TARTU ÜLIKOOL
MATEMAATIKA-INFORMAATIKATEADUSKOND
Matemaatilise statistika instituut

Kristjan Kokorev

# Erlangi jaotuste segude sobitamine kindlustuskahjudele

Magistritöö finants- ja kindlustusmatemaatika erialal (30 EAP)

Juhendaja: dotsent Meelis Käärik

TARTU
2015

## Erlangi jaotuste segude sobitamine kindlustuskahjudele

**Lühikokkuvõte:** Käesolevas magistritöös sobitame Eesti Liikluskindlustuse Fondist saadud kahjudele ühise skaalaparameetriga Erlangi jaotuste segusid. Anname ülevaate ühise skaalaparameetriga Erlangi jaotuste segudest ning nende parameetrite hindamisest EM algoritmiga. Meie eesmärgiks on võrrelda Erlangi jaotuste segude sobivust gamma-, lognormaalse, Weibulli ja Pareto jaotuste sobivusega. Näitame, et Erlangi jaotuste segud on heaks alternatiiviks eespool mainitud ja praktikas sagedasti kasutatavatele jaotustele.

**Märksõnad:** *kahjud, tõenäosusjaotus, jaotuste segu, Erlangi jaotus*

## Fitting mixtures of Erlang distributions to insurance claims

**Abstract:** In this master's thesis we fit mixtures of Erlang distributions with common scale parameter to loss data from Estonian Traffic Insurance Fund. We give an overview of mixtures of Erlang distributions with common scale parameter and estimation of parameters via the EM algorithm. Our goal is to compare the fit of mixtures of Erlang distributions with the fit of Gamma, lognormal, Weibull and Pareto distributions. We show that mixtures of Erlang distributions is a good alternative to those commonly used distributions.

**Keywords:** *loss, probability distribution, mixture of distributions, Erlang distribution*

# Sisukord

# Sissejuhatus

Pea iga inimtegevus on paratamatult seotud riskidega. Hommikusele lennule minnes on meil oht hiljaks jääda, sellest tulenevalt võime kiirustada ja hajutatud tähelepanu tõttu oleme avatud riskile tekitada liikluses ohtlik olukord või lausa õnnetus. Teisalt võib lend halva ilma tõttu ära jääda ja kui reis oli planeeritud uute töösuhete loomiseks, on risk potentsiaalsest kliendist või partnerist ilma jääda. See on vaid väike osa riskidest, mis selle pealtnäha ohutu ettevõtmise ajal realiseeruda võivad.

Ootamatute sündmuste võimalikul realiseerumisel tekkiv kahju puudutab meid kõiki. Kindlustamisest võib mõelda kui mehhanismist, mis jaotab tekkinud kahjud terve ühiskonna peale laiali, st kindlustatud osapoolele, kellel tekkis kahju, korvatakse see nende kindlustatute arvelt, kellel kahju ei tekkinud.

Aktuaari peamiseks ülesandeks on modelleerida kindlustusettevõtte rahavoogusid. Kindlustuspakkuja peamine sissetulekuallikas on kindlustuspreemiad, peamine kulu aga makstavad hüvitised. On selge, et esimene peab olema sõltuvuses teisest ning vastupidi. Kui me teaksime täpselt ette summat, mis me järgneval aastal peame hüvitisteks maksma, oleks meil väga lihtne leida nö ausad preemiamaksed. Riskide realiseerumine on aga oma olemuselt juhuslik ja seega võime me järgmise aasta hüvitiste kogumakset vaid hinnata. Ühelt poolt huvitab meid kahjude tekkimise sagedus ja teisalt tahame kirjeldada realiseerinud kahjude suurust. Antud töös keskendume just viimasele.

Kahjude suuruse modelleerimiseks kasutame mõne varasema perioodi andmeid ning eeldame, et need on teatud tõenäosusjaotusega juhuslike suuruste realisatsioonid. Kuna tekkinud kahjud on positiivsed reaalarvud, siis sobivateks kandidaatideks on mittenegatiivsed pidevad jaotused. Kahjude andmed sisaldavad sageli ebakorrapäraselt (harvasid) võrdlemisi suuri nõudeid ning seega tuleks vaatluse alla võtta raskema sabaga jaotused. Laialt kasutatud on gamma-, lognormaalne, Weibulli ja Pareto jaotused. Kui me oleme võimalikud jaotuste klassid valinud, peame klassidest valima sobivaimad jaotused.

Selleks hindame parameetrid. Antud töös kasutame parameetrite hindamiseks suurima tõepära meetodit ning selle modifikatsioone. Seejärel tahame saadud jaotuste sobivust omavahel võrrelda ning valida neist omakorda sobivaima. Lisaks sooviksime vastust küsimusele, kas meie poolt valitud jaotus kirjeldab andmeid piisavalt hästi. Selleks on võimalik teha statistilisi sobivuse (*goodness of fit*) teste. Mida aga teha olukorras, kus ükski teoreetiline jaotus ei läbi teste? Sellisel juhul ei jää aktuaaril üle muud, kui valida nö parim variant halbadest. Siinkohal on sobilik korrata George Boxi kuulsat motot: „Oma olemuselt on kõik mudelid valed, kuid mõned neist on kasulikud“ (Box ja Draper, 1987:424).

Antud töö peamiseks eesmärgiks on tutvustada ühise skaalaparameetriga Erlangi jaotuste segusid kui võimalikke kandidaatjaotusi kahjude suurusele ning võrrelda neid eelpool mainitud jaotustega. Selleks proovime sobitada jaotusi Eesti Liikluskindlustuse Fondist (ELF) saadud kahjude andmetele.

Pikemalt kirjeldame ühise skaalaparameetriga Erlangi jaotuste segude parameetrite hindamist, kuna just see on töö kõige keerulisem ning töömahukam osa. Jaotuste segude korral ei ole suurima tõepära meetod triviaalne, kuna tõepärafunktsiooni (või ka logaritmilist tõepärafunktsiooni) maksimiseerivate väärtuste leidmine analüütiliselt ei ole võimalik. Hinnangute leidmiseks kasutame EM algoritmi, mis on suurima tõepära meetodi iteratiivne modifikatsioon puuduvate andmete korral. Tutvustame EM algoritmi üldiselt ning kirjeldame EM algoritmi jaotuste lõplike segude ning ühise skaalaparameetriga Erlangi jaotuste segu korral.

Töö ülesehitus on järgnev. Esimeses peatükis anname lühikese ülevaate Erlangi jaotusest ja jaotuste lõplikest segudest üldiselt ning seejärel ühise skaalaparameetriga Erlangi jaotuste segudest. Teises tutvustame EM algoritmi parameetrite hindamiseks. Peatükis 3 kirjeldame EM algoritmi jaotuste segude korral ning seejärel teeme sama erijuhul ehk ühise skaalaparameetriga Erlangi jaotuste segu korral. Samuti toome selles osas ära mõned näited genereeritud andmete korral hinnangute leidmisest. Neljandas peatükis lähendame ELF-st saadud kahjude andmetele Erlangi jaotuste segusid ning teisi jaotusi ning võrdleme saadud tulemusi.

Kogu analüüs, sh näited ning reaalsetele kahjudele jaotuste sobitamine, on tehtud vabavaralise statistikapaketi R abil. Samuti on R-i funktsioonidena realiseeritud EM algoritm. Tähtsam osa programmikoodist on toodud lisades, kogu kood on lisatud CD-le.

Autor tänab siinkohal oma juhendajat Meelis Käärikut nõuannete ja konstruktiivse tagasiside eest.

# 1 Erlangi jaotus ja ühise skaalaparameetriga Erlangi jaotuste segu

## 1.1 Kahju suuruse jaotused

Meie eesmärgiks on modelleerida rahalist kahju, mille võib teatud kindlustatava riski realiseerumine tekitada kas era- või juriidilisele isikule. Antud töö kontekstis ei erista me kindlustatule tekkinud kahju kindlustusettevõtte kahjust, st me eeldame, et kõik nõuded makstakse kogu ulatuses välja. Kindlustuspakkuja kahju ühe poliisi pealt on ühest küljest alati mittenegatiivne, kuid silmas tuleb pidada ka seda, et kahju suurus võib potentsiaalselt olla väga suur. Seega peavad sobivad mudelid lubama praktikas realiseeruda väga suurtel väärtustel. Tõenäosusjaotused, mis seda lubavad, on nö raske sabaga. Need jaotused on asümmeetrilised ning nende parempoolsel sabal on võrdlemisi suur tõenäosusmass (Gray ja Pitts, 2012). Järgnevalt toome ära neli praktikas laialt kasutatavat jaotust kahjude suuruste modelleerimiseks.

- Me ütleme, et mittenegatiivne juhuslik suurus $X$ on gammajaotusega kujuparameetriga $\alpha > 0$ ja skaalaparameetriga $\theta > 0$ $(X \sim \Gamma(\alpha, \theta))$, kui ta tihedusfunktsioon on kujul

$$f_X(x;\alpha,\theta) = \begin{cases} \dfrac{x^{a-1}e^{-x/\theta}}{\theta^{\alpha}\Gamma(\alpha)}, & x > 0, \\ 0, & x \le 0, \end{cases}$$

kus $\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1}e^{-x}dx$ on gammafunktsioon.

- Me ütleme, et mittenegatiivne juhuslik suurus $X$ on lognormaalse jaotusega $(X \sim lnN(\mu,\sigma))$ asukohaparameetriga $\mu$ ja skaalaparameetriga $\sigma > 0$, kui ta tihedusfunktsioon on kujul

$$f_X(x;\mu,\sigma) = \begin{cases} \frac{1}{\sqrt{2\pi}x\sigma} e^{\frac{-(\ln x-\mu)^2}{2\sigma^2}}, & x > 0, \\ 0, & x \le 0. \end{cases}$$

- Me ütleme, et mittenegatiivne juhuslik suurus $X$ on Weibulli jaotusega kujuparameetriga $k > 0$ ja skaalaparameetriga $\lambda > 0$ $(X \sim We(k,\lambda))$, kui ta tihedusfunktsioon on kujul

$$f_X(x;k,\lambda) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}, & x > 0, \\ 0, & x \le 0. \end{cases}$$

- Me ütleme, et mittenegatiivne juhuslik suurus $X$ on Pareto jaotusega kujuparameetriga $\alpha > 0$ ja skaalaparameetriga $\lambda > 0$ $(X \sim Pa(\alpha,\lambda))$, kui ta tihedusfunktsioon on kujul

$$f_X(x;\alpha,\lambda) = \begin{cases} \frac{\alpha\lambda^\alpha}{(x+\lambda)^{\alpha+1}}, & x > 0, \\ 0, & x \le 0. \end{cases}$$

Toodud neli jaotust ei ole kindlasti ainukesed kandidaadid. Antud töö eesmärgiks on tutvustada ühise skaalaparameetriga Erlangi jaotuste segusid ning soovime näidata, et need on sobilikuks alternatiiviks eeltoodud ning ka teistele praktikas kasutatud jaotustele, kui meie eesmärgiks on modelleerida kahjude suurusi. Järgnevalt anname üldise ülevaate Erlangi jaotusest ning (ühise skaalaparameetriga) Erlangi jaotuste segudest.

## 1.2 Erlangi jaotus

Me ütleme, et mittenegatiivne juhuslik suurus $X$ on Erlangi jaotusest, kui ta tihedusfunktsioon on kujul

$$f_X(x;r,\theta)=\begin{cases}\dfrac{x^{r-1}e^{-x/\theta}}{\theta^r(r-1)!}, & x>0,\\ 0, & x\le 0,\end{cases} \tag{1.1}$$

kus $r$ on positiivne täisarvuline kujuparameeter $(r\in\mathbb{N})$ ja $\theta$ on positiivne reaalarvuline skaalaparameeter $(\theta>0)$. Seda, et juhuslik suurus $X$ on Erlangi jaotusest parameetritega $r$ ja $\theta$, tähistame $X\sim Erlang(r,\theta)$. Suurust $\lambda=1/\theta$ nimetame intensiivsusparameetriks ja eeltoodud tihedusfunktsiooni võib alternatiivse parametriseerimise korral kirjutada ka kui

$$f_X(x;r,\lambda)=\begin{cases}\dfrac{\lambda^r x^{r-1}e^{-\lambda x}}{(r-1)!}, & x>0,\\ 0, & x\le 0.\end{cases}$$

Paneme tähele, et kui valida $r=1$, saame eksponentjaotuse tihedusfunktsiooni ja seega on eksponentjaotus erijuht Erlangi jaotustest. Lisaks sellele saame Erlangi jaotust vaadelda kui gammajaotuse erijuhtu, täpsemalt naturaalarvulise kujuparameetriga gammajaotusena. Kui $Z_1,\ldots,Z_n$ on sõltumatud eksponentjaotusest juhuslikud suurused $(Z_i=Exp(\xi),\ i=1,\ldots,n)$, siis $S=\sum_{i=1}^{n}Z_i$ on gammajaotusega kujuparameetriga $n$ ja skaalaparameetriga $1/\xi$. Seega on sõltumatute eksponentjaotusest juhuslike suuruste summa jaotuseks Erlangi jaotus.

Jaotusfunktsiooni leiame tihedusfunktsiooni integreerimisel:

$$F_X(x;r,\theta)=\int_0^x\frac{s^{r-1}e^{-s/\theta}}{\theta^r(r-1)!}ds\ .$$

Antud juhul tuleb rakendada $r$ korda ositi integreerimist:

$$
\begin{aligned}
F_X(x;r,\theta) &= \int_0^x \frac{s^{r-1}e^{-s/\theta}}{\theta^r(r-1)!}ds = \frac{1}{\theta^{r-1}(r-1)!}\int_0^x \frac{1}{\theta}s^{r-1}e^{-s/\theta}ds \\
&= \frac{1}{\theta^{r-1}(r-1)!}\left[-x^{r-1}e^{-x/\theta} + \int_0^x (r-1)s^{r-2}e^{-s/\theta}ds\right] \\
&= -\frac{x^{r-1}e^{-x/\theta}}{\theta^{r-1}(r-1)!} + \frac{1}{\theta^{r-2}(r-2)!}\int_0^x \frac{1}{\theta}s^{r-2}e^{-s/\theta}ds \\
&= -\frac{x^{r-1}e^{-x/\theta}}{\theta^{r-1}(r-1)!} + \frac{1}{\theta^{r-2}(r-2)!}\left[-x^{r-2}e^{-x/\theta} + \int_0^x (r-2)s^{r-3}e^{-s/\theta}ds\right] \\
&= -\frac{x^{r-1}e^{-x/\theta}}{\theta^{r-1}(r-1)!} - \frac{x^{r-2}e^{-x/\theta}}{\theta^{r-2}(r-2)!} + \frac{1}{\theta^{r-3}(r-3)!}\int_0^x \frac{1}{\theta}s^{r-3}e^{-s/\theta}ds \\
&= \dots \\
&= \sum_{i=1}^{r-1} -\frac{x^i e^{-x/\theta}}{\theta^i i!} + \int_0^x \frac{1}{\theta}e^{-s/\theta}ds = 1 - \sum_{i=1}^{r-1}\frac{x^i e^{-x/\theta}}{\theta^i i!} - e^{-x/\theta} \\
&= 1 - \sum_{i=0}^{r-1}\frac{x^i e^{-x/\theta}}{\theta^i i!}.
\end{aligned}
$$

Keskväärtuse ja dispersiooni leidmiseks leiame esmalt Erlangi jaotusest juhusliku suuruse momente genereeriva funktsiooni:

$$
\begin{aligned}
M(t) = E(e^{tX}) &= \int_0^\infty e^{tx}\frac{x^{r-1}e^{-x/\theta}}{\theta^r(r-1)!}dx = \int_0^\infty \frac{x^{r-1}e^{-x(1/\theta - t)}}{\theta^r(r-1)!}dx \\
&= \int_0^\infty \frac{x^{r-1}e^{-x\frac{1-t\theta}{\theta}}}{\theta^r(r-1)!}dx = \frac{1}{(1-t\theta)^r}\int_0^\infty (1-t\theta)^r \cdot \frac{x^{r-1}e^{-x\frac{1-t\theta}{\theta}}}{\theta^r(r-1)!}dx = \\
&= (1-t\theta)^{-r}.
\end{aligned}
$$

Viimase võrduse juures kasutasime teadmist, et integraali alune avaldis on $X \sim Erlang(r, \theta/(1-t\theta))$ tihedusfunktsioon ja seega integreerub üheks. Momendid leiame seosest $E(X^n) = M^{(n)}(0)$. Seega saame

$$
\begin{aligned}
EX &= \frac{\partial}{\partial t}(1-t\theta)^{-r}\bigg|_{t=0} = -r(1-t\theta)^{-r-1}(-\theta)\bigg|_{t=0} = r\theta, \\
E\left(X^2\right) &= \frac{\partial^2}{\partial t^2}(1-t\theta)^{-r}\bigg|_{t=0} = r\theta(-r-1)(1-t\theta)^{-r-2}(-\theta)\bigg|_{t=0} = r(r+1)\theta^2, \\
DX &= E\left(X^2\right) - \left(EX\right)^2 = r^2\theta^2 + r\theta^2 - r^2\theta^2 = r\theta^2.
\end{aligned}
$$

## 1.3 Jaotuste lõplikud segud

Olgu $\boldsymbol{X} = (X_1, \ldots, X_n)$ juhuslik valim, kus $X_i$ on juhuslik suurus tõenäosustihedusega $f(x_i; \boldsymbol{\Psi})$ ja $\boldsymbol{\Psi} \in \Omega$ on parameetrite vektor parameeterruumist $\Omega$. Me eeldame, et tihedusfunktsioon $f(x_i; \boldsymbol{\Psi})$ on esitatav kujul

$$f(x_i; \boldsymbol{\Psi}) = \sum_{j=1}^{l} \alpha_j f_j(x_i; \boldsymbol{\theta}_j), \tag{1.2}$$

kus $f_j(x_i; \boldsymbol{\theta}_j)$ on tihedusfunktsioonid ja $\alpha_j$ on mittenegatiivsed üheks summeeruvad kaalud $(\alpha_j \geq 0,\ j = 1, \ldots, l,\ \sum_{j=1}^{l} \alpha_j = 1)$ ja $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_l),\ \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_l),\ \boldsymbol{\Psi} = (\boldsymbol{\theta}, \boldsymbol{\alpha})$. Seega $\boldsymbol{\Psi}$ on kõigi parameetrite vektor (nii komponentjaotuste parameetrid kui ka kaalud).

