnangu $f_{\hat{\theta}}$. Meie eesmärgiks on hinnata, kui hea või halb hinnatud mudel on. Käesolevas peatükis käsitleme mudeli headust kui selle võimet prognoosida. Seega tähistades tuleviku vaatlust kui $Z = z$, mis on pärit tundmatust tegelikust jaotusest g, on meie ülesandeks hinnata, kui hästi kirjeldab hinnatud mudel $f_{\hat{\theta}}$ tegelikku jaotust g puntkis $Z = z$. Hea valitud mudeli puhul peaks g ja $f_{\hat{\theta}}$ vaheline erinevus olema võimalikult väike. Kahe jaotuse läheduse kontrollimiseks saab leida K-L

informatsiooni:

$$I(g; f_{\hat{\theta}}) = E_g\left[\ln\left\{\frac{g(Z)}{f_{\hat{\theta}}(Z)}\right\}\right] = E_g\left[\ln g(Z)\right] - E_g\left[\ln f_{\hat{\theta}}(Z)\right],$$

kus võrduse parema poole esimene liige on konstant, mis sõltub ainult tegelikust jaotusest g. Seetõttu piisab $I(g; f_{\hat{\theta}})$ minimeerimiseks vaadelda vaid teist liiget $E_g\left[\ln f_{\hat{\theta}}(Z)\right]$. Peatüki 2 põhjal on $E_g\left[\ln f_{\hat{\theta}}(Z)\right]$ loomulikuks lähendiks

$$\frac{1}{n}\, l(\hat{\theta}; \mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \ln f_{\hat{\theta}}(x_i).$$

Hinnatud mudelite võrdlemiseks tahame võrrelda suurusi $I(g; f_{1,\hat{\theta}_1}), \ldots, I(g; f_{m,\hat{\theta}_m})$. Seega on oluline leida keskväärtustele $E_g\left[\ln f_{i,\hat{\theta}_i}(Z)\right]$, $i = 1, \ldots, m$ hea hinnang. Kui kasutaksime $nE_g\left[\ln f_{i,\hat{\theta}_i}(Z)\right]$ hinnanguna suurust $l_i(\hat{\theta}_i; \mathbf{x})$, saaksime nihkega hinnangu, seejuures nihke suurus sõltub parameetervektori $\hat{\theta}_i$ dimensioonist.

3.1 Log-tõepära $l(\hat{\theta}; \mathbf{x})$ kui keskväärtuse $nE_g\left[\ln f_{\hat{\theta}}(Z)\right]$ nihe

Olgu θ_0 parameetri θ väärtus, mis minimeerib K-L informatsiooni g ja f_θ vahel ehk teisisõnu maksimeerib log-tõepära keskväärtuse $E_g\left[\ln f_\theta(Z)\right]$. Maksimeerigu $\hat{\theta}$ log-tõepära funktsiooni $l(\theta; \mathbf{x})$. Mudeli $f_{\hat{\theta}}$ headust Kullback-Leibleri informatsiooni abil tuleks hinnata log-tõepära keskväärtuse $E_g\left[\ln f_{\hat{\theta}}(Z)\right]$ hinnangu kaudu. Keskväärtuste puhul kehtib alati seos $E_g\left[\ln f_{\hat{\theta}}(Z)\right] \leq E_g\left[\ln f_{\theta_0}(Z)\right]$, sest θ_0 maksimeerib keskväärtuse $E_g\left[\ln f_\theta(Z)\right]$. Log-tõepärade korral kehtib aga seos $l(\hat{\theta}; \mathbf{x}) \geq l(\theta_0; \mathbf{x})$, sest $\hat{\theta}$ maksimeerib log-tõepära funktsiooni $l(\theta; \mathbf{x})$. Seega kui kasutame $nE_g\left[\ln f_{\hat{\theta}}(Z)\right]$ hindamisel suurust $l(\hat{\theta}; \mathbf{x})$, saame ülehinnangu.

Kui $\mathbf{x} = \{x_1, ..., x_n\}$ on sõltumatud vaatlused jaotusest tihedusega $g(x)$, siis log-tõepära funktsiooni kui log-tõepära keskväärtuse hinnangu nihe on defineeritud

järgmiselt:

$$b(g) = E_{g(\mathbf{x})}\left[l(\hat{\theta};\mathbf{X}) - nE_g\left[\ln f_{\hat{\theta}}(Z)\right]\right], \text{ kus } \quad g(\mathbf{x}) = \prod_{i=1}^{n} g(x_i)$$

tähistab valimi ühisjaotust.

Informatsioonikriteeriumi saab üldisel kujul koos nihke parandusliikmega kirja panna kujul

$$\begin{aligned}\mathrm{IC}(\mathbf{X};\hat{g}) &= -2(\text{mudeli log-tõepära} - \text{nihke hinnang}) \\ &= -2l(\hat{\theta};\mathbf{x}) + 2\{b(g)\text{ hinnang}\}.\end{aligned}$$

Akaike informatsioonikriteerumi puhul on nihke $b(g)$ hinnanguks hinnatavate parameetrite arv. Olgu hinnatavate parameetrite arv p. Defineerime $p \times p$ maatriksid

$$I(\theta_0) = E_g\left[\frac{\partial \ln f_\theta(Z)}{\partial \theta}\frac{\partial \ln f_\theta(Z)}{\partial \theta^T}\bigg|_{\theta_0}\right], \; J(\theta_0) = -E_g\left[\frac{\partial^2 \ln f_\theta(Z)}{\partial \theta \partial \theta^T}\bigg|_{\theta_0}\right]. \quad (3)$$

Saab näidata, et suurte valimite korral avaldub nihe $b(g)$ järgmiselt:

$$b(g) = \mathrm{tr}\{I(\theta_0)J(\theta_0)^{-1}\}. \quad (4)$$

Kui tegelik mudel g kuulub valitud mudeli klassi $\{f_\theta\}$, st. $\exists\theta_0$ nii, et $g = f_{\theta_0}$, siis kehtib seos $I(\theta_0) = J(\theta_0)$ ja seega avaldub nihe $b(g)$ hinnatavate parameetrite arvu kaudu:

$$\mathrm{tr}\{I(\theta_0)J(\theta_0)^{-1}\} = \mathrm{tr}\{I_p\} = p. \quad (5)$$

Seeläbi jõuamegi **Akaike informatsioonikriteeriumi** kujuni:

$$\mathrm{AIC} = -2l(\hat{\theta};\mathbf{x}) + 2p = -2\sum_{i=1}^{n} \ln f_{\hat{\theta}}(x_i) + 2p.$$

Väikeste valimite korral on Akaike informatsioonikriteerium kujul

$$\mathrm{AIC}_v = -2\sum_{i=1}^{n} \ln f_{\hat{\theta}}(x_i) + 2p + \frac{2p(p+1)}{(n-p-1)}, \tag{6}$$

kus n tähistab valimimahtu. Informatsioonikriteeriumit AIC_v on soovitatav kasutada kui $n/p < 40$. (Burnham ja Anderson, 1998, lk. 322)

3.2 Nihke leidmine normaaljaotuse sobitamise korral

Eesmärk on näidata, et Akaike informatsioonikriteeriumi parandusliige annab üsna hea hinnangu nihkele $b(g)$, kui sobitatav mudel on piisavalt lähedal tegelikule mudelile.