**Definitsioon.** Seosega (1.2) antud jaotusi nimetame jaotuste lõplikeks segudeks (McLachlan ja Peel, 2001).

Tihedusfunktsioone $f_j(x_i; \boldsymbol{\theta}_j)$ nimetame segu komponenttihedusteks ning nende poolt määratud jaotusi komponentjaotusteks või lühemalt komponentideks. Väljendi „jaotuste lõplik segu“ all mõtleme me selliseid jaotuste segusid, mille komponentide arv on lõplik $(l < \infty)$. Kuna antud töö raames käsitlemegi vaid selliseid jaotuste segusid, siis edasises me lõplikust iga kord eraldi ei rõhuta ning viitame neile kui jaotuste segule, vahel kasutame ka lihtsalt väljendit segu. Silmas tuleb pidada, et sellisel juhul mõtleme siiski vaid lõplikke segusid.

Jaotuste segude peamine eelis modelleerimisel on nende paindlikkus ning seetõttu pühendutakse nende uurimisele järjest enam, seda nii teoreetilisest kui ka praktilisest vaatenurgast. Jaotuste segusid on tulemuslikult rakendatud astronoomias, bioloogias, geneetikas, meditsiinis, psühhiaatrias, majanduses, inseneerias, turunduses ning teistes bioloogia, füüsika ning sotsiaalteaduste valdkondades (McLachlan ja Peel, 2001).

## 1.4 Ühise skaalaparameetriga Erlangi jaotuste segu

Me ütleme, et mittenegatiivse juhusliku suuruse $X$ jaotuseks on $M$ Erlangi jaotuse segu ühise skaalaparameetriga $\theta$, kui ta tihedusfunktsioon on kujul

$$f_X(x;\boldsymbol{\alpha},\boldsymbol{r},\theta)=\begin{cases}\sum_{j=1}^{M}\alpha_j\dfrac{x^{r_j-1}e^{-x/\theta}}{\theta^{r_j}(r_j-1)!}, & x>0,\\ 0, & x\le 0,\end{cases} \tag{1.3}$$

kus $M$ on Erlangi jaotuste arv segus, $\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_M)$ $(\alpha_j>0,\ \forall j\ \sum_{j=1}^{M}\alpha_j=1)$ on vastavate Erlangi jaotuste tihedusfunktsioonide $f_X(x;r_j,\theta)=\dfrac{x^{r_j-1}e^{-x/\theta}}{\theta^{r_j}(r_j-1)!}$ kaalud ning $\boldsymbol{r}=(r_1,\ldots r_M)$ on täisarvulised kujuparameetrid $(r_j\in\mathbb{N}, j=1,\ldots M)$. Üldsust kitsendamata eeldame, et $r_1<\ldots<r_M$.

Antud töös käsitleme vaid ühise skaalaparameetriga Erlangi jaotuste segusid, st me ei luba skaalaparameetril üle erinevate komponenttiheduste varieeruda. Seega viitame neile edasises kohati ka lihtsalt kui Erlangi jaotuste segudele. On selge, et Erlangi jaotuste segude tihedusfunktsioon on kooskõlas eelnevas peatükis toodud definitsiooniga — tegu on lõplikke jaotuste seguga, mille komponenttihedusteks on Erlangi jaotuse tihedused.

Leiame ka Erlangi jaotuste segu jaotusfunktsiooni ning seejärel momente genereeriva funktsiooni ning keskväärtuse ja dispersiooni. Jaotusfunktsiooni saame kujul

$$\begin{aligned}F_X(x;\boldsymbol{\alpha},\boldsymbol{r},\theta)&=\int_0^x\sum_{j=1}^{M}\alpha_j\frac{s^{r_j-1}e^{-s/\theta}}{\theta^{r_j}(r_j-1)!}ds=\sum_{j=1}^{M}\alpha_j\int_0^x\frac{s^{r_j-1}e^{-s/\theta}}{\theta^{r_j}(r_j-1)!}\\&=\sum_{j=1}^{M}\alpha_jF_X(x;r_j,\theta)=\sum_{j=1}^{M}\alpha_j\left[1-\sum_{i=0}^{r_j-1}\frac{x^ie^{-x/\theta}}{\theta^i i!}\right]\\&=1-\sum_{j=1}^{M}\alpha_j\sum_{i=0}^{r_j-1}\frac{x^ie^{-x/\theta}}{\theta^i i!}.\end{aligned}$$

Momente genereeriv funktsioon on

$$M(t) = E(e^{tX}) = \int_0^\infty e^{tx} \sum_{j=1}^{M} a_j \frac{x^{r_j-1} e^{-x/\theta}}{\theta^{r_j}(r_j-1)!} dx =$$
$$= \sum_{j=1}^{M} a_j \int_0^\infty e^{tx} \frac{x^{r_j-1} e^{-x/\theta}}{\theta^{r_j}(r_j-1)!} dx$$
$$= \sum_{j=1}^{M} a_j (1-t\theta)^{-r_j}.$$

Siit leiame keskväärtuse ja dispersiooni:

$$EX = \frac{\partial}{\partial t} \sum_{j=1}^{M} \alpha_j (1-t\theta)^{-r_j} \Big|_{t=0} = \sum_{j=1}^{M} \alpha_j r_j \theta (1-t\theta)^{-r_j-1} \Big|_{t=0} = \theta \sum_{j=1}^{M} \alpha_j r_j,$$
$$E\left(X^2\right) = \frac{\partial^2}{\partial t^2} \sum_{j=1}^{M} \alpha_j (1-t\theta)^{-r_j} \Big|_{t=0} = \sum_{j=1}^{M} \alpha_j r_j (r_j+1) \theta^2 (1-t\theta)^{-r_j-2} \Big|_{t=0} = \theta^2 \sum_{j=1}^{M} \alpha_j r_j (r_j+1),$$
$$DX = E\left(X^2\right) - \left(EX\right)^2 = \theta^2 \left[ \sum_{j=1}^{M} \alpha_j r_j (r_j+1) - \left( \sum_{j=1}^{M} \alpha_j r_j \right)^2 \right].$$

Järgnevalt esitame teoreemi (Tijms, 1994), mis väidab, et iga pidev mittenegatiivne juhuslik suurus on Erlangi jaotuste seguga mistahes täpsuseni lähendatav. Olgu antud mittenegatiivne juhuslik suurus $X$ jaotusfunktsiooniga $F(x)$. Defineerime järgneva ühise skaalaparameetriga $\theta > 0$ Erlangi jaotuste segu (kumulatiivse) jaotusfunktsiooni:

$$F(x;\theta) = \sum_{j=1}^{\infty} \alpha_j(\theta) F(x; j, \theta),$$

kus $F(x; j, \theta)$ on kujuparameetriga $j$ ja skaalaparameetriga $\theta$ Erlangi jaotuse jaotusfunktsioon,

$$F(x; j, \theta) = 1 - \sum_{i=0}^{j-1} \frac{x^i e^{-x/\theta}}{\theta^i i!},$$

ja komponentide kaalud on

$$\alpha_j(\theta) = F(j\theta) - F((j-1)\theta) \quad j = 1, 2, \ldots .$$

**Teoreem.** Ühise skaalaparameetriga Erlangi jaotuste segu klass on tihe pidevate positiivsete jaotuste ruumis. Täpsemalt, olgu $F(x)$ positiivse pideva juhusliku suuruse jaotusfunktsioon. Siis $\lim_{\theta \to 0} F(x;\theta) = F(x)$ iga $F(\cdot)$ pidevuspunkti korral.

Teoreemi tõestust me siinkohal ära ei too, kuid see on erineval kujul mitmes allikas, nende seas (Tijms, 1994), (Lee ja Lin, 2010) ning Roel Verbeleni magistritöös *„Phase-type distributions & mixtures of Erlangs: a study of theoretical concepts, calibration techniques & actuarial applications“* (Verbelen, 2013). Tulemust kasutame EM algoritmi algväärtustamisel Peatükis 3.2.

# 2 EM algoritm

## 2.1 Sissejuhatus

Olgu meil $n$ sõltumatut samast jaotusest vaatlust $\boldsymbol{x} = (x_1, \ldots x_n)$. Eeldame, et meil on sobiv kandidaatjaotus kirjeldamaks huvialuseid andmeid. Me tähistame nii tõenäosusfunktsiooni (diskreetsel juhul) kui ka tihedusfunktsiooni (pideval juhul) $f_X(x;\boldsymbol{\Psi})$-ga, kus $\boldsymbol{\Psi} \in \Omega$ on tundmatute parameetrite vektor parameeterruumist $\Omega$. Meie eesmärgiks on hinnata antud jaotuse parameetreid. Laialdaselt kasutatud lähenemine on suurima tõepära meetod. Selle korral maksimiseeritakse valimi tõepärafunktsioon $L(\boldsymbol{\Psi}) = \prod_{i=1}^{n} f_X(x_i;\boldsymbol{\Psi})$ ning saadakse parameetritele suurima tõepära (STP) hinnang — $\hat{\boldsymbol{\Psi}}_{STP} = \arg\max_{\boldsymbol{\Psi}\in\Omega} L(\boldsymbol{\Psi})$. Meetodi ideed on väga lihtne mõista just diskreetsel juhul, sest sellisel juhul on valimi tõepära võrdne tõenäosusega saada antud valimit ja seega leiame parameetritele hinnangud selliselt, et need maksimiseeriksid tõenäosust saada antud valimit. Lähenemisel on palju häid omadusi — suurima tõepära hinnangud on mõjusad, invariantsed, asümptootiliselt normaaljaotusega ning efektiivsed. Paraku ei ole meetod aga alati otseselt rakendatav. Näiteks ei pruugi maksimiseerimisülesanne olla analüütiliselt lahendatav. Sellisel juhul on võimalik kasutada iteratiivseid lähenemisi, näiteks Newton-Raphsoni meetodit. Järgnevalt teeme tutvust EM algoritmiga, mille abil on võimalik leida STP hinnang olukorras, kus meie vaatlusandmed ei ole täielikud ehk ei sisalda endas kogu meile vajalikku informatsiooni.

## 2.2 EM algoritm lihtsa numbrilise näite põhjal

Tegemist on suurima tõepära meetodi modifikatsiooniga mittetäielike andmete jaoks, mille formaliseerisid Dempster, Laird ja Rubin oma 1977 a. töös „*Maximum Likelihood from Incomplete Data via the EM Algorithm*“ (Dempster jt, 1997). Nimi tuleb

inglisekeelsest sõnapaarist *Expectation Maximization* (keskväärtuse maksimiseerimine) ja kirjeldab oma napisõnalisuses küllaltki hästi meetodi olemust. Nimelt on tegu iteratiivse lähenemisega, kus igal iteratsioonil teostatakse kaks sammu: keskväärtustamise samm (*E-step*, *Expectation step,* E-samm) ja maksimiseerimissamm (*M-step*, *Maximization step,* M-samm).

Nagu eelnevalt mainitud, leiab EM algoritm kasutust mittetäielike andmete korral. Lisaks ilmsetele olukordadele, nagu näiteks puuduvate väärtuste olemasolu, grupeeritud andmed, tsenseeritud või lõigatud sabadega vaatlused, esineb veel mitmeid olukordi, kus andmete mittetäielikkus ei ole esmapilgul ilmne. Näidetena võib tuua juhuslike mõjudega mudelid, log-lineaarsed mudelid ja jaotuste segud. Antud töö raames on erilise huvi all just viimased.

Enne EM algoritmi formaalsemat kirjeldust proovime anda esmase ettekujutuse ideest ühe lihtsa numbrilise näite najal (Dempster jt, 1977). Uurimise all on $n = 197$ looma ning nad on multinomiaalselt jaotunud nelja klassi. Meile kättesaadavateks andmeteks on sageduste vektor $\boldsymbol{x} = (x_1, x_2, x_3, x_4) = (125, 18, 20, 34)$. Klasside tõenäosused määrab teatud geneetiline mudel ning need on vastavalt

$$\frac{1}{2} + \frac{1}{4}\pi,\ \frac{1}{4}(1-\pi),\ \frac{1}{4}(1-\pi) \text{ ja } \frac{1}{4}\pi,$$

kus $0 \le \pi \le 1$. Eeldame nüüd, et esimene klass jaguneb kaheks alamklassiks, mille tõenäosused on vastavalt $\frac{1}{2}$ ja $\frac{1}{4}\pi$ ja olgu vastavad sagedused klassides $y_1$ ning $y_2$. Sellega oleme tehislikult tekitanud olukorra, kus meil puudub kogu informatsioon andmete kohta —— me teame kahe klassi sageduste summat, kuid mitte sagedusi kahes uues klassis. Uus sageduste vektor on $\boldsymbol{y} = (y_1, y_2, y_3, y_4, y_5)$, kus $y_1 + y_2 = x_1,\ y_3 = x_2,\ y_4 = x_3,\ y_5 = x_4.$ . Valimi tõenäosusfunktsioon on kujul

$$f_Y(\boldsymbol{y}; \pi) = \frac{n!}{y_1!\, y_2!\, y_3!\, y_4!\, y_5!} \left(\frac{1}{2}\right)^{y_1} \left(\frac{\pi}{4}\right)^{y_2} \left(\frac{1-\pi}{4}\right)^{y_3} \left(\frac{1-\pi}{4}\right)^{y_4} \left(\frac{\pi}{4}\right)^{y_5}.$$

Eeldame korraks, et me teame suuruseid $y_1$ ja $y_2$. Sellisel juhul saame leida STP hinnangu otse. Pärast tõenäosusfunktsiooni logaritmist ja tuletise ($\pi$ järgi) võrdsustamist 0-ga saame võrduse $(y_2+y_5)\frac{1}{\pi}=(y_3+y_4)\frac{1}{1-\pi}$ ja siit $\hat{\pi}_{ST}=\frac{y_2+y_5}{y_2+y_3+y_4+y_5}$. Selleks, et antud olukorras EM algoritm defineerida, peame näitama, kuidas toimub üleminek väärtuselt $\pi^{(p)}$ väärtusele $\pi^{(p+1)}$, kus $\pi^{(p)}$ tähistab $\pi$ hinnangu väärtust pärast $p$-ndat iteratsioonisammu, $p=0,1,2,\ldots$.

**Keskväärtustamise samm ehk E-samm.** Me teame väärtusi $y_3$, $y_4$ ja $y_5$ ja seega peame hindama vaid suurusi $y_1$ ning $y_2$. Hinnangud leiame kujul

$$y_1^{(p)}=125\frac{\frac{1}{2}}{\frac{1}{2}+\frac{1}{4}\pi^{(p)}} \text{ ja } y_2^{(p)}=125\frac{\frac{1}{4}\pi^{(p)}}{\frac{1}{2}+\frac{1}{4}\pi^{(p)}}. \qquad (2.1)$$

**Maksimiseerimissamm ehk M-samm**. Me käsitleme väärtuseid $(y_1^{(p)},y_2^{(p)},y_3,y_4,y_5)$ kui teadaolevaid ja leiame hinnangu:

$$\pi^{(p+1)}=\frac{y_2^{(p)}+y_5}{y_2^{(p)}+y_3+y_4+y_5}=\frac{y_2^{(p)}+34}{y_2^{(p)}+18+20+34}. \qquad (2.2)$$

Seejärel korratakse kahte sammu, kuni on saavutatud soovitav täpsus.

Asendades $y_2^{(p)}$ avaldisest (2.1) avaldisse (2.2) ja valides $\pi^*=\pi^{(p)}=\pi^{(p+1)}$ saame avaldise lihtsustamise järel leida hinnangu $\pi$-le ruutvõrrandist $197(\pi^*)^2-15\pi^*-68=0$ ja selle positivseks lahendiks on $\pi^*=\frac{15+\sqrt{53809}}{394}\approx 0.6268215$. Antud näide on taotluslikult valitud selline, et me saame leida lahendi $\pi^*$, sest sellisel juhul on algoritmi tulemusi parem illustreerida. Tegelikult me aga üldjuhul $\pi^*$ leida ei oska ja seega peame kasutama EM algoritmi. Valime algväärtuseks $\pi^{(0)}=0.5$ ja rakendame EM algoritmi.

**Tabel 1.** EM algoritmi tulemused illustreeriva näite korral

| $p$ | $\pi^{(p)}$ | $\left\|\pi^{(p)} - \pi^*\right\|$ | $(\pi^{(p+1)} - \pi^*)/(\pi^{(p)} - \pi^*)$ |
|---|---|---|---|
| 0 | 0.50000000 | 0.12682150 | 0.14646 |
| 1 | 0.60824742 | 0.01857408 | 0.13462 |
| 2 | 0.62432105 | 0.00250045 | 0.13302 |
| 3 | 0.62648888 | 0.00033262 | 0.13281 |
| 4 | 0.62677732 | 0.00004418 | 0.13278 |
| 5 | 0.62681563 | 0.00000587 | 0.13278 |
| 6 | 0.62682072 | 0.00000078 | 0.13278 |
| 7 | 0.62682139 | 0.00000010 | 0.13278 |
| 8 | 0.62682148 | 0.00000001 | - |

Saadud esimese kaheksa iteratsioonisammu tulemused on toodud Tabelis 1. Näeme, et koondumine toimub külaltki kiiresti — juba 8 sammuga saavutame täpsuse $10^{-8}$. Järjestikustel sammudel saadud vigade suhe on praktiliselt konstantne juba alates neljandast iteratsioonisammust.

## 2.3 EM algoritm

Järgnev meetodi kirjeldus põhineb Geoffrey J. McLachlani ning Thriyambakam Krishnani raamatul „*The EM Algorithm and Extensions. 2nd Edition*“ (McLachlan ja Krishnan, 2008). Olgu $\boldsymbol{X}$ juhuslik vektor realisatsiooniga $\boldsymbol{x}$ ning tihedusfunktsiooniga (diskreetse juhusliku suuruse korral tõenäosusfunktsiooniga) $f_X(\boldsymbol{x};\boldsymbol{\Psi})$, kus $\boldsymbol{\Psi} \in \Omega$. Rõhutame siinkohal, et kasutame kompaktsemat tähistust ja viimane tihedusfunktsioon on valimi tihedusfunktsioon ehk ühistihedusfunktsioon. Valim $\boldsymbol{x}$ esindab meie kasutuses olevaid andmeid, mis antud eesmärgi kontekstis on ebapiisav, mittetäielik.

Olgu $\boldsymbol{Y}$ täielikele andmetele $\boldsymbol{y}$ vastav juhuslik vektor tihedusfunktsiooniga $f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\Psi})$. Eeldades, et $\boldsymbol{y}$ on teada, avalduvad tõepära- ja log-tõepärafunktsioonid järgnevalt:

$$L_T(\boldsymbol{\Psi}) = f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\Psi}) \quad \text{ja} \quad l_T(\boldsymbol{\Psi}) = \ln L_T(\boldsymbol{\Psi}) = \ln f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\Psi}),$$

kus alaindeks $T$ tähistab täieliku informatsiooni olemasolu. Tegelikkuses ei ole vektor $\boldsymbol{y}$ antud ja seega on viimased suurused juhuslikud. Meie eesmärgiks on leida lahend võrrandile $\frac{\partial L_{MT}(\boldsymbol{\Psi})}{\partial \boldsymbol{\Psi}} = \mathbf{0}$ või alternatiivselt $\frac{\partial \ln L_{MT}(\boldsymbol{\Psi})}{\partial \boldsymbol{\Psi}} = \mathbf{0}$, kus $L_{MT}(\boldsymbol{\Psi}) = f_{\boldsymbol{X}}(\boldsymbol{x};\boldsymbol{\Psi})$ on mittetäielikele $(MT)$ andmetele vastav tõepärafunktsioon. Pühendume edasises justnimelt log-tõepära maksimiseerimisele, kuna logaritmimine on monotoonne teisendus ja lihtsustab väga tihti maksimiseerimist ning on seetõttu laialt kasutatav võte. Samad põhimõtted töötaksid ka tõepärafunktsiooni korral.

EM algoritm on loodud töötama olukordades, kus vaadeldud andmete log-tõepärafunktsiooni on keeruline maksimisseerida, kuid täielike andmete korral on see lihtne. Suurus $l_T(\boldsymbol{\Psi})$ on juhuslik ja seega vaatleme selle tinglikku keskväärtust (tingimusel, et on antud $\boldsymbol{x}$) fikseeritud $\boldsymbol{\Psi}$ korral. Kuna parameetrite väärtused ei ole teada, lahendatakse ülesanne iteratiivselt. Esmalt algväärtustame parameetrid — $\boldsymbol{\Psi}^{(0)}$. Seejärel, esimesel iteratsioonisammul, leiame E-sammul tingliku keskväärtuse $Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(0)}) = E\left[ l_T(\boldsymbol{\Psi}) \middle| \boldsymbol{x};\boldsymbol{\Psi}^{(0)} \right]$. M-sammul maksimiseerime saadud keskväärtust $\boldsymbol{\Psi}$ suhtes, st leiame $\boldsymbol{\Psi}^{(1)}$ nii, et $Q(\boldsymbol{\Psi}^{(1)};\boldsymbol{\Psi}^{(0)}) \geq Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(0)}) \ \ \forall \boldsymbol{\Psi} \in \Omega$. Seejärel võtame uueks lähendiks $\boldsymbol{\Psi}^{(1)}$ ja kordame E- ning M-samme. Üldiselt, iteratsioonisammul $(k+1)$, on meetod kirjeldatud järgnevalt.

**Keskväärtustamise samm.** Leiame $Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(k)})$, kus

$$Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(k)}) = E\left[ l_T(\boldsymbol{\Psi}) \middle| \boldsymbol{x};\boldsymbol{\Psi}^{(k)} \right]. \tag{2.3}$$

**Maksimiseerimissamm.** Valime parameeterruumist $\Omega$ sellise väärtuse $\boldsymbol{\Psi}^{(k+1)}$ nii, et ta maksimiseerib $Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(k)})$, st

$$Q(\boldsymbol{\Psi}^{(k+1)};\boldsymbol{\Psi}^{(k)}) \geq Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(k)}) \ \forall \ \boldsymbol{\Psi} \in \Omega. \tag{2.4}$$

Samme korratakse soovitud täpsuse saavutamiseni, st kuni erinevus $l_{MT}(\boldsymbol{\Psi}^{(k+1)}) - l_{MT}(\boldsymbol{\Psi}^{(k)})$ on piisavalt väike. Dempster, Laird ja Rubin (1977) näitasid, et EM algoritmi tulemusel saadud parameetrite jada korral on vastav log-tõepärafunktsiooni väärtuste jada kasvav — $l_{MT}(\boldsymbol{\Psi}^{(k+1)}) \geq l_{MT}(\boldsymbol{\Psi}^{(k)})$. See tähendab, et logaritmilised tõepärad iteratsioonide käigus ei kahane.

# 3 Ühise skaalaparameetriga Erlangi jaotuste segude parameetrite hindamine

## 3.1 EM algoritm jaotuste lõplike segude korral

Käesolevas osas, mis põhineb raamatul „*Finite Mixture Models*“ (McLachlan ja Peel, 2001), anname ülevaate mittetäielike andmete esinemisest jaotuste lõplike segude kontekstis ning konstrueerime EM algoritmi antud olukorras.

Olgu $\boldsymbol{X} = (X_1, \ldots, X_n)$ sõltumatu sama jaotusega juhuslik valim, kus $X_i$ on juhuslik suurus tõenäosustihedusega (1.2). Oletame, et me tahame genereerida juhusliku suuruse $X_i$ jaotuste segust (1.2). Olgu $Z_i$ diskreetne juhuslik suurus, mis võib omandada väärtusi $1, ..., l$ tõenäosustega $\alpha_1, \ldots, \alpha_l$ (vastavalt) ning eeldame, et $X_i$ tinglik tihedus tingimusel $Z_i = j$ on $f_j(x_i; \boldsymbol{\theta}_j)$ $(j = 1, \ldots, l)$. Seega genereerime juhusliku suuruse $Z_i$ ning saadud väärtuse $j$ korral genereerime $X_i$ tihedusest $f_j(x_i; \boldsymbol{\theta}_j)$ $(j = 1, \ldots, l)$. Sellisel juhul on juhusliku suuruse $X_i$ marginaalne tihedus $f(x_i; \boldsymbol{\Psi})$. Antud interpretatsiooni kasutame edasises EM algoritmiga parameetrite hindamisel, kuid suuruse $Z_i$ asemel defineerime vektorid $\boldsymbol{Z_i} = (Z_{i1}, \ldots Z_{il})$, kus

$$Z_{ij} = \begin{cases} 1, \text{ kui vaatluse } x_i \text{ genereeris } j\text{-s komponent } f_j(x_i; \boldsymbol{\theta}_j), \\ 0, \text{ vastasel korral,} \end{cases}$$

$i = 1, \ldots, n$ ja $j = 1, \ldots, l$. Juhuslik suurus $Z_i$ ja vektor $\boldsymbol{Z_i}$ täidavad sama rolli. Paneme tähele, et $P(\boldsymbol{Z_i} = \boldsymbol{z_i}; \boldsymbol{\alpha}) = \alpha_1^{z_{i1}} \cdot \ldots \alpha_l^{z_{il}}$ ehk $\boldsymbol{Z_i}$ on multinomiaalse jaotusega — $\boldsymbol{Z_i} \sim Mult_l(1, \boldsymbol{\alpha})$. Juhul kui populatsioon $G$ on jaotunud $l$ gruppi $G_1, \ldots, G_l$ proportsioonidega $\alpha_1, \ldots, \alpha_l$ ja juhusliku suuruse $X_i$ tihedus grupis $j$ on $f_j(x_i; \boldsymbol{\theta}_j)$ $j = 1, \ldots, l,$ saame tiheduse (1.2) komponendid reaalselt siduda eksisteerivate gruppidega.

Viimane näide on kooskõlas siin välja toodud interpretatsiooniga jaotuste segudest. Alati ei ole aga sisuliselt sobilik jaotuste segudest sääraselt mõelda, kuid vaatluste sidumine indikaatorvektoritega $\boldsymbol{Z}_i$ võib olla praktiliselt väga kasulik.

Olgu antud juhusliku valimi $\boldsymbol{X} = (X_1, \ldots, X_n)$ realisatsioon $\boldsymbol{x} = (x_1, \ldots, x_n)$. Valimit $\boldsymbol{x}$ vaatleme kui mittetäielikku, sest vastavad komponentide indikaatorid $\boldsymbol{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$, kus $\boldsymbol{z}_i$ on $\boldsymbol{Z}_i$ realisatsioon $i = 1, \ldots, n$, on meile teadmata. Me eeldame ka, et valimi $\boldsymbol{x}$ korral ei ole puuduvaid väärtusi või tsenseerimist jms ehk kogu mittetäielikkus tuleneb indikaatorite puudumisest. Juhusliku suuruse $X_i$ ning vastavate indikaatorite $\boldsymbol{Z}_i$ ühisjaotus on kujul

$$\begin{aligned} f_{X_i,\boldsymbol{Z}_i}(x_i, \boldsymbol{z}_i; \boldsymbol{\Psi}) &= f_{X_i|\boldsymbol{Z}_i}(x_i \mid \boldsymbol{z}_i; \boldsymbol{\theta}) P(\boldsymbol{Z}_i = \boldsymbol{z}_i; \boldsymbol{\alpha}) \\ &= \prod_{j=1}^{l} f_j(x_i; \boldsymbol{\theta}_j)^{z_{ij}} \prod_{j=1}^{l} \alpha_j^{z_{ij}} \\ &= \prod_{j=1}^{l} \left[ \alpha_j f_j(x_i; \boldsymbol{\theta}_j) \right]^{z_{ij}}. \end{aligned}$$

Ning seega saab täielike andmete logaritmiline tõepära kuju

$$\begin{aligned} l_T(\boldsymbol{\Psi}) &= \ln \left\{ \prod_{i=1}^{n} \prod_{j=1}^{l} \left[ \alpha_j f_j(x_i; \boldsymbol{\theta}_j) \right]^{z_{ij}} \right\} \\ &= \sum_{i=1}^{n} \sum_{j=1}^{l} z_{ij} \left\{ \ln(\alpha_j) + \ln(f_j(x_i; \boldsymbol{\theta}_j)) \right\}. \end{aligned} \tag{3.1}$$

Tuletame meelde, et EM algoritmi keskväärtustamise sammul leitakse $Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)})$, mis on antud seosega (2.3). Kuna täielike andmete logaritmiline tõepära on $z_{ij}$ suhtes lineaarne, taandub E-samm iteratsioonil $(k+1)$ tinglike keskväärtuste $E(Z_{ij} \mid \boldsymbol{x}; \boldsymbol{\Psi}^{(k)})$ leidmisele:

$$\begin{aligned}
\mathrm{z}_{ij}^{(k+1)} := E(Z_{ij} \mid \boldsymbol{x}; \boldsymbol{\Psi}^{(k)}) &= P(Z_{ij} = 1 \mid x_i; \boldsymbol{\Psi}^{(k)}) \\
&= \frac{f(x_i \mid Z_{ij} = 1; \boldsymbol{\Psi}^{(k)}) P(Z_{ij} = 1; \boldsymbol{\Psi}^{(k)})}{f(x_i; \boldsymbol{\Psi}^{(k)})} \\
&= \frac{f_j(x_i; \boldsymbol{\theta}_j^{(k)}) P(Z_{ij} = 1; \boldsymbol{\alpha}^{(k)})}{f(x_i; \boldsymbol{\Psi}^{(k)})} \\
&= \frac{\alpha_j^{(k)} f_j(x_i, \boldsymbol{\theta}_j^{(k)})}{\sum_{h=1}^{l} \alpha_h^{(k)} f_h(x_i, \boldsymbol{\theta}_h^{(k)})},
\end{aligned}$$

$i = 1, \ldots, n$ ja $j = 1, \ldots, l$. Suurusi $\mathrm{z}_{ij}^{(k+1)}$ nimetame järeltõenäosusteks, et $i$. vaatlus kuulub $j$. segu komponenti, ning analoogselt nimetame suurusi $\alpha_j$ eeltõenäosusteks, et $i$. vaatlus kuulub $j$. segu komponenti (sarnaselt Bayesi statistika põhimõtetele). On lihtne näha, et suurused $\mathrm{z}_{ij}^{(k+1)}$ summeeruvad fikseeritud $i$ korral üheks . Kokkuvõtvalt saame

$$Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)}) = \sum_{i=1}^{n} \sum_{j=1}^{l} z_{ij}^{(k+1)} \left\{ \ln(\alpha_j) + \ln(f_j(x_i; \boldsymbol{\theta}_j)) \right\}.$$