Olgu $\mathbf{X} = \{X_1, ..., X_n\}$ sõltumatud juhuslikud suurused jaotusest tihedusega $g(x)$ ning olgu vaatlused $\mathbf{x} = \{x_1, ..., x_n\}$ nende realisatsioonid. Olgu X juhuslik suurus jaotusest tihedusega $g(x)$. Olgu sobitatud mudeliks normaaljaotuste klass $N(\mu, \sigma^2)$ tihedusfunkstiooniga $f_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, kus $\theta = (\mu, \sigma^2)$. Defineerime **Takeuchi informatsioonikriteeriumi**

$$\mathrm{TIC} = -2\sum_{i=1}^{n} \ln f_{\hat{\theta}}(x_i) + 2\mathrm{tr}(\hat{I}\hat{J}^{-1}),$$

kus $\hat{I}$ ja $\hat{J}$ on avaldises (3) defineeritud maatriksite $I(\theta_0)$ ning $J(\theta_0)$ mõjusad hinnangud.

Normaaljaotuse korral saab nihke $b(g)$ asümptootilise avaldise $\mathrm{tr}\{I(\theta_0)J(\theta_0)^{-1}\}$ kergesti välja arvutada. Tähistame mudeli parameetrid $\theta_0 = (\mu_g, \sigma_g^2)$ mis maksimeerivad keskväärtuse $E_g[\ln f_\theta(x)]$. Keskväärtus $E_g[\ln f_\theta(x)]$ on avaldise (1) põhjal kujul

$$E_g[\ln f_\theta(x)] = -\frac{1}{2}\ln\left(2\pi\sigma^2\right) - \frac{\sigma_g^2}{2\sigma^2} - \frac{(\mu-\mu_g)^2}{2\sigma^2},$$

kus μ_g ja σ_g^2 on tegeliku jaotuse $g(x)$ keskväärtus ja dispersioon. Veendume, et

parameetrid θ_0 maksimeerivad selle keskväärtuse:

$$\frac{\partial}{\partial\mu}\left(-\frac{1}{2}\ln\left(2\pi\sigma^2\right)-\frac{\sigma_g^2}{2\sigma^2}-\frac{(\mu-\mu_g)^2}{2\sigma^2}\right)=\frac{\mu-\mu_g}{\sigma^2}=0\Rightarrow\mu=\mu_g,$$
$$\frac{\partial}{\partial\sigma^2}\left(-\frac{1}{2}\ln\left(2\pi\sigma^2\right)-\frac{\sigma_g^2}{2\sigma^2}-\frac{(\mu-\mu_g)^2}{2\sigma^2}\right)=\frac{-\sigma^2+\sigma_g^2+(\mu-\mu_g)^2}{2\sigma^4}=0\Rightarrow\sigma^2=\sigma_g^2.$$

Osatuletised log-tõepara funktsioonist $\ln f_\theta(x)$ parameetrite μ ja σ^2 suhtes on

$$\frac{\partial}{\partial\mu}\ln f_\theta(x)=\frac{x-\mu}{\sigma^2},\quad \frac{\partial}{\partial\sigma^2}\ln f_\theta(x)=-\frac{1}{2\sigma^2}+\frac{(x-\mu)^2}{2\sigma^4},$$
$$\frac{\partial^2}{\partial\mu^2}\ln f_\theta(x)=-\frac{1}{\sigma^2},\quad \frac{\partial^2}{\partial\mu\partial\sigma^2}\ln f_\theta(x)=-\frac{x-\mu}{\sigma^4},$$
$$\frac{\partial^2}{(\partial\sigma^2)^2}\ln f_\theta(x)=\frac{1}{2\sigma^4}-\frac{(x-\mu)^2}{\sigma^6}.$$

Kasutades avaldisi (3) ja leitud osatuletisi, saame leida maatriksid $I(\theta_0)$ ja $J(\theta_0)$:

$$\begin{aligned}I(\theta_0)&=E_g\left[\begin{pmatrix}\frac{X-\mu_g}{\sigma_g^2}\\ -\frac{1}{2\sigma_g^2}+\frac{(X-\mu_g)^2}{2\sigma_g^4}\end{pmatrix}\left(\frac{X-\mu_g}{\sigma_g^2},-\frac{1}{2\sigma_g^2}+\frac{(X-\mu_g)^2}{2\sigma_g^4}\right)\right]\\&=E_g\begin{bmatrix}\frac{(X-\mu_g)^2}{\sigma_g^4} & -\frac{(X-\mu_g)}{2\sigma_g^4}+\frac{(X-\mu_g)^3}{2\sigma_g^6}\\ -\frac{(X-\mu_g)}{2\sigma_g^4}+\frac{(X-\mu_g)^3}{2\sigma_g^6} & \frac{1}{4\sigma_g^4}-\frac{(X-\mu_g)^2}{2\sigma_g^6}+\frac{(X-\mu_g)^4}{4\sigma_g^8}\end{bmatrix}\\&=\begin{bmatrix}\frac{\sigma_g^2}{\sigma_g^4} & \frac{\mu_3}{2\sigma_g^6}\\ \frac{\mu_3}{2\sigma_g^6} & -\frac{\sigma_g^2}{4\sigma_g^6}+\frac{\mu_4}{4\sigma_g^8}\end{bmatrix}=\begin{bmatrix}\frac{1}{\sigma_g^2} & \frac{\mu_3}{2\sigma_g^6}\\ \frac{\mu_3}{2\sigma_g^6} & \frac{\mu_4}{4\sigma_g^8}-\frac{1}{4\sigma_g^4}\end{bmatrix},\end{aligned}$$

kus $\mu_j=E_g\left[(X-\mu_g)^j\right]$ tähistab tegeliku jaotuse $g(x)$ j-ndat tsentraalset momenti.

$$J(\theta_0)=-\begin{bmatrix}-\frac{1}{\sigma_g^2} & -\frac{E_g[(X-\mu_g)]}{\sigma_g^4}\\ -\frac{E_g[(X-\mu_g)]}{\sigma_g^4} & \frac{1}{2\sigma_g^4}-\frac{E_g[(X-\mu_g)^2]}{\sigma_g^6}\end{bmatrix}=\begin{bmatrix}\frac{1}{\sigma_g^2} & 0\\ 0 & \frac{1}{\sigma_g^4}-\frac{1}{2\sigma_g^4}\end{bmatrix}=\begin{bmatrix}\frac{1}{\sigma_g^2} & 0\\ 0 & \frac{1}{2\sigma_g^4}\end{bmatrix}.$$

Leitud suuruste põhjal saame avaldise (4) abil välja arvutada nihke $b(g)$

$$\mathrm{tr}\{I(\theta_0)J(\theta_0)^{-1}\} = 1 + \frac{\mu_4}{2\sigma_g^4} - \frac{1}{2} = \frac{1}{2}\left(1 + \frac{\mu_4}{\sigma_g^4}\right).$$

Seega on loomulik valimi põhjal hinnata suurust $\mathrm{tr}\{I(\theta_0)J(\theta_0)^{-1}\}$ järgmiselt:

$$\mathrm{tr}(\hat{I}\hat{J}^{-1}) = \frac{1}{2} + \frac{\hat{\mu}_4}{2\hat{\sigma}^4},$$

kus $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^2$ ja $\hat{\mu}_4 = \frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^4$. Seega Takeuchi informatsioonikriteerium on antud juhul kujul

$$\mathrm{TIC} = -2\sum_{i=1}^n \ln f_{\hat{\theta}}(x_i) + 2\left(\frac{1}{2} + \frac{\hat{\mu}_4}{2\hat{\sigma}^4}\right).$$

Kui $\exists\theta_0$ nii, et $g = f_{\theta_0}$, siis $g(x)$ on normaaljaotusega ning seega $\mu_3 = 0$ ja $\mu_4 = 3\sigma_g^4$. Sellisel juhul saame nihkeks

$$\frac{1}{2} + \frac{\mu_4}{2\sigma_g^4} = \frac{1}{2} + \frac{3\sigma_g^4}{2\sigma_g^4} = \frac{1}{2} + \frac{3}{2} = 2,$$

mis on antud juhul hinnatavate parameetrite arv ning seega oleme jõudnud juba varasemalt leitud Akaike informatsioonikriteeriumini

$$\mathrm{AIC} = -2\sum_{i=1}^n \ln f_{\hat{\theta}}(x_i) + 2 \cdot 2.$$

Eeldame nüüd, et tegelik jaotus $g(x)$ on kahe normaaljaotuse segu

$$g(x) = (1-\pi)f_{\theta_1}(x) + \pi f_{\theta_2}(x), \quad 0 \le \pi \le 1,$$

kus $f_{\theta_i}(x)$ $(i = 1, 2)$ tähistab normaaljaotuse tihedusfunktsiooni parameetritega $\theta_i = (\mu_i, \sigma_i^2)$ ning π tähistab segukaalu.

Järgnevalt uurime TIC parandusliikme $\frac{1}{2}\left(1 + \frac{\hat{\mu}_4}{\hat{\sigma}^4}\right)$ käitumist erinevate segujao-

tuste ja valimimahtude korral. Vaadeldud on kolme erinevat segujaotuse juhtu, kus komponentide keskväärtused on võrdsed, kuid teise komponendi standardhälve on erinev. Järgnevates tabelites on toodud 10 000 simulatsiooni parandusliikme keskmised ja standarhälbed erinevate segukaalude π ning vaatluste arvu n puhul. Sobitatud mudeliks on normaaljaotuste klass $N(\mu, \sigma^2)$.

Tabel 2: TIC parandusliikme keskmine ja standardhälve kahe normaaljaotuse segu korral, kui $\mu_1 = \mu_2 = 0, \sigma_1 = 1, \sigma_2 = \sqrt{2}$.

π	$n = 25$	$n = 100$	$n = 400$	$n = 1600$
0	1,89 (0,37)	1,97 (0,23)	1,99 (0,12)	2,00 (0,06)
0,1	1,94 (0,43)	2,06 (0,32)	2,10 (0,18)	2,11 (0,09)
0,4	1,99 (0,45)	2,13 (0,32)	2,17 (0,18)	2,18 (0,09)
0,7	1,95 (0,41)	2,07 (0,27)	2,10 (0,15)	2,11 (0,07)
1	1,88 (0,36)	1,97 (0,23)	1,99 (0,12)	2,00 (0,06)

Tabelist 2 on näha, et $n = 25$ korral, kui $\pi = 0$ või $\pi = 1$, siis parandusliige on märgatavalt väiksem kui 2 ehk AIC parandusliige, mis viitab sellele, et vähese vaatluste arvu korral me potentsiaalselt alahindame nihet. Antud juhul ei tule TIC parandusliikme väärtused väga suured, sest teise komponendi standardhälve on väike.

Tabel 3: TIC parandusliikme keskmine ja standardhälve kahe normaaljaotuse segu korral, kui $\mu_1 = \mu_2 = 0, \sigma_1 = 1, \sigma_2 = 2$.

π	$n = 25$	$n = 100$	$n = 400$	$n = 1600$
0	1,88 (0,37)	1,97 (0,23)	1,99 (0,12)	2,00 (0,06)
0,1	2,15 (0,68)	2,52 (0,72)	2,66 (0,47)	2,70 (0,25)
0,4	2,26 (0,62)	2,56 (0,50)	2,64 (0,27)	2,66 (0,14)
0,7	2,07 (0,46)	2,24 (0,33)	2,28 (0,18)	2,29 (0,09)
1	1,88 (0,37)	1,97 (0,23)	1,99 (0,12)	2,00 (0,06)

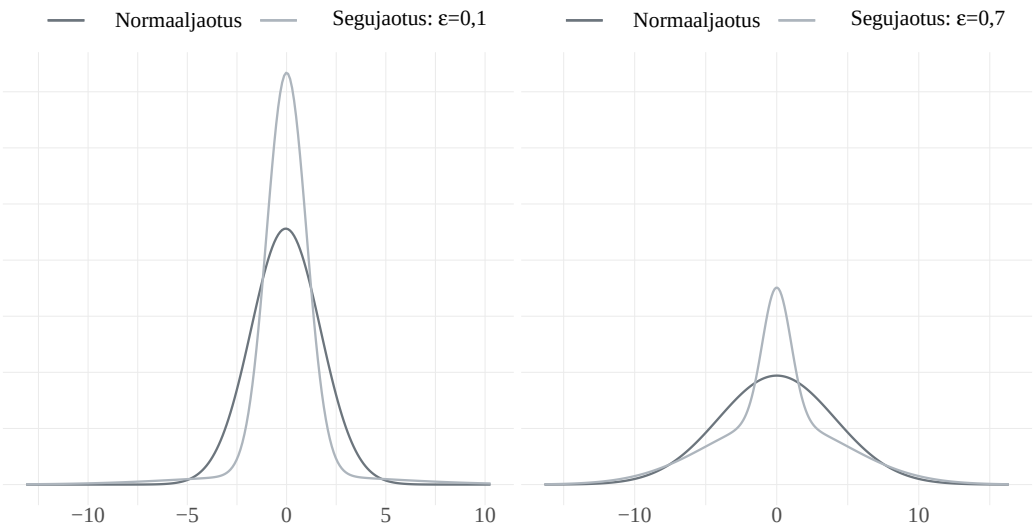
Tabelist 3 on näha, et $\pi = 0{,}1, \pi = 0{,}4$ ja $\pi = 0{,}7$ korral on TIC parandusliikme väärtused suuremad kui tabelis 2, sest segujaotus ja sobitatud normaaljaotus erinevad üksteisest rohkem, kuna segujaotuse teise komponendi standardhälve on suurem.

Tabel 4: TIC parandusliikme keskmine ja standardhälve kahe normaaljaotuse segu korral, kui $\mu_1 = \mu_2 = 0, \sigma_1 = 1, \sigma_2 = 5$.

π	$n = 25$	$n = 100$	$n = 400$	$n = 1600$
0	1,88 (0,37)	1,97 (0,23)	1,99 (0,12)	2,00 (0,06)
0,1	3,86 (1,90)	6,84 (2,78)	8,19 (1,99)	8,58 (1,14)
0,4	3,14 (0,98)	3,66 (0,79)	3,80 (0,45)	3,83 (0,23)
0,7	2,32 (0,58)	2,51 (0,38)	2,56 (0,21)	2,57 (0,10)
1	1,88 (0,36)	1,97 (0,23)	1,99 (0,12)	2,00 (0,06)

Tabelist 4 on näha, et $\pi = 0{,}1$ puhul on TIC parandusliige kõikidest esitatud π väärtustest suurim, sest sellisel juhul on jaotus raskemate sabadega, mille tõttu on normaaljaotust raskem sobitada. Kaalu $\pi = 0{,}7$ puhul on suurema standardhälbega $\sigma_2 = 5$ komponendil suurem mõju ning seega on jaotus rohkem sarnane suurema standardhälbega normaaljaotusele. Samuti on näha, et suuremate TIC parandusliikme väärtuste korral on standardhälve samuti suurem.

Eelnevatest tabelitest nägime, et piisavalt suure valimimahu korral normaaljaotuse puhul (juhud $\pi = 0$ ja $\pi = 1$) tuli parandusliige väga lähedale tegelikule nihkele (parameetrite arvule) 2. Samuti nägime, et valimimahu kasvades vähenes ka parandusliikme standardhälve.