Maksimiseerimissammul leiame uue väärtuse $\boldsymbol{\Psi}^{(k+1)}$ funktsiooni $Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)})$ maksimiseerimisel $\boldsymbol{\Psi}$ suhtes üle parameeterruumi $\Omega$. Jaotuste segude korral leitakse iteratsioonidel hinnangud $\alpha_j^{(k+1)}$ suurustele $\alpha_j$ sõltumatult hinnangutest $\boldsymbol{\theta}_j^{(k)}$ jaotuste parameetritele $\boldsymbol{\theta}_j$. Leiame esmalt eeskirja, kuidas uuendada jaotuste segu kaale. Silmas pidades, et kaalud peavad summeeruma üheks, kirjutame $Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)})$ alternatiivselt

$$Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)}) = \sum_{i=1}^{n} \sum_{j=1}^{l-1} z_{ij}^{(k+1)} \ln(\alpha_j) + \sum_{i=1}^{n} z_{il}^{(k+1)} \ln(1 - \sum_{j=1}^{l-1} \alpha_j) + \sum_{i=1}^{n} \sum_{j=1}^{l} z_{ij}^{(k+1)} \ln(f_j(x_i; \boldsymbol{\theta}_j))$$

ja võrdsustame osatuletised $\alpha_j$ $(j = 1, \ldots, l)$ suhtes nulliga

$$\left. \frac{\partial Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)})}{\partial \alpha_j} \right|_{\boldsymbol{\alpha} = \boldsymbol{\alpha}^{(k+1)}} = \left. \frac{\sum_{i=1}^{n} z_{ij}^{(k+1)}}{\alpha_j} - \frac{\sum_{i=1}^{n} z_{il}^{(k+1)}}{1 - \sum_{h=1}^{l-1} \alpha_j} \right|_{\boldsymbol{\alpha} = \boldsymbol{\alpha}^{(k+1)}} = \left. \frac{\sum_{i=1}^{n} z_{ij}^{(k+1)}}{\alpha_j} - \frac{\sum_{i=1}^{n} z_{il}^{(k+1)}}{\alpha_l} \right|_{\boldsymbol{\alpha} = \boldsymbol{\alpha}^{(k+1)}} = 0,$$

$j = 1, \ldots, l-1$. Siit saame, et

$$\alpha_j^{(k+1)} = \alpha_l^{(k+1)} \frac{\sum_{i=1}^{n} z_{ij}^{(k+1)}}{\sum_{i=1}^{n} z_{il}^{(k+1)}}, \quad j = 1, \ldots, l-1, \tag{3.2}$$

ja arvestades, et kaalud peavad üheks summeeruma, saame

$$\begin{aligned}\sum_{j=1}^{l} \alpha_j^{(k+1)} &= \frac{\sum_{j=1}^{l} \alpha_l^{(k+1)} \sum_{i=1}^{n} z_{ij}^{(k+1)}}{\sum_{i=1}^{n} z_{il}^{(k+1)}} = \frac{\alpha_l^{(k+1)} \sum_{j=1}^{l} \sum_{i=1}^{n} z_{ij}^{(k+1)}}{\sum_{i=1}^{n} z_{il}^{(k+1)}} \\ &= \frac{\alpha_l^{(k+1)} \sum_{i=1}^{n} \sum_{j=1}^{l} z_{ij}^{(k+1)}}{\sum_{i=1}^{n} z_{il}^{(k+1)}} = \frac{\alpha_l^{(k+1)} \sum_{i=1}^{n} 1}{\sum_{i=1}^{n} z_{il}^{(k+1)}} \\ &= \frac{n\alpha_l^{(k+1)}}{\sum_{i=1}^{n} z_{il}^{(k+1)}} = 1\end{aligned}$$

ning siit leiame

$$\alpha_l^{(k+1)} = \frac{\sum_{i=1}^{n} z_{il}^{(k+1)}}{n}.$$

Asendades saadud suuruse seosesse (3.2) saame

$$\alpha_j^{(k+1)} = \frac{\sum_{i=1}^{n} z_{ij}^{(k+1)}}{n} \qquad j = 1, \ldots, l. \tag{3.3}$$

Viimane tuletuskäik on läbi tehtud Roel Verbeleni töös (2013). Samuti näitas ta, et tegu on tõepoolest maksimumpunktiga. Saadud tulemus ütleb, et iteratsioonisammul $(k+1)$ on $\alpha_j$ hinnanguks segu komponenti kuulumise järeltõenäosuste keskmine üle valimi. Paneme tähele, et uuenduseeskiri (3.3) on jaotuste segude korral universaalne, st ei sõltu sellest, mis jaotustega parajasti tegu on.

Lisaks eelnevalt leitud kaalude hinnangutele, peame igal iteratsioonil uuendama ka komponenttiheduste parameetrite hinnanguid. Viimased leitakse järgnevate võrrandite lahendamisel:

$$\sum_{i=1}^{n} \sum_{j=1}^{l} z_{ij} \, \partial \ln(f_j(x_i; \boldsymbol{\theta}_j)) / \partial \boldsymbol{\theta} = 0 \, .$$

## 3.2 EM algoritm Erlangi jaotuste segude korral

Antud peatükis lähtume eelmises selgitatud põhimõtetest ning kirjeldame EM algoritmi ühise skaalaparameetriga Erlangi jaotuste segude korral. Tuletuskäigud põhinevad Roel Verbeleni magistritööl (2013). Olgu $\boldsymbol{X} = (X_1, \dots X_n)$ sõltumatu juhuslik valim ühise skaalaparameetriga Erlangi jaotuste segust tihedusfunktsiooniga (1.3) ning $\boldsymbol{x} = (x_1, \dots x_n)$ selle realisatsioon. Eeldame, et Erlangi jaotuste arv segus $M$ ja kujuparameetrid $\boldsymbol{r} = (r_1, \dots r_M)$ on fikseeritud. Seega tahame hinnata kaale $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_M)$ $(\alpha_j > 0,\ \forall j \sum_{j=1}^{M} \alpha_j = 1)$ ning ühist skaalaparameetrit $\theta$. Et luua seos eelnevas peatükis toodud üldisemate tähistustega võime kirjutada $\boldsymbol{\Psi} = (\theta, \boldsymbol{\alpha})$, sest parameetrite vektori $\boldsymbol{\theta}$ asemel vaatleme skalaarset suurust $\theta$. Märgime siinkohal ära, et parameetrite hindamisülesande korral ei ole aga suurus $M$ ning kujuparameetrid meile teada ning need tuleb leida tuginedes andmetele. Kuidas seda täpsemalt teha, kirjeldame edasises.

**Algväärtustamine.** Algväärtustamine põhineb Peatükis 1.4 toodud Tijmsi teoreemil. Fikseeritud $M$ korral anname esialgselt kujuparameetritele väärtused $r_j = j$, kus $j = 1, \dots, M$. Esialgse skaalaparameetri $\theta^{(0)}$ valime selliselt, et $\theta^{(0)} r_M$ võrduks vaatluste maksimaalse väärtusega ning esialgseteks kaaludeks võtame vaatluste suhtelised sagedused intervallides $(r_{j-1}\theta^{(0)}, r_j\theta^{(0)}]$. Seega leiame algväärtused kujul

$$\theta^{(0)} = \frac{\max(\boldsymbol{x})}{r_M},\ \alpha_j^{(0)} = \frac{\sum_{i=1}^{n} I(r_{j-1}\theta^{(0)} < x_i \le r_j\theta^{(0)})}{n}\ \ j = 1, \dots, M,\ r^{(0)} = 0, \quad (3.4)$$

kus $I(\cdot)$ on indikaatorfunktsioon. Selliselt valitud algväärtused garanteerivad, et juba esialgne hinnang on küllaltki hea, ja seetõttu väheneb arvutusteks vajaminev aeg.

**Keskväärtustamise samm.** Nagu eelnevalt mainisime, taandub jaotuste segude korral keskväärtustamise samm segu komponentidesse kuulumise järeltõenäosuste $z_{ij}^{(k+1)}$ $(i = 1, \dots, n$ ja $j = 1, \dots, M)$ arvutamisele. Erlangi jaotuste segu korral on need kujul

$$z_{ij}^{(k+1)} = E(Z_{ij} \mid \boldsymbol{x}; \boldsymbol{\Psi}^{(k)}) = \frac{\alpha_j^{(k)} f_X(x_i; r_j, \theta^{(k)})}{\sum_{h=1}^{M} \alpha_h^{(k)} f_X(x_i; r_h, \theta^{(k)})} \quad i = 1, \ldots n,\ j = 1, \ldots, M,$$

kus $f_X(x_i; r_j, \theta^{(k)})$ on Erlangi jaotuse tihedusfunktsioon (1.1) kujuparameetri $r_j$ ja skaalaparameetri $\theta^{(k)}$ korral. Ühise skaalaparameetriga Erlangi jaotuste segu korral saab täielike andmete logaritmiline tõepära (3.1) kuju

$$l_T(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \sum_{j=1}^{M} z_{ij} \left\{ \ln(\alpha_j) + (r_j - 1)\ln(x_i) - \frac{x_i}{\theta} - r_j \ln(\theta) - \ln((r_j - 1)!) \right\}$$

ja seega

$$Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)}) = \sum_{i=1}^{n} \sum_{j=1}^{M} z_{ij}^{(k+1)} \left\{ \ln(\alpha_j) + (r_j - 1)\ln(x_i) - \frac{x_i}{\theta} - r_j \ln(\theta) - \ln((r_j - 1)!) \right\}.$$

**Maksimiseerimissamm.** Kaalude hinnangute muutmine iteratsioonidel ei sõltunud jaotustest ja need leiame seostest (3.3). Skaalaparameetri hinnangu saame $Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)})$ osatuletise $\theta$ järgi võrdsustamisel 0-ga:

$$\begin{aligned}
\left. \frac{\partial Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)})}{\partial \theta} \right|_{\theta=\theta^{(k+1)}} &= \left. \sum_{i=1}^{n} \sum_{j=1}^{M} z_{ij}^{(k+1)} \left\{ \frac{x_i}{\theta^2} - \frac{r_j}{\theta} \right\} \right|_{\theta=\theta^{(k+1)}} \\
&= \left. \frac{1}{\theta^2} \sum_{i=1}^{n} \sum_{j=1}^{M} z_{ij}^{(k+1)} x_i - \frac{n}{\theta} \sum_{i=1}^{n} \sum_{j=1}^{M} \frac{z_{ij}^{(k+1)}}{n} r_j \right|_{\theta=\theta^{(k+1)}} \\
&= \left. \frac{1}{\theta^2} \sum_{i=1}^{n} x_i \sum_{j=1}^{M} z_{ij}^{(k+1)} - \frac{n}{\theta} \sum_{j=1}^{M} \sum_{i=1}^{n} \frac{z_{ij}^{(k+1)}}{n} r_j \right|_{\theta=\theta^{(k+1)}} \\
&= \left. \frac{1}{\theta^2} \sum_{i=1}^{n} x_i - \frac{n}{\theta} \sum_{j=1}^{M} \alpha_j^{(k+1)} r_j \right|_{\theta=\theta^{(k+1)}} = 0.
\end{aligned}$$

Seega

$$\theta^{(k+1)} = \frac{\frac{1}{n} \sum_{i=1}^{n} x_i}{\sum_{j=1}^{M} \alpha_j^{(k+1)} r_j} = \frac{\overline{\boldsymbol{x}}}{\sum_{j=1}^{M} \alpha_j^{(k+1)} r_j}.$$

Verbelen (2013) on näidanud, et $\theta^{(k+1)}$ tõepoolest maksimiseerib $Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(k)})$.

Iteratsioone kordame kuni vajaliku (eelnevalt määratud) täpsuse saavutamiseni. Enim kasutatud algoritmi lõpetamise kriteerium põhineb logaritmilise tõepära paranemisel, st jätkame iteratsioone, kuni log-tõepära paraneb piisavalt. Antud lähenemise nõrgaks küljeks on see, et sellisel juhul antakse suurem tähtsus sobivusele andmete põhiosas. Isegi kui sabaosas paraneb sobivus arvestataval määral, võib logaritmiline tõepära kasvada liialt vähe, sest sabaosal on väiksem kaal. Kindlustusmatemaatika rakendustes on sageli aga väga suur tähtsus just sabaosal ning sellistel juhtudel võiks kaaluda alternatiivseid lähenemisi. Üheks võimaluseks oleks logaritmilise tõepära asemel kasutada kaalutud summat ja anda sabaosale suurem kaal. Sellisel juhul jätkaksime algoritmi, kuni antud kaalutud summa paraneb piisavalt (Lee ja Lin, 2010). Antud töö raames kasutame piiratud aja tõttu esimest lähenemist, kuid tulevikus võiks põhjalikumalt uurida ka alternatiivseid võimalusi, seda eriti olukordades, kus hinnangud sabaosas on suure tähtsusega.

Märgime veelkord ära, et eelkirjeldatud iteratiivne algoritm maksimiseerib kaalude ja skaalaparameetri hinnanguid fikseeritud Erlangi jaotuste arvu $M$ ja kujuparameetrite $\boldsymbol{r} = (r_1, \ldots r_M)$ korral. Kuna kujuparameetrid on algoritmi vältel fikseeritud ja me neid ei muuda, saavutame me antud hinnangutega log-tõepära lokaalse maksimumi. Peatükis 3.4 kirjeldame kujuparameetrite muutmise protseduuri, mis aitab saadud hinnanguid parandada.

## 3.3 EM algoritm kokkuvõtvalt

**Algväärtustamine:**

$$\theta^{(0)} = \frac{\max(\boldsymbol{x})}{r_M} \text{ ja } \alpha_j^{(0)} = \frac{\sum_{i=1}^{n} I(r_{j-1}\theta^{(0)} < x_i \le r_j\theta^{(0)})}{n} \; j = 1, \ldots, M \text{ ning } r^{(0)} = 0.$$

**E-samm:**

$$z_{ij}^{(k+1)} = \frac{\alpha_j{}^{(k)} f_X(x_i; r_j, \theta^{(k)})}{\sum_{h=1}^{M} \alpha_h{}^{(k)} f_X(x_i; r_h, \theta^{(k)})} \quad i = 1, \ldots n, \; j = 1, \ldots, M.$$

**M-samm:**

$$\begin{cases} \alpha_j^{(k+1)} = \dfrac{\sum_{i=1}^{n} z_{ij}^{(k+1)}}{n}, \; j = 1, \ldots, M \\ \theta^{(k+1)} = \dfrac{\bar{x}}{\sum_{j=1}^{M} \alpha_j^{(k+1)} r_j}. \end{cases}$$

Märgime siinkohal ära, et kui intervallis $(r_{j-1}\theta^{(0)}, r_j\theta^{(0)}\rceil$ ei ole ühtegi vaatlust, siis vastav esialgne kaal võrdub nulliga $(\alpha_j^{(0)} = 0)$ ja antud kaal jääb nulliks terve algoritmi vältel. Viimast on lihtne näha siin toodud E- ning M-sammu hinnangute uuenduseeskirjadest. Seega saame praktikas kõik kaaludele $\alpha_j$, mille korral $\alpha_j^{(0)} = 0,$ vastavad komponendid vaatluse alt välja jätta (programmeerimise kontekstis saame antud kaalud ja nende kujuparameetrid eemaldada).

## 3.4 Kujuparameetrite kohandamine ning Erlangi jaotuste arvu valik

Nagu juba eelnevalt mainisime, leiame me fikseeritud jaotuste arvu $M$ ja kujuparameetrite korral hinnangud, mis maksimiseerivad logaritmilist tõepära lokaalselt. Ideaalis on meie eesmärgiks leida globaalne maksimum. Ilmne võimalus logaritmilise tõepära maksimiseerimiseks üle suurema piirkonna on kaasata segusse suurem arv Erlangi jaotusi ja seega ka suurem arv kujuparameetreid. Äärmuslikul juhul saaksime kujuparameetritena vaadelda tervet naturaalarvude hulka. Praktikas seab meile aga piirangud arvutusteks kuluv aeg. Kõige enam piirab arvutusmaht jaotuste arvu (segus) valikut. Eeldame, et on fikseeritud selline jaotuste arv segus, mille korral algoritm suudab hinnangud leida meile aktsepteeritava aja jooksul. Seega tahame leida sellised hinnangud, mis on antud jaotuste arvu $M$ korral parimad. Selleks tuleks muuta kujuparameetreid ja võrrelda logaritmilisi tõepärasid erinevate kujuparameetrite korral. Järgnevalt tutvustame Lee ja Lini (2010) poolt kirjeldatud ning Verbeleni (2013) poolt täiendatud kujuparameetrite kohandamise protseduuri.

1. Leiame EM algoritmiga hinnangud parameetritele fikseeritud komponentide arvu $M$ ja algsete kujuparameetrite $\{r_1, r_2, \ldots r_M\}$ korral.

2. Suurendame viimast kujuparameetrit ühe võrra, st võtame vaatluse alla kujuparameetrid $\{r_1, r_2, \ldots r_M + 1\}$, ning leiame EM algoritmiga uued hinnangud, kusjuures algväärtusteks võtame eelneval sammul EM algoritmiga saadud hinnangud. Kui logaritmiline tõepära suureneb, siis asendame vanad parameetrid uutega. Sammu kordame, kuni log-tõepära paraneb.

3. Rakendame sammu 2 eelviimasele kujuparameetrile jne, kuni oleme kõiki kujuparameetrid muutnud.

4. Vähendame esimest kujuparameetrit ühe võrra, st võtame vaatluse alla kujuparameetrid $\{r_1 - 1, r_2, \ldots r_M\}$, ning käitume analoogselt sammus 2 kirjeldatule.

5. Rakendame sammu 4 teisele kujuparameetrile jne, kuni oleme kõiki kujuparameetreid muutnud.

6. Pöördume tagasi sammu 2 juurde ja kordame samme 2 – 4 kuni logaritmiline tõepära enam ei parane.

Jaotuste arvu (segus) suurendamisega saame logaritmilist tõepära suurendada. Samas liiga suure jaotuste arvu korral peame arvestama ka ülesobitamisega. Seega on vajalik mingi otsustuseeskiri, mille abil määrata Erlangi jaotuste arv segus. Antud töös kasutame selleks Akaike informatsioonikriteeriumit (AIC) ning Schwarzi (alternatiivselt ka Bayesi) informatsioonikriteeriumit (BIC), mis on antud valemitega

$$AIC = -2l_{MT}(\hat{\boldsymbol{\Psi}}) + 2k, \quad BIC = -2l_{MT}(\hat{\boldsymbol{\Psi}}) + k\ln(n),$$

kus $l_{MT}(\hat{\boldsymbol{\Psi}})$ on meie kasutuses olevate (mittetäielike) andmete logaritmiline tõepära saadud hinnangute korral ja $k$ on parameetrite arv ning $n$ on valimi maht. Nimetatud informatsioonikriteeriumid koondame tähistuse IC (*information criterion*) alla. Toodud valemitest on lihtne näha, et BIC on konservatiivsem — „trahv“ parameetrite arvu pealt

on suurem, sest juba kümnevaatluselise valimi korral on parameetrite arvu kordajaks $\ln(10) \approx 2.303$. Tahame leida sellist jaotuste arvu, mille korral IC on minimaalne. Lee ja Lin (2010) kirjeldasid järgmise protseduuri. Esmalt fikseerime jaotuste arvu $M$ ning seejärel käitume järgnevalt.

1. Leiame EM algoritmiga parameetritele hinnangud, rakendame eelnevalt kirjeldatud kujuparameetrite kohandamise protseduuri ja arvutame IC.

2. Eemaldame segust sellise Erlangi jaotuse (kujuparameetriga $r_j$), mille kaal $\alpha_j$ on vähim, standardiseerime kaalud, võtame algväärtusteks viimased hinnangud (sh standardiseerimisel saadud kaalud) ja kordame sammus 1 kirjeldatut.

Samme korratakse seni kuni IC väheneb. Kui IC enam ei vähenenud, siis hinnanguks on viimasel korral saadud sammu 1 tulemus.

Võrreldes standardse lähenemisega vähendab antud protseduur arvutuseks kuluvat aega märgatavalt. Standardse lähenemise korral leiame esmalt IC ühe Erlangi jaotuse korral ja seejärel suurendame arvu $M$ igal sammul ühe võrra kuni IC enam ei parane. Kuid kui me suurendame Erlangi jaotuste arvu segus, siis eelmisel sammul saadud parameetrite hinnangud ei sisalda endas kasulikku informatsiooni algväärtustamiseks ja seega peame algväärtustamisel kasutame seoseid (3.4). Kui me aga eemaldame vähima kaaluga Erlangi jaotuse, siis logaritmiline tõepära ei vähene märkimisväärselt. Seega eelneval sammul saadud skaalaparameeter ja standardiseeritud kaalud on küllaltki lähedased suurima tõepära hinnangutele uute kujuparameetrite korral. Kasutades sellist algväärtustamist vähendame me EM algoritmi iteratsioonide arvu märkimisväärselt.

Lisaks muudame algväärtusi tuginedes Verbeleni (2013) töös toodud kirjeldusele ning võrdleme sobivuse headust erinevate algväärtuste korral. Nimelt kasutame kujuparameetrite algväärtustamisel nö kattefaktorit $s$ ja anname kujuparameetritele väärtused $r_j = s \cdot j$, kus $j = 1, \ldots, M$. Sääraselt toimides võimaldame me algoritmil läbida suuremat osa parameeterruumist. Erinevate algväärtustega saadud tulemusi võrdleme omavahel IC baasil.

## 3.5 Näiteid jaotuste lähendamisest Erlangi jaotuste segudega

Järgnevalt näitlikustame eelnevalt kirjeldatud EM algoritmi koos protseduuridega kujuparameetrite kohandamiseks, jaotuste arvu (segus) valikuks ning algväärtuste muutmiseks. Selleks genereerime juhuslikke andmeid pidevatest mittenegatiivsetest jaotustest ning hindame Erlangi jaotuste segu parameetrid. Seejärel uurime saadud tulemusi.

EM algoritm koos vajalike protseduuridega on realiseeritud statistikapaketi R funktsioonidena ning toodud Lisas 1. Funktsioonide kirjutamisel oli suureks abiks Roel Verbeleni (2013) magistritöös toodud programmikood, eriliselt kasulikuks osutus seal kasutatud funktsioon *outer()*, millega endal varasemad kokkupuuted puudusid. Nimelt lubab viimane leida erinevate väärtuste ning parameetrite korral tihedusfunktsiooni väärtused maatriksina ning kuna maatriksarvutused on R-s kiiremad kui tsüklite läbimine, võidame märkimisväärselt arvutusteks kuluvas ajas. Eriti vajalikuks osutub see töö peamist eesmärki, milleks on Eesti Liikluskindlustuse Fondist saadud kahjude andmetele Erlangi jaotuste segu sobitamine, silmas pidades. Antud andmestikus on 39 306 vaatlust ja seega on väga vajalik, et programmi tööaeg oleks optimaalselt väike.

Käesolevate näidete puhul kasutame informatsioonikriteeriumina Akaike informatsioonikriteeriumit. Iga järgneva näite korral genereerime $n = 2000$ vaatlust etteantud jaotusest. Esialgseks jaotuste arvuks segus valime $M = 10$ ning maksimaalseks kattefaktoriks $s_{\max} = 5$ (leiame hinnangud väärtuste $s = m,\ m = 1,\ldots,5$ korral ja valime neist parima).

Alustame Erlangi jaotusest ja kahe Erlangi jaotuse segust. Sellisel juhul on meil tõde teada ja seega saame kontrollida, kui hästi või halvasti meie algoritm töötab. Esmalt genereerime vaatlused Erlangi jaotusest kujuparameetriga $r = 3$ ja skaalaparameetriga $\theta = 12$. Algoritm hindab parimaks segu, kus on vaid üks komponent. Saadud hinnangud on toodud Tabelis 2. Nagu näeme, töötab antud olukorras algoritm hästi — piisas vaid ühest komponendist ning kujuparameetri hinnang vastab tegelikule väärtusele ja ka skaalaparameetri hinnang on küllaltki täpne.

**Tabel 2.** Parameetrite hinnangud Erlangi jaotusest genereeritud andmete korral

| Parameetrite hinnangud | | |
|---|---|---|
| $r_j$ | $\alpha_j$ | $\theta$ |
| 3 | 1 | 12.173 |

Järgnevalt genereerime vaatlused kahe Erlangi jaotuse segust kujuparameetritega $\boldsymbol{r} = (r_1, r_2) = (5, 25)$, kaaludega $\boldsymbol{\alpha} = (\alpha_1, \alpha_2) = (0.7, 0.3)$ ning skaalaparameetriga $\theta = 1.2$. Ka antud olukorras hindab algoritm hästi (tulemused Tabelis 3). Kui aga genereerida valim kahe Erlangi jaotuse segust, mille korral kujuparameetrite absoluutne vahe on väike, siis algoritm ei suuda tuvastada kahe komponendi olemasolu. Juhul kui me eelmises segus võtaksime teiseks kujuparameetriks $r_2 = 7,$ siis algoritm hindab parimaks ühekomponendilise segu ehk lihtsalt Erlangi jaotuse, mille kujuparameeter on 5 ning skaalaparameetri hinnanguks on 1.36.

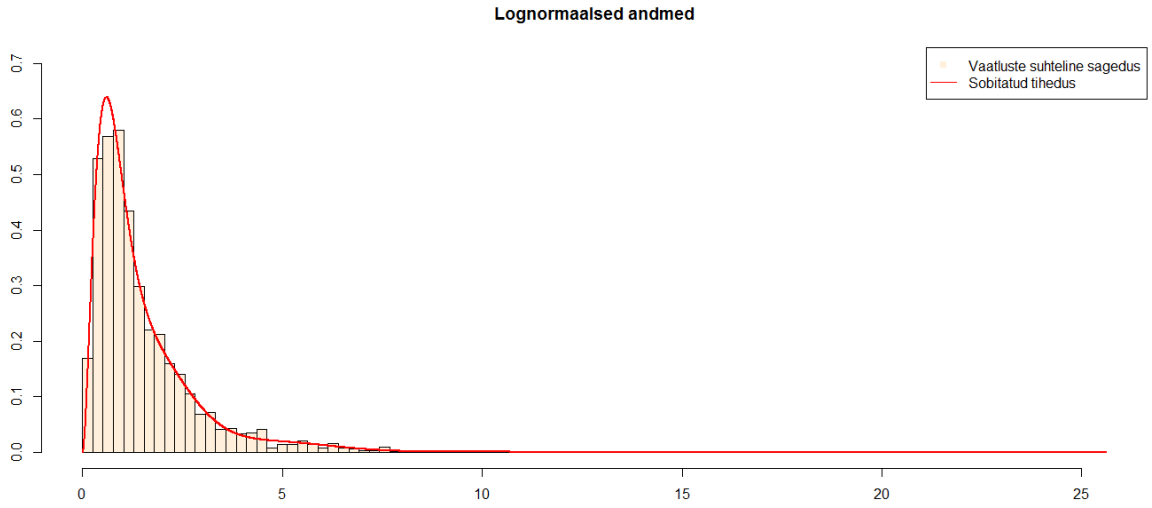
**Tabel 3.** Parameetrite hinnangud kahe Erlangi jaotuse segust genereeritud andmete korral

| Parameetrite hinnangud | | |
|---|---|---|
| $r_j$ | $\alpha_j$ | $\theta$ |
| 5 | 0.69820 | 1.212 |
| 25 | 0.30180 | |

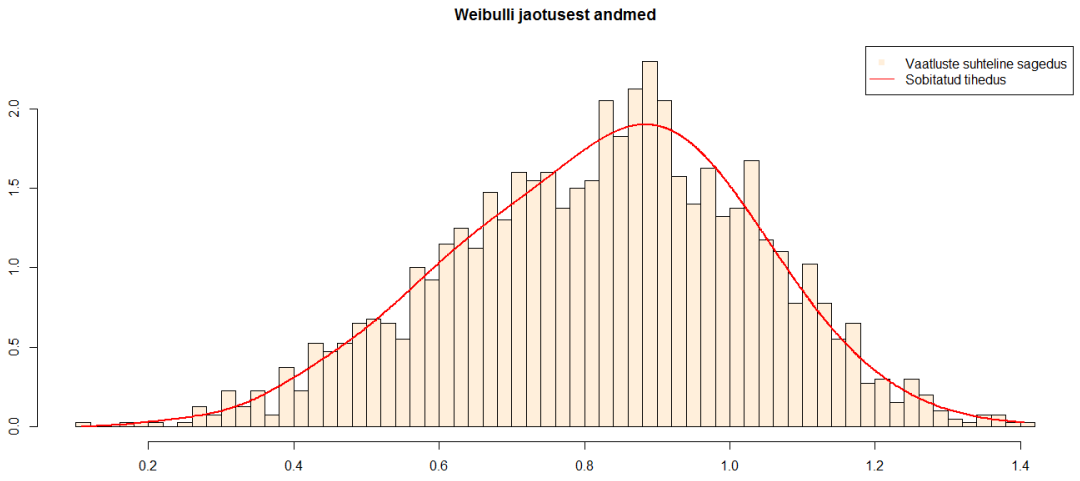
Uurime, kuidas suudab Erlangi jaotuste segu lähendada lognormaalseid ja Weibulli jaotusest andmeid. Selleks genereerime juhusliku valimi jaotustest $lnN(\mu = 0.1, \sigma = 0.8)$ ning $We(k = 4.5, \lambda = 0.9)$. Lognormaalse jaotuse parameetrid valisime tahtlikult sellised, et genereerimisel saaksime ka mingis osas keskmisest märgatavalt suuremaid väärtusi. Eesmärk on imiteerida raskema sabaga andmeid.

Saadud parameetrite hinnangud on toodud Lisas 2. Lognormaalsetele andmetele sobitatud segus on 5 komponenti $(M = 5)$. Nagu juba mainisime, lubasime me kattefaktoril võtta

maksimaalselt väärtuse $s_{\max} = 5$. Parimad väärtused lognormaalse jaotuse korral saime kattefaktoriga $s = 3$. Küllaltki suur komponentide arv on ilmselt põhjustatud justnimelt raskest sabast. Weibulli jaotuse korral on sobitatud segus, mille saime kattefaktoriga $s = 3$, komponentide arvuks $M = 4$.




**Joonis 1.** Lognormaalsest jaotusest genereeritud andmete histogramm ja sobitatud Erlangi jaotuste segu tihedusfunktsioon




**Joonis 2.** Weibulli jaotusest genereeritud andmete histogramm ja sobitatud Erlangi jaotuste segu tihedusfunktsioon

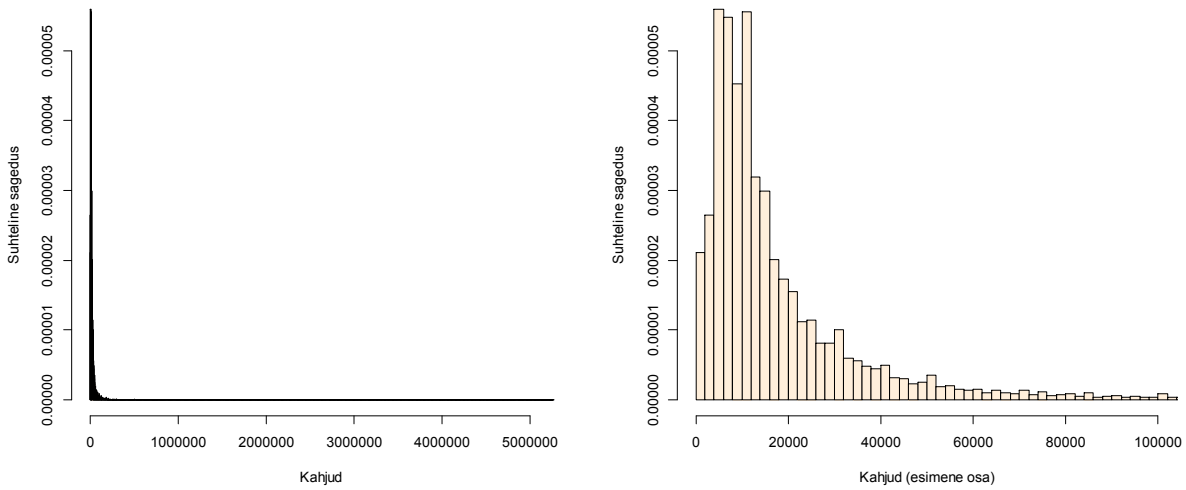
Andmete histogrammid ning sobitatud Erlangi jaotuste segu tihedusfunktsioonid on toodud Joonistel 1 ja 2. Näeme, et sobitatud jaotuste tihedused käituvad andmetele väga sarnaselt nii lognormaalsete kui ka Weibulli jaotusest juhuslike suuruste korral. Lognormaalse jaotuse korral näeme, et kohati imiteerib tihedusfunktsioon andmeid ehk isegi liialt (väärtuse 5 ümbruses) ja meil võib olla tegu ka ülesobitamisega. Samad hinnangud saime me aga ka Schwarzi informatsioonikriteeriumi baasil. Seega võib andmete iseloomust sõltuvalt tekkida vahel vajadus veel rangema otsustuskriteeriumi järele. See on aspekt, mida võiks kindlasti süvenenumalt uurida.