Joonis 2: Võrdlus valimi ($n = 1600$) põhjal hinnatud parameetritega normaaljaotuse ja segujaotuse ($\mu_1 = \mu_2 = 0, \sigma_1 = 1, \sigma_2 = 5$) korral, kui kaalud on vastavalt $\varepsilon = 0{,}1$ ja $\varepsilon = 0{,}7$.

Jooniselt 2 on näha, et $\pi = 0{,}1$ korral on segujaotusel rasked sabad, mis ei ole omane normaaljaotusele ning seetõttu võib olla segujaotusele normaaljaotust raskem sobitada kui juhul $\pi = 0{,}7$.

Kokkuvõttes nägime, et kui meie sobitatud mudel on piisavalt lähedal tegelikule mudelile, siis Akaike informatsioonikriteeriumi parandusliige tuleb üsna lähedale tegelikule nihkele $b(g)$.

4 Normaaljaotuste segu komponentide arvu tuvastamine

Käesolevas peatükis sobitame andmestikule erineva komponentide arvuga normaaljaotuste segujaotusi eesmärgiga tuvastada sobiv komponentide arv Akaike informatsioonikriteeriumi ja suurimate vahemike meetodi abil.

Kasutatud andmestikuks on 1872. aasta Mehhiko postmarkide paksused. Andmestik sisaldab 485 kasutatud vesimärgita margi paksust millimeetrites. Mitmes varasemas artiklis on juba käsitletud sobivaima komponentide arvu küsimust antud andmestiku korral. Izenman ja Sommer (1988) leidsid tõepärasuhte testi abil, et vähimaks sobivaks komponentide arvuks võiks olla kolm. Basford, McLachlan ja York (1997) näitasid, et määrates segukomponentide hajuvused võrdseks, on tõepärasuhte testi põhjal sobivaimaks komponentide arvuks seitse. Izenman ja Sommer (1988) uurisid markide paksuse jaotuse multimodaalsust ka mitteparameetriliste tuumatiheduste hinnangute abil ja jõudsid järeldusele, et sobiv normaaljaotuste segu komponentide arv võiks olla seitse (McLachlan ja Peel, 2004, lk. 179).

Sobivaima komponentide arvu tuvastamiseks peame kõigepealt andmetele sobitama erinevate komponentide arvuga normaaljaotuste segusid. Segujaotuste parameetrite hindamise viime läbi **EM-algoritmi** abil.

4.1 EM-algoritm normaaljaotuste segu parameetrite hindamiseks

Alapeatükk põhineb raamatul *The Elements of Statistical Learning* (Hastie, Tibshirani ja Friedman, 2009, lk. 272–275).

Olgu $f(x)$ juhusliku suuruse X tihedusfunktsioon, mis on K komponendiga nor-

maaljaotuste segu:

$$f(x) = \sum_{k=1}^{K} \pi_k f_{\theta_k}(x), \quad 0 \leq \pi_k \leq 1, \quad \sum_k \pi_k = 1,$$

kus $f_{\theta_k}(x)(k = 1, \ldots, K)$ tähistab k-nda komponendi normaaljaotuse tihedusfunktsiooni parameetritega $\theta_k = (\mu_k, \sigma_k^2)$ ning π_k tähistab k-nda komponendi segukaalu. Olgu Z latentne juhuslik suurus, mis võtab väärtuseid $1, \ldots, K$ tõenäosustega $P(Z = k) = \pi_k$ ning mis näitab, millisest komponendist vaatlus pärit on. Tinglik tõenäosus $P(Z = k|X = x)$ avaldub Bayes'i valemi järgi kujul

$$P(Z = k|X = x) = \frac{\pi_k f_{\theta_k}(x)}{\sum_{j=1}^{K} \pi_j f_{\theta_j}(x)} := \gamma_k.$$

Olgu meil vaatlused $X_1, \ldots, X_n$ ning neile vastavad latentsed juhuslikud suurused $Z_1, \ldots, Z_n$. Tahame andmetele $x_1, \ldots, x_n$ sobitada mitmest normaaljaotusest koosnevat segujaotust kasutades suurima tõepära meetodit. Hinnatavad parameetrid on $\theta = \left[\left(\pi_1, \mu_1, \sigma_1^2\right), \ldots, \left(\pi_K, \mu_K, \sigma_K^2\right)\right]$, seejuures $\sum_k \pi_k = 1$ ehk hinnatavate parameetrite arv on $2K + (K - 1)$. Log-tõepära on kujul

$$l(\theta; x_1, \ldots, x_n) = \sum_{i=1}^{n} \ln \left[\sum_{k=1}^{K} \pi_k f_{\theta_k}(x_i) \right].$$

Leiame iga vaatluse x_i korral tõenäosuse, et see pärineb komponendist k. Seega tähistagu γ_{ik} tinglikku tõenäosust $P(Z_i = k|X_i = x_i)$:

$$\gamma_{ik} = \frac{\pi_k f_{\theta_k}(x_i)}{\sum_{j=1}^{K} \pi_j f_{\theta_j}(x_i)}, \quad k = 1, \ldots, K; \quad i = 1, \ldots, n.$$

Kui tahaksime leida funktsiooni $l(\theta; x_1, \ldots, x_n)$ maksimeerivad parameetrid nagu tavaliselt suurima tõepära meetodi korral (võtame log-tõepärast parameetrite suh-

tes tuletised ja võrdsustame nulliga), jõuaksime järgmiste avaldisteni:

$$\mu_k = \frac{\sum_{i=1}^n \gamma_{ik} x_i}{\sum_{i=1}^n \gamma_{ik}}, \quad \sigma_k^2 = \frac{\sum_{i=1}^n \gamma_{ik}(x_i - \mu_k)^2}{\sum_{i=1}^n \gamma_{ik}}, \quad \pi_k = \frac{\sum_{i=1}^n \gamma_{ik}}{n}. \tag{7}$$

Me ei saa aga neid parameetreid analüütiliselt välja arvutada, sest tinglikud tõenäosused γ_{ik} sõltuvad omakorda parameetritest μ_k ja $\sigma_k^2, k = 1, \ldots, K$. Avaldistest (7) on aga näha, et parameetrite hinnangute leidmiseks saame kasutada iteratiivset lähenemist järgnevalt kirjeldatud **EM-algoritmi** abil.

Valime esialgsed väärtused parameetritele, olgu need $\hat{\mu}_k, \hat{\sigma}_k^2, \hat{\pi}_k$, $k = 1, \ldots, K$. Rakendame korda-mööda järgmiseid samme.

E-samm: kasutame praeguseid parameetrite väärtuseid, et arvutada välja

$$\hat{\gamma}_{ik} = \frac{\hat{\pi}_k f_{\hat{\theta}_k}(x_i)}{\sum_{j=1}^K \hat{\pi}_j f_{\hat{\theta}_j}(x_i)}, \quad k = 1, \ldots, K; \quad i = 1, \ldots, n.$$

M-samm: eelmises sammus arvutatud $\hat{\gamma}_{ik}$ väärtuste abil hindame uuesti parameetrid

$$\hat{\mu}_k = \frac{\sum_{i=1}^n \hat{\gamma}_{ik} x_i}{\sum_{i=1}^n \hat{\gamma}_{ik}}, \quad \hat{\sigma}_k^2 = \frac{\sum_{i=1}^n \hat{\gamma}_{ik}(x_i - \hat{\mu}_k)^2}{\sum_{i=1}^n \hat{\gamma}_{ik}}, \quad \hat{\pi}_k = \frac{\sum_{i=1}^n \hat{\gamma}_{ik}}{n}, \; k = 1, \ldots, K.$$

Iga parameetrite uuendus E-sammu ja sellele järgneva M-sammu tulemusena garanteerib log-tõepära suurenemise. Kordame algoritmi E- ja M-samme seni kuni log-tõepära koondub soovitud täpsuseni või on tehtud etteantud arv iteratsioonisamme.

Kirjeldatud EM-algoritmi abil saame valitud komponentide arvu korral hinnata segujaotuse parameetrid ning leida hinnatud mudeli log-tõepära. Kasutades leitud log-tõepärasid, saame kõikide erinevate komponentide arvuga segujaotuste puhul välja arvutada Akaike informatsioonikriteeriumi. Järgnevas alapeatükis toome sisse **suurimate vahemike meetodi**.

4.2 Suurimate vahemike meetod

Alapeatükk põhineb artiklil *The Maximum Spacing Method. An Estimation Method Related to the Maximum Likelihood Method* (Ranneby, 1984, lk. 94–95).

Suurimate vahemike meetodit kutsutakse inglise keele eeskujul ka MSP-meetodiks, mis tuleneb väljendist *maximum spacing method*. Olgu $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ sõltumatud juhuslikud suurused tundmatust jaotusest tihedusega $g(x)$ ning olgu vaatlused $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ nende realisatsioonid. Olgu $\{f_\theta, \theta \in \Theta\}$ ning g sama kandjaga pidevate jaotuste tihedusfunktsioonid, tähistame neile vastavad jaotusfunktsioonid kui $F_\theta(x)$ ja $G(x)$. Kasutades kahe järjestikuse vaatluse abil moodustatud jaotusfunktsiooni vahemikke, saame leida lähendi Kullback-Leibleri informatsioonile $I(g; f_\theta)$. Olgu

$$-\infty = x_{(0)} < x_{(1)} < \ldots < x_{(n)} < x_{(n+1)} = \infty$$

järjestatud vaatlused. Vaatleme vahemikke $F_\theta(x_{(j)}) - F_\theta(x_{(j-1)})$ ja $G(x_{(j)}) - G(x_{(j-1)})$, siis Lagrange'i keskväärtusteoreemi kohaselt

$$\begin{aligned} F_\theta(x_{(j)}) - F_\theta(x_{(j-1)}) &= (x_{(j)} - x_{(j-1)}) \cdot f_\theta(x'_j), \\ G(x_{(j)}) - G(x_{(j-1)}) &= (x_{(j)} - x_{(j-1)}) \cdot g(x''_j), \end{aligned}$$

kus x'_j ja x''_j asuvad $x_{(j-1)}$ ja $x_{(j)}$ vahel. Peatüki 2 põhjal teame, et $\frac{1}{n}\sum_{i=1}^{n} \ln\left(\frac{g(x_i)}{f_\theta(x_i)}\right)$ koondub Kullback-Leibleri informatsiooniks $I(g; f_\theta)$. Seetõttu on intuitiivselt selge, et ka

$$\frac{1}{n+1}\sum_{j=1}^{n+1} \ln \frac{G(x_{(j)}) - G(x_{(j-1)})}{F_\theta(x_{(j)}) - F_\theta(x_{(j-1)})} \tag{8}$$

koondub suuruseks $I(g; f_\theta)$. Avaldise (8) minimeerimine on võrdväärne avaldise

$$\frac{1}{n+1}\sum_{j=1}^{n+1} \ln\left(F_\theta(x_{(j)}) - F_\theta(x_{(j-1)})\right)$$

maksimeerimisega, mis omakorda on võrdväärne avaldise

$$S_n(\theta) = \frac{1}{n+1} \sum_{j=1}^{n+1} \ln \left[\left(F_\theta(x_{(j)}) - F_\theta(x_{(j-1)}) \right) (n+1) \right]$$

maksimeerimisega. Tähistame funktsiooni $S_n(\theta)$ maksimeeriva parameetri $\hat{\theta}_n$, nimetame seda **suurimate vahemike hinnanguks (MSP-hinnanguks)**.

Kui tegelik mudel kuulub sobitatud mudelite klassi, st $\exists \theta_0$ nii et $f_{\theta_0} = g$, siis juhuslik suurus $S_n(\theta_0)$ on asümptootiliselt normaaljaotusega, st kui $n \to \infty$, siis

$$\sqrt{n} S_n(\theta_0) \xrightarrow{d} N\left(-\gamma, \pi^2/6 - 1\right).$$

kus $\gamma \approx 0,577$ on Euleri konstant. Seega mudeli headust saab hinnata selle järgi, kui lähedal on MSP funktsiooni väärtus $S_n(\hat{\theta}_n)$ suurusele $-\gamma$.

MSP-meetod on alternatiivne meetod suurima tõepära meetodile parameetrite hindamiseks pidevate jaotuste korral. MSP-funktsiooni väärtuse arvutamine võimaldab meil hinnata erinevate komponentide arvuga segujaotuste mudeli sobivust antud andmete jaoks ja öelda, kui kaugel oleme tegelikust mudelist.

4.3 Sobiva komponentide arvu tuvastamine

Markide andmestikule sobitasime komponentide arvuga $k = 2, \ldots, 8$ normaaljaotuste segu, parameetrite suurima tõepära hinnangud $\hat{\theta}$ leidsime kasutades EM-algoritmi. Esimesel juhul ei teinud me ühtegi kitsendust EM-algoritmi poolt hinnatavate parameetrite kohta. Kuna algoritmi alglähendid valitakse juhuslikult, siis iga komponentide arvu puhul hindasime segujaotuse parameetreid 20 korda ning valisime parameetrid, mille puhul oli log-tõepära kõige suurem. EM-algoritmi teostamiseks kasutasime rakendustarkvara R versiooni 4.2.1 lisapaketi *mixtools* funktsiooni *normalmixEM* (Benaglia *et al.*, 2009). Realiseeritud kood ja hinnatud segujaotuse parameetrid on toodud lisas 1. Kasutades erinevate komponentide arvuga hinnatud segujaotuste log-tõepärasid, saame avaldisega (6) arvutada Akaike informatsioonikriteeriumi väärtused.

Tabel 5: Hinnatud segujaotuste log-tõepära, AIC_v ja $S_n(\hat{\theta})$ väärtused.

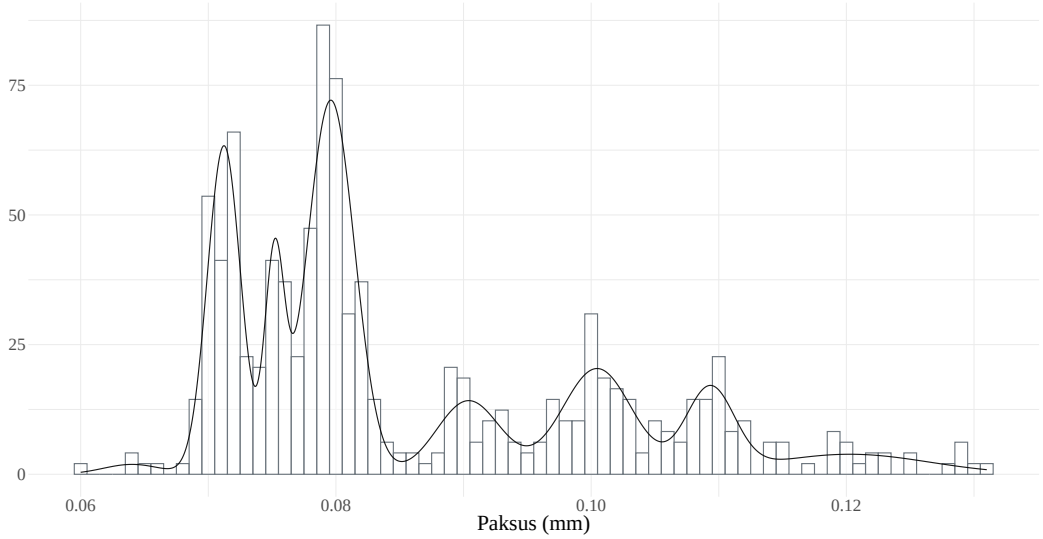
k	Log-tõepära	AIC_v	$S_n(\hat{\theta})$
2	1484,8	-2959,4	-0,702
3	1518,8	-3021,4	-0,638
4	1529,9	-3037,2	-0,621
5	1532,9	-3036,8	-0,617
6	1537,3	-3039,2	-0,611
7	1543,5	-3045,1	-0,593
8	1547,9	-3047,3	-0,585

Tabelist 5 on näha, et mida suurem on hinnatava mudeli komponentide arv, seda väiksem tuleb AIC_v väärtus, kõige väiksem on AIC_v väärtus $k = 8$ komponendiga normaaljaotuse segu korral.