# 4 Liikluskahjudele jaotuste sobitamine

Antud töö peamiseks eesmärgiks on Eesti Liikluskindlustuse Fondist (LKF) saadud liikluskahjude andmetele sobitada Erlangi jaotuste segu. Varasemalt on proovinud samadele andmetele lähendada praktikas kasutatud jaotusi Merili Umbleja oma magistritöös „Kahjude jaotuse ja kindlustuspreemiate hindamine Eestis 2006/2007 toimunud liikluskahjude põhjal“ (Umbleja, 2008). Oma töös käsitles ta gamma-, lognormaalset, Weibulli, Pareto ning beetajaotust. Meie vaatleme neist nelja esimest ning võrdleme saadud tulemusi meie poolt sobitatud Erlangi jaotuste segu korral saaduga.

## 4.1 Andmed

Andmestikus on toodud kahjunõuete suurused ajavahemikust 01.07.2006 – 30.06.2007. Sellel perioodil on esitatud kokku 39306 nõuet. Lisaks kahjude suurustele on iga kahju korral toodud kuupäev ning enamike korral kindlustatud sõiduki liik, vaid 4.74 protsendil vaatlustest ei ole sõidukiliiki märgitud. Kahjude suurused varieeruvad 80-st 5.26 miljoni Eesti kroonini, keskmine kahju on 22 448 krooni, samas suur osa vaatlustest asub vahemikus 5 000 – 15 000 EEK (48.5% vaatlustest). Joonisel 3 on toodud kahjude histogrammid. Vasakpoolne histogramm käsitleb kahjusid kogu ulatuses, kuid kuna see on küllaltki mitteinformatiivne, siis parempoolsel on eraldi ära toodud andmete põhiosa käitumine (kahjud, mis on alla 100 000 krooni). Kui me edasises kujutame graafikutel sobitatud jaotuste tihedusfunktsioone, siis teeme seda samuti 100 000 kroonist väiksemate kahjude korral.




**Joonis 3.** LKF-st saadud liikluskahjude suuruste histogrammid

Histogrammidelt on näha, et kahjud käituvad kindlustusandmetele omaselt — suur osa kahjudest on väiksemad ning on ebaregulaarselt suuri kahjunõudeid, mis venitavad saba pikaks.

## 4.2 Liikluskahjudele jaotuste sobitamine

Umbleja (2008) oli oma töös kahjude suurused teisendanud miljonitesse kroonidesse ning analüüs viidi läbi statistikapaketiga SAS. Antud töö praktiline osa on tehtud vabavaralise statistikapaketiga R. Gamma-, Weibulli ja Pareto jaotuse parameetrid hindame paketis *MASS* sisalduva funktsiooni *fitdistr* abil. Viimane funktsioon kasutab mitme muutuja funktsioonide optimiseerimiseks Broyden-Fletcher-Goldfarb-Shanno (BFGS) algoritmi. SASi protseduur *proc univariate* kasutab selleks aga Newton-Raphsoni meetodit ning seega erinevad osad hinnangud veidi Umbleja töös tooduist. Erinevused on aga küllaltki väikesed. Nimetatud kolme jaotuse parameetrite hindamisel teisendame kahjud miljonitesse, kuna sellisel korral koondub iteratiivne optimiseerimisalgoritm paremini. Pärast hinnangute leidmist teisendame need vastavusse esialgsete andmetega kasutades järgnevaid seoseid:

$$X \sim \Gamma(\alpha,\theta) \Leftrightarrow cX \sim \Gamma(\alpha,c\theta)\ \forall c>0,$$
$$X \sim We(k,\lambda) \Leftrightarrow cX \sim We(k,c\lambda)\ \forall c>0,$$
$$X \sim Pa(\alpha,\lambda) \Leftrightarrow cX \sim Pa(\alpha,c\lambda)\ \forall c>0.$$

Lognormaalse jaotuse parameetrite hindamiseks kasutame samuti funktsiooni *fitdistr*, kuid antud juhul on STP hinnangud analüütiliselt leitavad ning funktsioon kasutab neid teadaolevaid seoseid. Saadud suurima tõepära hinnangud on toodud Tabelis 4.

**Tabel 4.** Gamma-, lognormaalse, Weibulli ja Pareto jaotuse parameetrite hinnangud kogu andmestiku korral

| Jaotus | Parameetrite hinnangud | |
|---|---|---|
| Gamma | $\alpha = 0.95$ | $\theta = 2.38 \cdot 10^4$ |
| Lognormaalne | $\sigma = 1.06$ | $\mu = 9.40$ |
| Weibull | $k = 0.88$ | $\lambda = 2.05 \cdot 10^4$ |
| Pareto | $\alpha = 4.00$ | $\lambda = 6.35 \cdot 10^4$ |

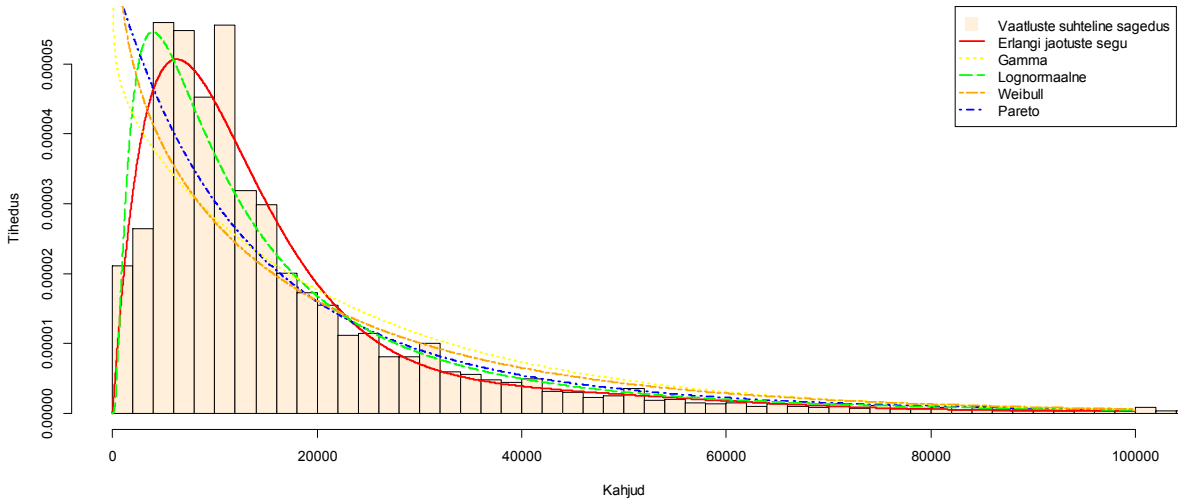
**Tabel 5.** Erlangi jaotuste segu parameetrite hinnangud kogu andmestiku korral

| Parameetrite hinnangud | | |
|---|---|---|
| $r_j$ | $\alpha_j$ | $\theta$ |
| 2 | 0.86329 | 6263.816 |
| 8 | 0.10457 | |
| 20 | 0.02321 | |
| 41 | 0.00675 | |
| 84 | 0.00162 | |
| 188 | 0.00046 | |
| 564 | 0.00010 | |

Erlangi jaotuste segu korral saame suurima tõepära hinnangud vastavalt Peatükis 4 kirjeldatule. Parameetrite hinnangud leidsime 2 korda — Erlangi jaotuste arvu valikul ning erinevate algväärtuste korral saadud hinnangute võrdlemiseks kasutasime nii Akaike kui ka Schwarzi informatsioonikriteeriumeid. Hinnangute saamiseks kuluv aeg (koos

kõikide vajalike protseduuridega) oli tundides. Mõlema hindamisprotsessi käigus jõudsime aga täpselt samade hinnanguteni. Tulemused on toodud Tabelis 5.

Parimad hinnangud saime kattefaktori $s = 5$ korral ning Erlangi jaotuste arv segus on $M = 7$. Kokku on seega antud Erlangi jaotuste segul 15 parameetrit. Paneme tähele, et suurematele kujuparameetritele vastavad väiksemad kaalud. Selline tulemus on oodatav andmete iseloomust tingituna — suurematel vaatlustel on valimis väiksem osakaal.




**Joonis 4.** Kogu andmestikule sobitatud jaotuste tihedusfunktsioonid

Joonisel 4 on toodud sobitatud jaotuste tihedusfunktsioonid. Esmasel visuaalsel vaatlusel tundub, et sobivaimad kandidaadid võiksid olla lognormaalne jaotus ning Erlangi jaotuste segu. Teiste jaotuste sobivus väiksemate kahjude, mis aga moodustavad suure osa andmetest, korral on märgatavalt halvem. Ka Umbleja (2008) hindas visuaalselt sobivaimaks just lognormaalse jaotuse. Samuti viis ta läbi Kolmogorov-Smirnovi ning hii-ruut testi kontrollimaks, kas andmed võiksid olla leitud teoreetilistest jaotustest. Ükski teoreetiline jaotus antud teste ei läbinud. Käesolevas töös on testide korral olulisuse nivooks 0.05. Erlangi jaotuste segu korral saame Kolmogorov-Smirnovi testi teststatistiku väärtuseks $D = 0.0305$ ja olulisustõenäosus on nii väike, et R ei erista seda arvutuslikust nullist. Märgime siinkohal ära, et teiste jaotuste korral on teststatistikute väärtused veel suuremad. Antud juhul on meil tegu aga väga suure valimiga ja seega võivad ka võrdlemisi väikesed erinevused empiirilise ja teoreetilise jaotusfunktsiooni väärtustes nullhüpoteesi (andmed on etteantud teoreetilise jaotusega) ümber lükata.

Tähelepanu tuleb pöörata ka sellele, et antud olukorras teeme Kolmogorov-Smirnovi teste hinnatud parameetrite korral. Selline lähenemine ei ole teoreetiliselt korrektne ja saadud olulisustõenäosused on tegelikest väärtustest suuremad. Teststatistiku tulemusi saab kasutada esmaseks võrdluseks, kuid statistiliste otsuste tegemine jaotuste sobivuse kohta ei ole õigustatud. Kui aga olulisustõenäosused on etteantud olulisuse nivoost väiksemad, siis võime väita, et andmed ei ole antud jaotusest. Hii-ruut testi tegemiseks jaotame andmed 20-ks sama tõenäosusega vahemikuks. Teststatistiku väärtuseks saame $\chi^2 = 1102.2$ ning vastav kriitiline väärtus (vabadusastmete arv on 5) on 11.07. Seega lükkavad mõlemad testid ümber väite, et andmete jaotuseks võiks olla eeltoodud Erlangi jaotuste segu.

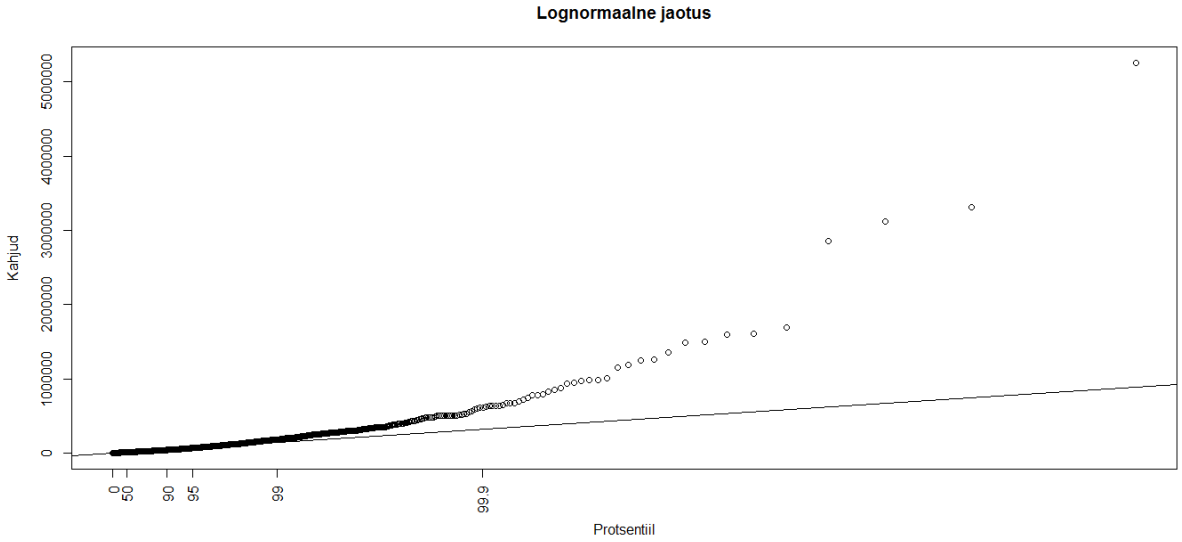
Milline antud jaotustest sobib aga andmetele kõige paremini? Sellele küsimusele leiame vastuse Akaike ning Schwarzi informatsioonikriteeriumi väärtustele tuginedes. Viimased kirjeldavad sobivust läbi logaritmilise tõepära ning samas võtavad arvesse ka jaotuse parameetrite arvu. Viimane on eriliselt tähtis just antud olukorras, kus me võrdleme kohati küllaltki suure parameetrite arvuga jaotuste segusid jaotustega, millel on kaks parameetrit. Informatsioonikriteeriumite väiksemad väärtused vastavad paremale sobivusele. Tabelist 6 näeme, et mõlemad kordajad on vähimad just Erlangi jaotuste segu korral. Seega on parimaks lähendiks Erlangi jaotuste segu ning oodatult järgneb lognormaalne jaotus.

**Tabel 6.** Informatsioonikriteeriumite väärtused kogu andmestiku korral

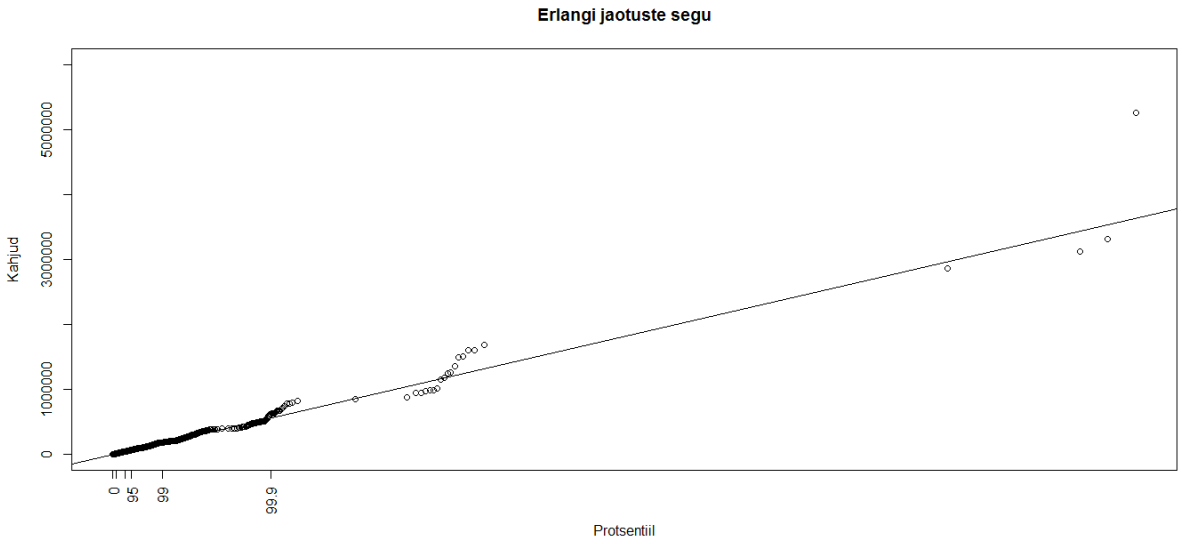
| Jaotus | Gamma | Lognormaalne | Weibull | Pareto | Erlangi jaotuste segu |
|---|---|---|---|---|---|
| AIC | 866 143 | 855 400 | 864 399 | 858 631 | **853 193** |
| BIC | 866 161 | 855 417 | 864 416 | 858 648 | **853 322** |

Uurime lähemalt kahte paremini sobivat jaotust, milleks on Erlangi jaotuste segu ja lognormaalne jaotus. Selleks toome ära kvantiil-kvantiil graafikud (Joonised 5 ja 6). Ordinaatteljel kujutame valimi kvantiile ning abstsissteljel sobitatud jaotuse teoreetilisi kvantiile. Hea sobivuse korral peaksid saadud punktid asuma sirgel $y(x) = x$. Erlangi

jaotuste segude korral ei ole võimalik kvantiile analüütiliselt leida. Küll aga on see võimalik arvutuslikult. Me leiame soovitud $p-$kvantiilid võrrandite $F(x;\boldsymbol{\alpha},\boldsymbol{r},\theta)-p=0$ lahenditena. Seega fikseerime teatud hulga tõenäosusi ja lahendame iga tõenäosuse korral võrrandi. Kasutades interpoleerimist saame leida ka vahepealsed kvantiilid. Kuna antud lähenemine võib mitmel juhul kasulikuks osutuda, on programmikood toodud Lisas 3.