Tabel 6: $k = 8$ komponendiga segujaotuse hinnatud parameetrid.

$\pi_1, \ldots, \pi_8$	0,01	0,20	0,08	0,34	0,09	0,14	0,08	0,06
$\mu_1, \ldots, \mu_8$	0,064	0,071	0,075	0,080	0,090	0,100	0,109	0,120
$\sigma_1, \ldots, \sigma_8$	0,0023	0,0013	0,0008	0,00186	0,0025	0,0028	0,0019	0,0063

Tabelist 6 näeme, et kolme suurima hinnatud kaaluga komponendi segukaalud annavad kokku suurema kaalu kui ülejäänud komponentide kaalud kokku. Samuti on näha, et kõige suurema hinnatud keskväärtusega komponendi standardhälve (parempoolses sabas) on samuti kõige suurem.



Joonis 3: Markide paksuse jaotus ja $k = 8$ komponendiga sobitatud segujaotuse tihedusfunktsioon.

Jooniselt 3 on näha, et markide paksuse jaotusel on mitu eristatavat tippu ning hinnatud segujaotuse kuue suurema segukaaluga komponendid järgivad jaotuse tippusid. Ülejäänud kaks komponenti on määratud vaatlustele, mis asuvad jaotuse sabades.

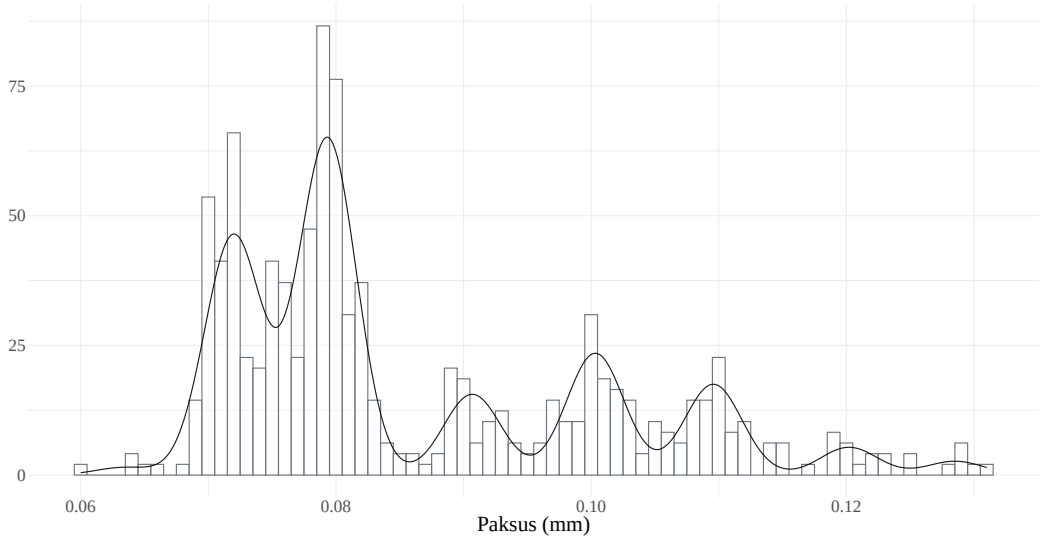
Järgnevalt teeme mudeli hindamisel kitsenduse, et komponentide standardhälbed on võrdsed. Selline kitsendus tundub mõistlik arvestades, et markide paksuse haju-

vus võiks tootmisprotsessis mistahes konkreetse paksusega paberi korral olla konstantne. Kuna võrdsete standardhälvete puhul on komponentide eristamine keerulisem, siis iga komponentide arvu puhul hindasime segujaotuse parameetreid EM-algoritmi abil 1000 korda ning valisime parameetrid, mille puhul oli log-tõepära kõige suurem. Realiseeritud kood ja hinnatud segujaotuse parameetrid on toodud lisas 2.

Tabel 7: Hinnatud segujaotuste log-tõepära, AIC_v ja $S_n(\hat{\theta})$ väärtused võrdsete standardhälvetega komponentide korral.

k	Log-tõepära	AIC_v	$S_n(\hat{\theta})$
2	1442,6	-2877,2	-0,823
3	1475,7	-2939,2	-0,753
4	1487,5	-2958,6	-0.730
5	1489,6	-2958,7	-0.727
6	1512,9	-3001,2	-0,688
7	1525,4	-3021,8	-0,660
8	1535,2	-3037,2	-0,632

Tabelist 7 näeme samuti, et mida suurem on hinnatud segujaotuse komponentide arv, seda väiksem tuleb AIC_v väärtus ning seega kõige väiksem on AIC_v väärtus $k = 8$ komponendiga normaaljaotuse segu korral. Samuti on näha, et $k = 5$ log-tõepära on ainult natukene suurem $k = 4$ log-tõepära väärtusest. See tuleneb sellest, et segujaotuse komponentide ühine standardhälve pole piisavalt väike, et $k = 5$ puhul lisandunud komponenti mõnele jaotuse tipule määrata. Võrreldes tabeli 7 AIC_v väärtusi tabelis 5 olevate AIC_v väärtustega, näeme, et ilma standardhälvete kitsenduseta sobitatud segujaotus annab kõikide komponentide puhul madalama AIC_v väärtuse kui võrdsete standardhälvetega sobitatud segujaotuste puhul.



Joonis 4: Markide paksuse jaotus ja $k = 8$ komponendiga sobitatud σ kitsendusega segujaotuse tihedusfunktsioon.

Jooniselt 4 on näha, et võrreldes joonisega 3 on üks komponent määratud jaotuse vasakul pool paikneva komponendi asemel hoopis jaotuse parempoolsesse sabasse. Ülejäänud komponendid on määratud samadele tippudele nagu kitsenduseta sobitatud segujaotuse puhul.

Lisaks AIC_v väärtustele arvutasime ka MSP funktsiooni väärtused suurima tõepära hinnangute korral, nende abil saame täiendavat informatsiooni erinevate komponentide arvuga segujaotuste andmetele sobivuse kohta. Meenutame, et MSP-meetod on parameetrite hindamise meetod pidevate jaotuste korral. Markide paksuste andmestik on aga eriline selle poolest, et selles esineb palju korduvaid vaatluseid. Andmestikus on ainult 62 unikaalset vaatlust ning paksused on mõõdetud millimeetrites kolmanda komakohani. Taoline eripära näib loomulik kuna mõõteskaala on väga väike ning tollel ajal ei olnud ilmselt võimalik mõõtmiseid täpsemalt teostada. Seega tundub samuti loomulik eeldada, et tegelik jaotus võiks olla normaaljaotuste segu, kuigi täpne komponentide arv on teadmata. Korduvate mõõtmiste probleemi lahendamiseks (et oleks võimalik välja arvutada suurus $S_n(\hat{\theta})$) lisame korduvatele vaatlustele müra ühtlasest jaotusest $U(-0,00049; 0,00049)$. See

garanteerib, et katame kogu jaotuse kandja ning vaatlused jäävad kuni kolmanda komakohani samaks. Kasutades järjestatud vaatluseid koos lisatud müraga, oleme välja arvutanud MSP-funktsiooni väärtused normaaljaotuste segude korral komponentide arvuga vastavalt $k = 2, \ldots, 8$ ilma standardhälvete kitsenduseta tabelis 5 ning võrdsete standardhälvete korral tabelis 7.