**Joonis 5.** Sobitatud lognormaalse jaotuse kvantiil-kvantiil graafik kogu andmestiku korral




**Joonis 6.** Sobitatud Erlangi jaotuste segu kvantiil-kvantiil graafik kogu andmestiku korral

Ka kvantiil-kvantiil graafikud kinnitavad Erlangi jaotuste segu paremat sobivust, seda nii andmete põhi- kui ka sabaosas. Parem sobivus sabaosas on esmapilgul kohati üllatav, sest Erlangi jaotuste segu korral on sisuliselt tegemist gammajaotuste seguga ja gammajaotus ise on kergema sabaga kui lognormaalne. Siin tuleb aga mängu jaotuste segu dünaamilisus. Kuna andmete põhiosas on väga palju vaatlusi, siis lognormaalse jaotuse parameetrite hindamisel STP meetodil leitakse nad selliselt, et hea sobivus oleks andmetega põhiosas ja selle arvelt tehakse järelandmisi sobivuses sabaosas. Erlangi jaotuste segu korral saame aga põhi- ning sabaosa lähendada erinevate komponentidega.

Saadud tulemused viitavad sellele, et Erlangi jaotuste segul võib olla praktiline väärtus kindlustusandmete modelleerimisel, mille korral üldjuhul on väga tähtis just sobivus sabaosas. See, et meie käsutuses olevate kahjude andmete korral Erlangi jaotuste segu hästi sobib, ei tähenda veel seda, et ta üleüldiselt sobib hästi kahjude suurusi lähendama. Tulemused viitavad aga sellele, et Erlangi jaotuste segu kasutamine kahjude modelleerimisel võib osutuda kasulikuks ning antud lähenemist võiks põhjalikumalt uurida.

Nagu eelnevalt mainisime, on vaatlustest enamuse korral ära toodud ka sõidukiliik (puudub vaid 4.74 protsendil vaatlustest). Sõidukiliigid on jagatud 8-sse kategooriasse, mis on koos nende osakaaludega toodud Tabelis 7.

**Tabel 7.** Sõidukiliigid ja nende osakaalud (%)

| Sõidukiliik | Osakaal (%) |
|---|---|
| Sõiduautod | 77.08% |
| Väikeveokid (kuni 31.12.1997 ka väikebussid) | 6.86% |
| Veoautod | 4.51% |
| Vedukautod | 3.90% |
| Bussid, trollid, trammid | 1.76% |
| Traktorid | 0.64% |
| Mootorrattad ja motorollerid | 0.44% |
| Haagised | 0.07% |

Järgnevalt proovime modelleerida kahjusid erinevate sõidukiliikide korral. Antud töös uurime eraldi kahte sõidukiliiki, milleks on sõiduautod ning bussid, trammid, trollid. Liigid valisime selliselt, et ühe korral oleks valim suur ning teise korral väike. Kuna kogu andmestiku korral saime nii Akaike kui ka Schwarzi informatsioonikriteeriumit kasutades samad tulemused, siis järgnevas kasutame jaotuste arvu valikul ja erinevate algväärtuste korral saadud tulemuste põhjal otsuste tegemiseks Schwarzi informatsioonikriteeriumit (BIC), mis on konservatiivsem.

## 4.3 Sõiduautod

Sobitatud jaotuste parameetrite hinnangud on toodud Tabelites 8 ja 9. Esmalt paneme tähele, et need on küllaltki sarnased eelnevalt kogu andmestiku korral leitutele. See on aga oodatav, sest sõiduautod moodustavad suure osa kõigist sõidukitest. Väikseima informatsioonikriteeriumi väärtusega Erlangi jaotuste segu saime kattefaktoriga $s = 10$, jaotuste arv segus on 8. Esimeste komponenttiheduste kujuparameetrid on väga lähedased kogu andmestikule sobitatud segu omadele, viimane on aga pea kaks korda väiksem. Ilmselt on selle põhjuseks see, et just sõiduautode poolt tekitatud kahjud moodustavad suure osa väiksematest kahjudest, aga samas võiks sõiduautodel olla vähem väga suuri kahjusid, sest kahjude suurus on piiratud sõiduki väärtuse poolt ning sõiduautod on üldjuhul odavamad kui näiteks veoautod või traktorid.

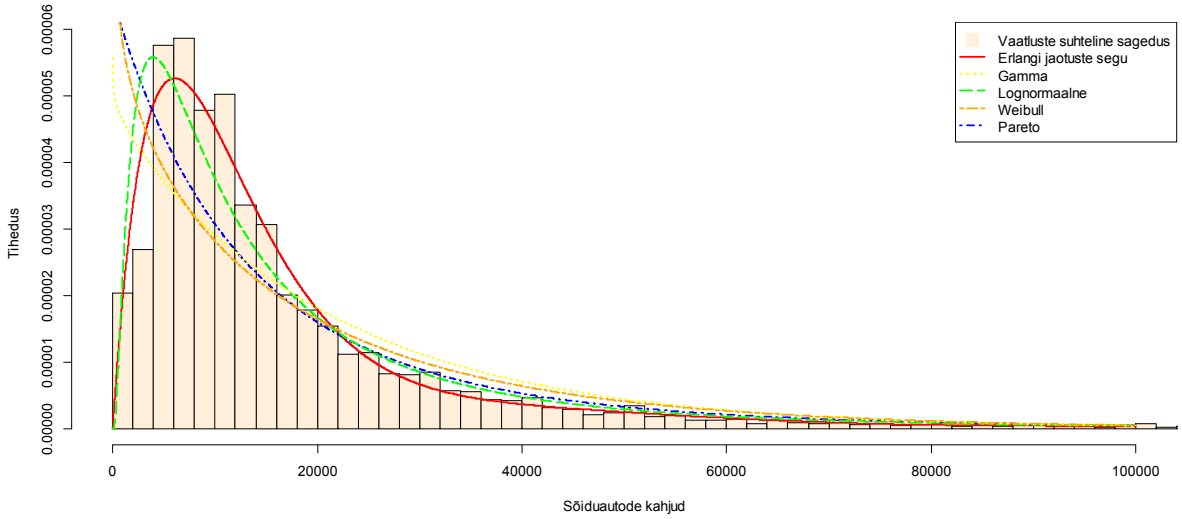
**Tabel 8**. Gamma-, lognormaalse, Weibulli ja Pareto jaotuse parameetrite hinnangud sõiduautode kahjude korral

| Jaotus | Parameetrite hinnangud | |
|---|---|---|
| Gamma | $\alpha = 0.98$ | $\theta = 2.17 \cdot 10^4$ |
| Lognormaalne | $\sigma = 1.05$ | $\mu = 9.38$ |
| Weibull | $k = 0.90$ | $\lambda = 1.98 \cdot 10^4$ |
| Pareto | $\alpha = 4.28$ | $\lambda = 6.63 \cdot 10^4$ |

**Tabel 9** Erlangi jaotuste segu parameetrite hinnangud sõiduautode kahjude korral

| Parameetrite hinnangud | | |
|---|---|---|
| $r_j$ | $\alpha_j$ | $\theta$ |
| 2 | 0.86722 | 6058.878 |
| 8 | 0.09980 | |
| 18 | 0.02106 | |
| 31 | 0.00690 | |
| 50 | 0.00323 | |
| 86 | 0.00113 | |
| 156 | 0.00046 | |
| 285 | 0.00020 | |

Nii Kolmogorov-Smirnovi kui ka hii-ruut test lükkavad kõigi teoreetiliste jaotuste sobivuse ümber ka sõiduautode korral. Erlangi jaotuste segu teststatistikute väärtused kahe uuritava sõidukiliigi korral on toodud Lisas 4, teiste jaotuste korral tehtud testide tulemusi saab huviline vaadata Umbleja (2008) tööst. Märgime siinkohal veelkord ära, et hii-ruut testi tegemiseks jaotasime andmed 20-ks sama tõenäosusega vahemikuks (sarnaselt teeme ka busside, trollide ja trammide kahjude korral järgnevas peatükis).

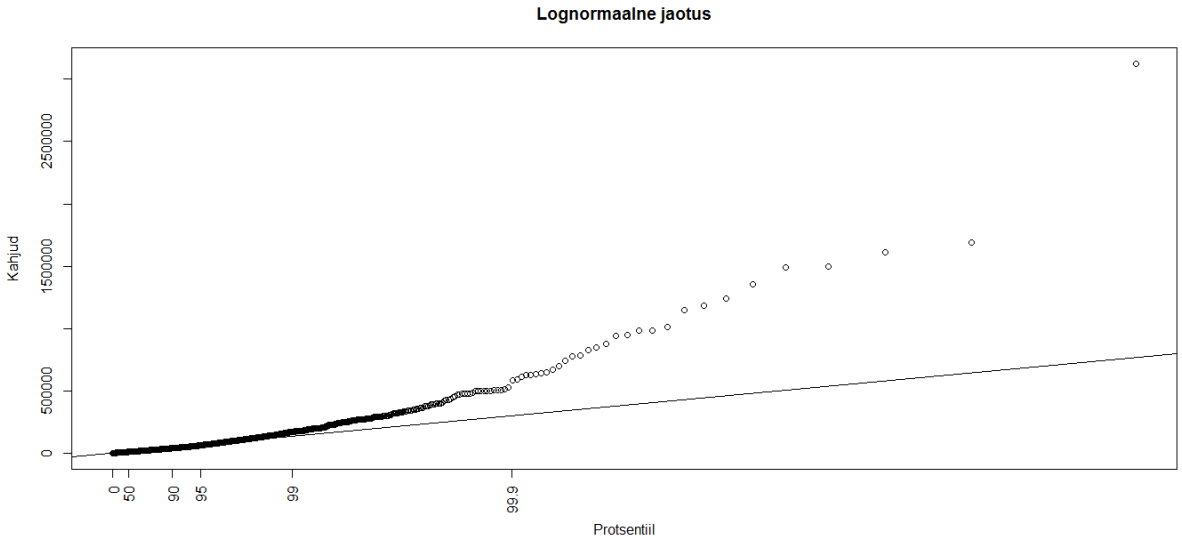



**Joonis 7.** Sõiduautode kahjudele sobitatud jaotuste tihedusfunktsioonid

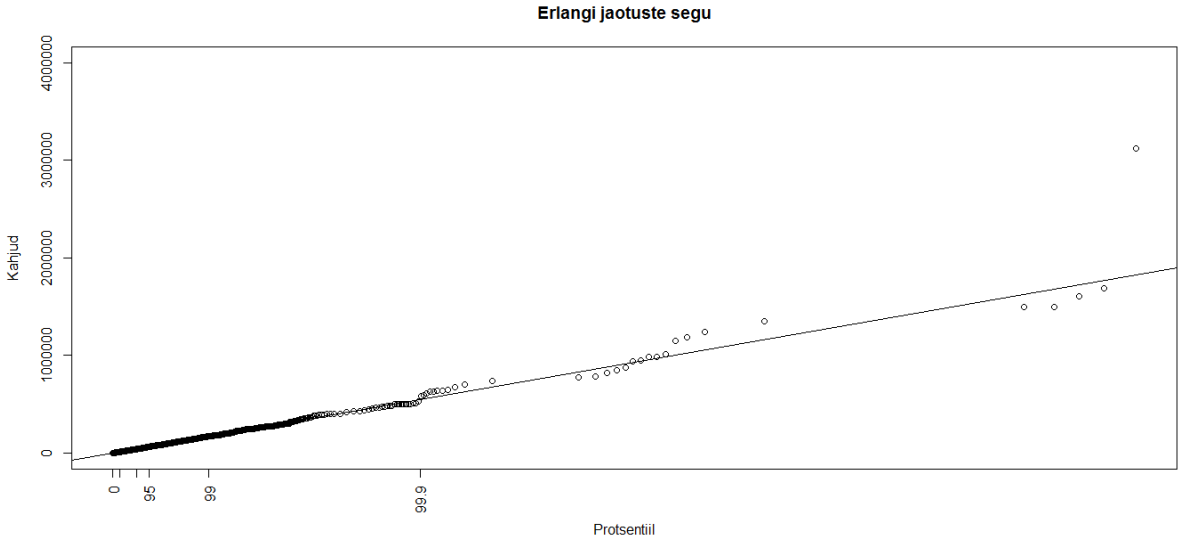
**Tabel 10.** Informatsioonikriteeriumite väärtused sõiduautode kahjude korral

| Jaotus | Gamma | Lognormaalne | Weibull | Pareto | Erlangi jaotuste segu |
|---|---|---|---|---|---|
| AIC | 664 439 | 656 907 | 663 488 | 659 427 | **654 835** |
| BIC | 664 455 | 656 924 | 663 505 | 659 444 | **654 977** |

Joonisel 7 on toodud sobitatud jaotuste tihedusfunktsioonid. Ka siin on pilt küllaltki sarnane eelnevas osas toodule. Viimast joonist täiendavad hästi informatsioonikriteeriumite väärtused (Tabel 10). Nimelt viitavad mõlemad, et andmetele sobivuse vaatenurgast on parim Erlangi jaotuste segu, järgnevad lognormaalne, Pareto, Weibulli ja gammajaotus.




**Joonis 8.** Sobitatud lognormaalse jaotuse kvantiil-kvantiil graafik sõiduautode kahjude korral




**Joonis 9.** Sobitatud Erlangi jaotuste segu kvantiil-kvantiil graafik sõiduautode kahjude korral

Ka sõiduautode korral on kvantiil-kvantiil graafikutelt (Joonised 8 ja 9) näha Erlangi jaotuste segu paremat sobivust antud andmetele nii väiksemate kui ka suuremate väärtuste korral. Kõik tulemused viitavad sellele, et Erlangi jaotuste segu lähendab sõiduautode kahjusid vaadeldavatest kandidaatjaotustest kõige paremini.

## 4.4 Bussid, trollid, trammid

Järgnevalt proovime jaotusi sobitada busside, trollide ning trammide poolt tekitatud kahjudele. Antud sõidukiliigi valisime uurimaks Erlangi jaotuste sobivust ka väiksemate valimite korral. Kindlasti pakub huvi, kas ka väiksema vaatluste arvu korral saame Erlangi jaotuste seguga parema lähendi kui teiste jaotustega.