Meenutame, et kui tegelik mudel kuulub sobitatud klassi, st $\exists\theta_0$ nii et $f_{\theta_0} = g$, siis suure n korral $\sqrt{n}S_n(\theta_0) \approx N\left(-\gamma, \pi^2/6 - 1\right)$. Kuna EM-algoritmist saadud suurima tõepära hinnang $\hat{\theta}$ on asümptootiliselt samade omadustega nagu suurimate vahemike hinnang, siis saame $S_n(\hat{\theta})$ väärtuseid kasutada mudelite võrdlemiseks. Sobiva mudeli korral peaks $S_n(\hat{\theta})$ väärtus olema lähedal $-\gamma$ väärtusele.

$S_n(\hat{\theta})$ väärtused tabelis 5 ja tabelis 7 viitavad, et segujaotus $k = 8$ komponendiga ilma standardhälvete kitsenduseta on antud andmetele kõige sobivam ($S_n(\hat{\theta}) \approx -0,585$). Kuna kehtib

$$P\left(S_n(\hat{\theta}) \leq -\gamma - z_{1-\alpha}\sqrt{\frac{\pi^2/6-1}{n}}\right) \leq$$
$$P\left(S_n(\theta_0) \leq -\gamma - z_{1-\alpha}\sqrt{\frac{\pi^2/6-1}{n}}\right) \approx \alpha,$$

kus $z_{1-\alpha}$ tähistab standardse normaaljaotuse $(1-\alpha)$ kvantiili, siis saame arvesse võtta ka juhuslikkust. Seega $S_n(\hat{\theta})$ väärtused väiksemad kui $-\gamma - z_{1-\alpha}\sqrt{(\pi^2/6-1)/n}$ viitavad sellele, et mudel ei ole sobiv. Antud näites $n = 485$ ning seega $\alpha = 0,05$ korral $S_n(\hat{\theta})$ väärtused väiksemad kui $-0,6372$ näitavad, et hinnatud mudel ei ole sobiv. Selle kriteeriumi põhjal sobib võrdsete standardhälvete puhul andmete kirjeldamiseks ainult $k = 8$ komponendiga segujaotus. $S_n(\hat{\theta})$ väärtused tabelis 5 annavad tulemuseks, et segujaotused komponentide arvuga $k = 4, \ldots, 8$ võiks kõik olla vastuvõetavad mudelid.

Kokkuvõttes nägime, et Akaike informatsioonikriteeriumi puhul, mida suurem oli komponentide arv, seda väiksem oli ka AIC_v ning seega näis kõige paremini sobivat vaadeldud mudelitest kõige suurema komponentide arvuga segujaotus. MSP-

meetod andis aga selgema tulemuse, sest MSP-funktsiooni väärtuste põhjal nägime, et võrdsete standardhälvete puhul sobis andmete kirjeldamiseks vaid kaheksa komponendiga segujaotus ning ilma kitsenduseta juhul nelja kuni kaheksa komponendiga segujaotus.

Kokkuvõte

Töö eesmärk oli tutvuda Kullback-Leibleri informatsiooni mõistega, Akaike informatsioonikriteeriumiga ning uurida normaaljaotuste segujaotuste andmetele sobitamise abil, millist lisainformatsiooni lisaks Akaike informatsioonikriteeriumile saab uuritavate mudelite kohta suurimate vahemike meetodi abil.

Esmalt käsitleti Kullback-Leibleri informatsiooni ja selle käitumist normaaljaotuste korral. Kullback-Leibleri informatsiooni puhul toodi välja, et tegemist on statistilise kaugusmõõduga, mis mõõdab, kui erinevad kaks jaotust on. Kuna tegemist ei ole tavalise kaugusmõõduga, siis on selle väärtuste kohta raske anda mingit konkreetset hinnangut. Saab vaid väita, et mida väiksem on Kullback-Leibleri informatsioonimõõt, seda lähemal on valitud mudel tegelikule mudelile.

Kullback-Leibleri informatsiooni lähendamisel log-tõepäraga on mudeli hindamise kontekstis tulemuseks nihkega hinnang. Akaike informatsioonikriteeriumi puhul näidati, et suurte valimite korral avaldub nihe parameetrite arvu kaudu ning seega on Akaike informatsioonikriteeriumi parandusliikmeks kahekordne hinnatavate parameetrite arv. Samuti näitlikustati, et Akaike parandusliige annab üsna hea hinnangu nihkele, kui sobitatav mudel on piisavalt lähedal tegelikule mudelile.

Vaadeldud andmestikule erinevate komponentide arvuga normaaljaotuste segujaotuste sobitamise puhul oli näha, et mida suurem oli komponentide arv, seda väiksem oli ka Akaike informatsioonikriteeriumi väärtus, ning seega näis kõige paremini sobivat vaadeldud mudelitest kõige suurema komponentide arvuga segujaotus. MSP-meetod andis aga selgema tulemuse, sest MSP-funktsiooni väärtuste põhjal nägime konkreetsemalt, milliste komponentide arvu puhul ei olnud hinnatud mudel sobiv andmeid kirjeldama.

Akaike informatsioonikriteeriumi abil saame mudelid ainult järjestada, väärtuse suurust üksinda on raske tõlgendada. Suurimate vahemike meetodi abil saame aga täpsemalt anda hinnangu, milline mudel sobib andmetele ja milline mitte.

Kasutatud allikad

Basford, K. E., G. J. Mclachlan ja M. G. York (1997). "Modelling the distribution of stamp paper thickness via finite normal mixtures: The 1872 Hidalgo stamp issue of Mexico revisited". *Journal of Applied Statistics* 24.2, lk. 169–180. DOI: https://doi.org/10.1080/02664769723783.

Benaglia, T., D. Chauveau, D. R. Hunter ja D. S. Young (2009). "mixtools: An R Package for Analyzing Mixture Models". *Journal of Statistical Software* 32.6, lk. 1–29. DOI: https://doi.org/10.18637/jss.v032.i06.

Burnham, K.P. ja D.R. Anderson (1998). *Model Selection and Inference: A Practical Information-Theoretic Approach*. Springer, New York.

Hastie, T., R. Tibshirani ja J.H. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.

Izenman, A. J. ja C. J. Sommer (1988). "Philatelic Mixtures and Multimodal Densities". *Journal of the American Statistical Association* 83.404, lk. 941–953. DOI: https://doi.org/10.2307/2290118.

Konishi, S. ja G. Kitagawa (2008). *Information Criteria and Statistical Modeling*. Springer, New York.

McLachlan, G.J. ja D. Peel (2004). *Finite Mixture Models*. Wiley.

Ranneby, B. (1984). "The Maximum Spacing Method. An Estimation Method Related to the Maximum Likelihood Method". *Scandinavian Journal of Statistics* 11.2, lk. 93–112. URL: https://www.jstor.org/stable/4615946.

Lisa 1. Markide näide: EM-algoritmi kood ja hinnatud parameetrid ilma standardhälvete kitsenduseta

Tabel 8: Hinnatud segujaotuse kaalud.

k	π_1	π_2	π_3	π_4	π_5	π_6	π_7	π_8
2	0,61	0,39						
3	0,19	0,37	0,44					
4	0,20	0,08	0,28	0,44				
5	0,01	0,20	0,08	0,30	0,41			
6	0,01	0,20	0,10	0,24	0,04	0,41		
7	0,20	0,08	0,28	0,32	0,05	0,04	0,03	
8	0,01	0,20	0,08	0,34	0,09	0,14	0,08	0,06

Tabel 9: Hinnatud segujaotuse keskväärtused.

k	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_7	μ_8
2	0,077	0,102						
3	0,071	0,079	0,099					
4	0,071	0,075	0,080	0,099				
5	0,064	0,071	0,075	0,080	0,101			
6	0,064	0,071	0,075	0,079	0,082	0,101		
7	0,071	0,075	0,080	0,094	0,101	0,110	0,123	
8	0,064	0,071	0,075	0,080	0,090	0,100	0,109	0,120

Tabel 10: Hinnatud segujaotuse standardhälbed.

k	σ_1	σ_2	σ_3	σ_4	σ_5	σ_6	σ_7	σ_8
2	0,005	0,012						
3	0,013	0,002	0,01					
4	0,001	0,001	0,002	0,014				
5	0,002	0,001	0,001	0,002	0,012			
6	0,002	0,001	0,001	0,001	0,0005	0,012		
7	0,001	0,001	0,002	0,014	0,001	0,001	0,005	
8	0,002	0,001	0,001	0,002	0,003	0,003	0,002	

```
library(mixtools)
set.seed(100)

data(Stamp)

# Kasutame EM algoritmi et hinnata seguajaotuse parameetrid k=2 kuni k=8 komponendi korral
# Iga komponentide arvu jaoks arvutame AIC väärtuse
lambdas <- list()
mus <- list()
sigmas <- list()
loglik_values <- c()
AIC_values <- c()

# Sooritame iga komponentide arvu kohta 20 korda EM algoritmi ning valimi nende seast suurima log-tõ
    epära
# Ilma hajuvuse kitsenduseta
max_k <- 8
num_iterations <- 20

for (k in 2:max_k) {
  best_loglik <- -Inf
  best_parameters <- list()
  for (i in 1:num_iterations) {
    p = k*2 + (k - 1)
    fit_k <- normalmixEM(Stamp, k = k, maxit = 10000)
    AIC_k <- -2 * fit_k$loglik + 2 * p + (2*p*(p+1))/(length(Stamp)-p-1)

    # Kas on siiani suurim log-tõepära?
    if (fit_k$loglik > best_loglik) {
      best_loglik <- fit_k$loglik
      best_parameters <- list(
        lambda = fit_k$lambda,
        mu = fit_k$mu,
        sigma = fit_k$sigma,
        loglik = fit_k$loglik,
```

```
        AIC = AIC_k
      )
    }
  }
  # 20 EM-algoritmi suurima log-tõepära puhul saadud parameetrid
  lambdas[[k-1]] <- best_parameters$lambda
  mus[[k-1]] <- best_parameters$mu
  sigmas[[k-1]] <- best_parameters$sigma
  loglik_values <- c(loglik_values, best_parameters$loglik)
  AIC_values <- c(AIC_values, best_parameters$AIC)
}
```

Lisa 2. Markide näide: EM-algoritmi kood ja hinnatud parameetrid võrdsete standardhälvetega komponentide korral

Tabel 11: Hinnatud segujaotuse kaalud.

k	π_1	π_2	π_3	π_4	π_5	π_6	π_7	π_8
2	0,71	0,29						
3	0,68	0,26	0,06					
4	0,65	0,15	0,16	0,04				
5	0,64	0,09	0,13	0,10	0,04			
6	0,26	0,38	0,09	0,13	0,10	0,04		
7	0,26	0,38	0,09	0,13	0,10	0,03	0,01	
8	0,01	0,26	0,37	0,09	0,13	0,09	0,03	0,02

Tabel 12: Hinnatud segujaotuse keskväärtused.

k	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_7	μ_8
2	0,077	0,107						
3	0,077	0,102	0,121					
4	0,080	0,095	0,106	0,123				
5	0,076	0,091	0,101	0,110	0,124			
6	0,072	0,079	0,091	0,100	0,109	0,124		
7	0,072	0,080	0,090	0,100	0,110	0,120	0,129	
8	0,063	0,072	0,080	0,091	0,100	0,110	0,120	0,129

Tabel 13: Hinnatud segujaotuse standardhälbed.

k	σ_1	σ_2	σ_3	σ_4	σ_5	σ_6	σ_7	σ_8
2	0,007	0,007						
3	0,006	0,006	0,006					
4	0,005	0,005	0,005	0,005				
5	0,005	0,005	0,005	0,005	0,005			
6	0,003	0,003	0,003	0,003	0,003	0,003		
7	0,002	0,002	0,002	0,002	0,002	0,002	0,002	
8	0,002	0,002	0,002	0,002	0,002	0,002	0,002	0,002

```
library(mixtools)
set.seed(100)

data(Stamp)
setwd(dirname(rstudioapi::getActiveDocumentContext()$path))
load("mus.RData")
mus_for_EM <- mus
load("lambdas.RData")
lambdas_for_EM <- lambdas


# Kasutame EM algoritmi et hinnata seguajaotuse parameetrid k=2 kuni k=8 komponendi korral
# Iga komponentide arvu jaoks arvutame AIC väärtuse
lambdas <- list()
mus <- list()
sigmas <- list()
loglik_values <- c()
AIC_values <- c()

# Sooritame iga komponentide arvu kohta 1000 korda EM algoritmi ning valimi nende seast suurima log-t
    õepära
# Eeldame, et standardhälbed on võrdsed
max_k <- 8
num_iterations <- 1000

for (k in 2:max_k) {
  best_loglik <- -Inf
  best_parameters <- list()
  sd_constr <- rep("a", k)
  for (i in 1:num_iterations) {
    p = 2*k
    if (k == 4 | k==5){
      fit_k <- normalmixEM(Stamp, k = k, maxit = 10000, lambda = lambdas_for_EM[[k-1]], sd.constr =
          sd_constr)
    } else {
```

```
      fit_k <- normalmixEM(Stamp, k = k, maxit = 10000, lambda = lambdas_for_EM[[k-1]], mu = mus_for_
          EM[[k-1]], sd.constr = sd_constr)
    }
    AIC_k <- -2 * fit_k$loglik + 2 * p + (2*p*(p+1))/(length(Stamp)-p-1)

    # Kas on siiani parim log-tõepära?
    if (fit_k$loglik > best_loglik) {
      best_loglik <- fit_k$loglik
      best_parameters <- list(
        lambda = fit_k$lambda,
        mu = fit_k$mu,
        sigma = fit_k$sigma,
        loglik = fit_k$loglik,
        AIC = AIC_k
      )
    }
  }
  # 1000 EM-algoritmi suurima log-tõepära puhul saadud parameetrid
  lambdas[[k-1]] <- best_parameters$lambda
  mus[[k-1]] <- best_parameters$mu
  sigmas[[k-1]] <- best_parameters$sigma
  loglik_values <- c(loglik_values, best_parameters$loglik)
  AIC_values <- c(AIC_values, best_parameters$AIC)
}
```

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Eric Jakobson,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose "Statistiline mudeli valik Akaike informatsioonikriteeriumi ja suurimate vahemike meetodi abil", mille juhendaja on Kristi Kuljus, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 4.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaaloman-di ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Eric Jakobson

15.05.2024