**Tabel 11.** Gamma-, lognormaalse, Weibulli ja Pareto jaotuse parameetrite hinnangud busside, trollide ja trammide kahjude korral

| Jaotus | Parameetrite hinnangud | |
|---|---|---|
| Gamma | $\alpha = 0.84$ | $\theta = 2.70 \cdot 10^4$ |
| Lognormaalne | $\sigma = 1.19$ | $\mu = 9.33$ |
| Weibull | $k = 0.84$ | $\lambda = 2.02 \cdot 10^4$ |
| Pareto | $\alpha = 3.34$ | $\lambda = 5.08 \cdot 10^4$ |

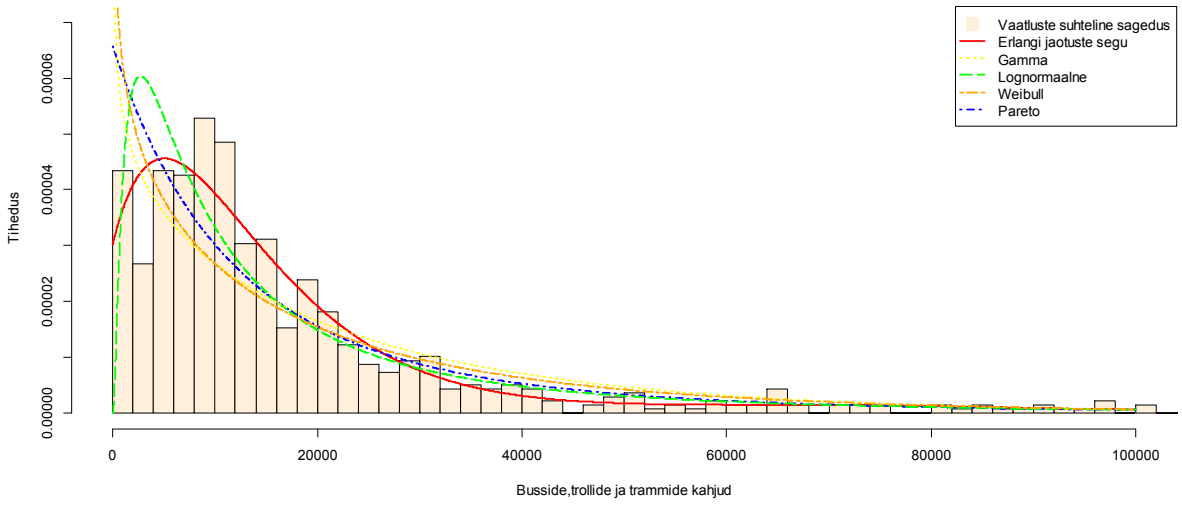
**Tabel 12.** Erlangi jaotuste segu parameetrite hinnangud busside, trollide ja trammide kahjude korral

| Parameetrite hinnangud | | |
|---|---|---|
| $r_j$ | $\alpha_j$ | $\theta$ |
| 1 | 0.23247 | 7682.707 |
| 2 | 0.67467 | |
| 10 | 0.07710 | |
| 29 | 0.01287 | |
| 80 | 0.00289 | |

Parameetrite hinnangud on toodud Tabelites 11 ja 12. Vähima BIC väärtusega hinnangud Erlangi jaotuste segu parameetritele saime kattefaktoriga $s = 3$. Nagu varsemalt, on ka siin suurimad kaalud just väiksemate kujuparameetritega komponentidel. Hinnatud Erlangi jaotuse segus on 5 komponenti, suurim kujuparameetri väärtus on 80.

Arvestades et valimi maht on suhteliselt väike (võrreldes kogu andmestiku mahuga), valisime esialgseks Erlangi jaotuste arvuks $M = 30$ ja maksimaalseks kattefaktoriks $s_{\max} = 15$. Et anda mingi ülevaade arvutusteks kuluvast ajast, märgime ära, et antud valimimahu $n = 691$ korral võttis kõikide protseduuride läbimine aega ligikaudu 25 minutit. Kuna aga parimad hinnangud saadi kattefaktori $s = 3$ korral, läbis programm suuremad kattefaktorite väärtused nö kasutult. Valides $s_{\max} = 3$, kulub hinnangute saamiseks umbes 4 minutit. Kuna me aga enne tulemuste saamist ei tea, milliste kattefaktori väärtuste korral hinnangud paranevad, töötabki algoritm kohati „põhjendamatult“ kauem kui tegelikult oleks olnud tarvis. See on kindlasti üks aspekt, mida sügavamalt uurida.

Vaatluste histogramm ning sobitatud jaotuste tihedusfunktsioonid on toodud Joonisel 10. Võrreldes eelmise kahe olukorraga (kogu andmed ja sõiduautod) võime joonise põhjal suurema enesekindlusega eeldada, et Erlangi jaotuste segu võiks anda kandidaatidest parima lähendi. Nii AIC kui ka BIC väärtused (Tabel 13) on segu korral vähimad. Paneme veel tähele, et informatsioonikriteeriumid lognormaalse ning Pareto jaotuse korral on küllaltki lähedased. Arvestades et Pareto jaotus andmeid alguse osas väga hästi ei lähenda, võime eeldada, et Pareto jaotuse sobivus sabaosas on küllaltki hea.

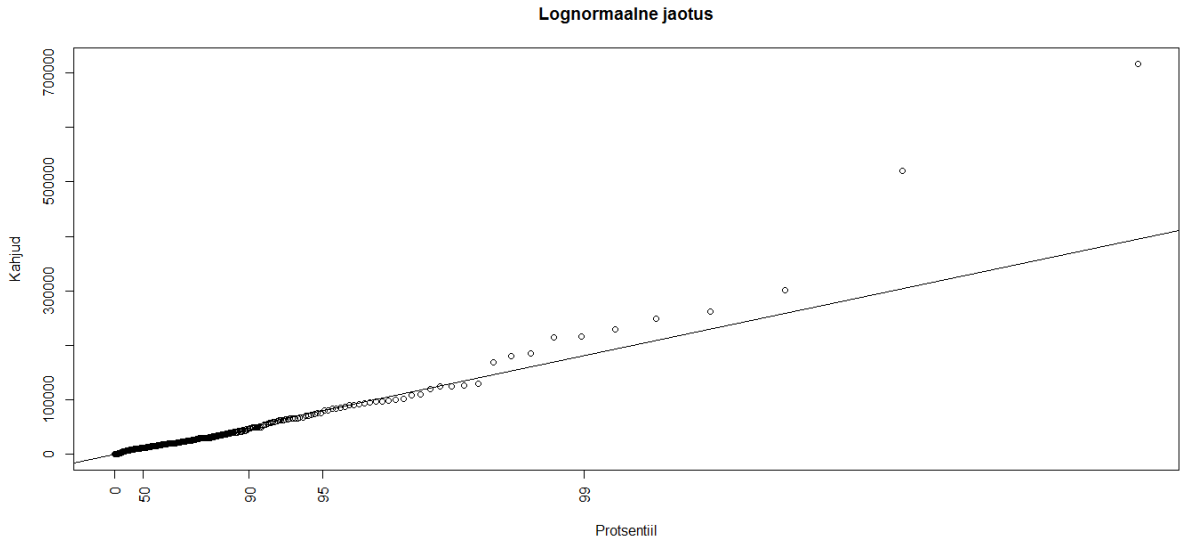



**Joonis 10**. Busside, trollide ja trammide kahjudele sobitatud jaotuste tihedusfunktsioonid

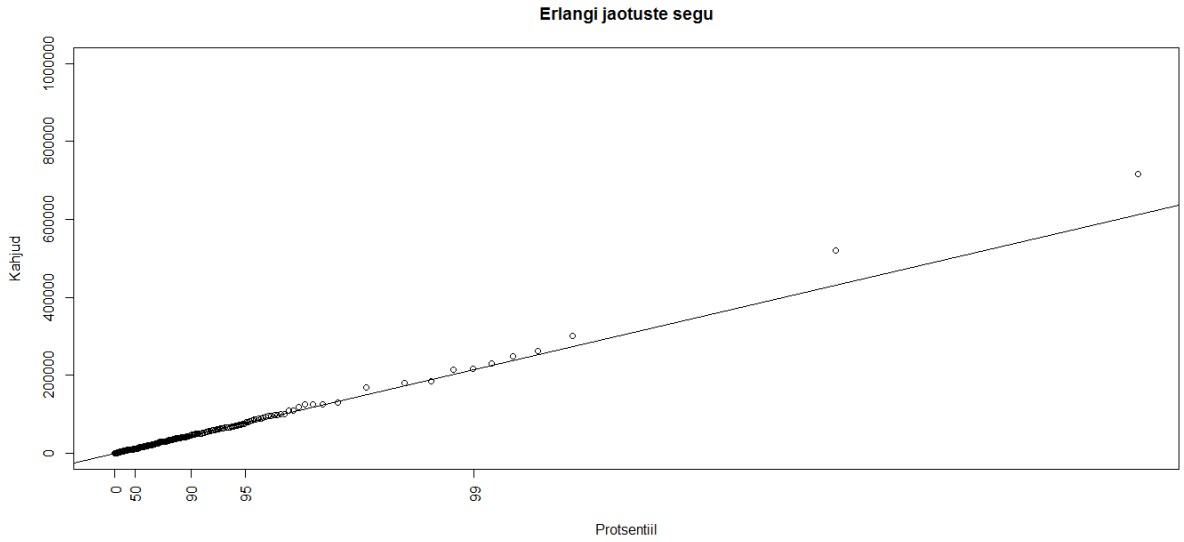
**Tabel 13.** Informatsioonikriteeriumite väärtused busside, trollide ja trammide kahjude korral

| Jaotus | Gamma | Lognormaalne | Weibull | Pareto | Erlangi jaotuste segu |
|---|---|---|---|---|---|
| AIC | 15 235 | 15 105 | 15 198 | 15 108 | **15 058** |
| BIC | 15 244 | 15 115 | 15 207 | 15 117 | **15 108** |

Viisime läbi ka statistilised testid sobivuse kontrollimiseks ning tulemused on toodud Lisas 3. Kolmogorov-Smirnovi testi olulisustõenäoususeks saime $p = 0.2575$. Nagu me aga juba eelnevalt mainisime, ei saa me selle tulemuse põhjal jääda nullhüpoteesi juurde ja eeldada, et andmete jaotuseks on Erlangi jaotuste segu. Hii-ruut testi korral on teststatistiku väärtus suurem kui kriitiline väärtus ja seega võime me pigem väita, et busside, trollid ning trammide kahjud ei ole sobitatud Erlangi jaotuste seguga. Umbleja (2008) töös on toodud testide tulemused ka teiste jaotuste korral. Hii-ruut testi ei läbinud ükski jaotus.




**Joonis 11.** Sobitatud lognormaalse jaotuse kvantiil-kvantiil graafik busside, trollide ja trammide kahjude korral




**Joonis 12.** Sobitatud Erlangi jaotuste segu kvantiil-kvantiil graafik busside, trollide ja trammide kahjude korral

Joonistelt 11 ja 12 näeme, et ka sõidukiliigi bussid, trollid, trammid korral sobib Erlangi jaotuste segu andmetele paremini kui lognormaalne jaotus. Kuigi me Pareto jaotuse korral kvantiil-kvantiil graafikut ära ei too, märgime, et see on küllaltki sarnane lognormaalse jaotuse omale ning suurt paremust sobivuses sabaosas näha ei ole. Nägime, et Erlangi jaotus lähendas andmeid teistest paremini ka väiksema valimimahu korral.

## 4.5 Tulemuste kokkuvõte

Antud peatükis sobitasime Eesti Liikluskindlustuse Fondist saadud kahjude suurustele ühise skaalaparameetriga Erlangi jaotuste segusid ning erinevaid praktikas kasutatud jaotusi. Meie eesmärgiks oli leida jaotus, mis lähendab andmeid kõige paremini.

Me nägime, et ükski jaotustest ei läbinud statistilisi sobivuse teste. Võrdlemaks erinevate jaotuste sobivust, leidsime Akaike ning Schwarzi informatsioonikriteeriumite väärtused. Kõikidel juhtudel saime vähimad väärtused sobitatud Erlangi jaotuste segude korral. Kindlasti jääb õhku küsimus, et mis piirist me loeme lisaparameetri kaasamist õigustatuks ja kas antud informatsioonikriteeriumite korral on trahv parameetrite arvu pealt piisav. See aga sõltub konkreetsest eesmärgist ja üks-ühest vastust pole võimalik anda. Silmas tuleb aga pidada, et antud töös tegelesime me reaalsete kahjudega ning ükski

klassikalistest jaotustest häid tulemusi ei andnud. Antud olukorras ei jäägi aktuaaril muud üle kui keerulisemaid lähenemisi kaaluda.

Saadud tulemuste seas tõusis esile üks väga atraktiivne aspekt, nimelt saime Erlangi jaotuste segude korral küllaltki hea sobivuse kahjude sabaosas (ja seda tegemata järelandmisi sobivuses põhiosas). Hea lähend suurtele kahjudele on kindlustuses kriitilise tähtsusega. Kui me sobitame andmetele liiga kerge sabaga jaotuse ning arvutame selle põhjal preemiad, siis mustema stsenaariumi korral ei pruugi meil enam isegi reservist piisata.

Kahjudele Erlangi jaotuste segude sobitamist võiks kindlasti tulevikus põhjalikumalt uurida. On mitmeid aspekte, mis vajaksid tähelepanu ja sügavamat analüüsi. Praktilisest vaatenurgast huvitab meid kindlasti, kas kuidagi oleks võimalik hinnangute leidmist ajaliselt optimiseerida. Tähelepanu tuleks pöörata ka ülesobitamise vältimisele ja erinevate otsustuskriteeriumite rakendamisele, et seda teha.

# Kasutatud kirjandus

[1] Box, G.E.P. ja Draper, N.R. (1987). *Empirical model-building and response surfaces*. New York: Wiley.

[2] Dempster, A.P., Laird, N.M. ja Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1), 1-38.

[3] Gray, R.J. ja Pitts, S.M. (2012). *Risk modelling in general insurance*. Cambridge: Cambridge University Press.

[4] Lee, S.C.K. ja Lin, X.S. (2010). Modeling and evaluating insurance losses via mixtures of Erlang distributions. *North American Actuarial Journal*, 14(1), 107-130.

[5] McLachlan, G.J. ja Krishnan, T. (2008). *The EM algorithm and extensions*. 2nd ed. Hoboken, New Jersey: Wiley.

[6] McLachlan, G.J. ja Peel, D. (2001). *Finite mixture models*. New York: Wiley.

[7] Tijms, H.C. (1994). *Stochastic models, an algorithmic approach*. Chichester: Wiley.

[8] Umbleja, M. (2008). Kahjude jaotuse ja kindlustuspreemiate hindamine Eestis 2006/2007 toimunud liikluskahjude põhjal. Msc. Tartu Ülikool.

[9] Verbelen, R. (2013). *Phase-type distributions & mixtures of Erlangs*. MSc. KU Leuven.

# Lisa 1 EM algoritmi programmikood

```
library(MASS)

Jaotus_erlang=function(x,r,alpha,theta){
    jaotused=outer(x,r,pgamma,scale=theta)
    a=t(t(jaotused)*alpha)
    jaotus=rowSums(a)
    return(jaotus)
}

Erlangite_tihedus=function(x,r,alpha,theta){
    tihedused=outer(x,r,dgamma,scale=theta)
    kaalutud_tih=t(t(tihedused)*alpha)
    return(rowSums(kaalutud_tih))
}

Tiheduse_kaalutud_komp=function(x,r,alpha,theta){
    tihedused=outer(x,r,dgamma,scale=theta) #teeb nö risttabeli
    kaalutud_tih=t(t(tihedused)*alpha) #kaalumine
    return(kaalutud_tih)
}

ErlangiteSegu_logtoepara=function(kaalutud_komp){
    vaatlus=rowSums(kaalutud_komp) #vaatluse panus
    log_vaatlus=ifelse((vaatlus>0), log(vaatlus),-1000)
    return(sum(log_vaatlus))

}

E_samm=function(kaalutud_komp,l){
    z= kaalutud_komp/rowSums(kaalutud_komp)#nimetajaga läbi jagamine
    z[is.nan(z)]=1/l
    return(z)
}

algvaartustamine=function(x,M,katte_fak){
    n=length(x)
    r=seq(1,M)
    alpha=rep(NA,M)

    #ALGVÄÄRTUSED:
    theta=max(x)/M
    for (i in 1:M){
        alpha[i]=sum(theta*(i-1)<x&x<=theta*i)/n
    }
    #EEMALDAME KUJUPARAMEETRID, MILLE KORRAL VASTAVA ERLANGI KAAL ON 0
    #NING SAMUTI KAALUD
    r=r[alpha>0]
    #et arvutustäpsus ei segaks kaalude summeerumist 1-ks:
```

```
    alpha=alpha[alpha>0]/sum(alpha)
    #Kasutame kattefaktorit:
    r=r*katte_fak
    theta=theta/katte_fak #algväärtustamine korrektne
    return(list(r=r,alpha=alpha,theta=theta))
}



EM=function(x,r,alpha,theta,epsilon){   #x on kasutuses olevad andmed
    n=length(x)
    kaalutud_komp=Tiheduse_kaalutud_komp(x,r,alpha,theta)
    #log-tõepära algväärtustega:
    log_tp=ErlangiteSegu_logtoepara(kaalutud_komp)
    log_tp_vana=-Inf #vajalik, et while tsükkel alustaks
    toeparad=log_tp #hoiame kontrolli mõttes ka tõepärad meeles
    #ALGORITMI PÕHIOSA
    while (log_tp-log_tp_vana>epsilon){
        log_tp_vana=log_tp
        #E-SAMM:
        l=length(r)
        #jareltõenäosus segu kompnenti kuulumiseks:
        z=E_samm(kaalutud_komp,l)
        #M-SAMM:
        alpha=colSums(z)/n
        theta=mean(x)/(sum(alpha*r))
        kaalutud_komp=Tiheduse_kaalutud_komp(x,r,alpha,theta)
        log_tp=ErlangiteSegu_logtoepara(kaalutud_komp) #uus tõepara
        toeparad=c(toeparad,log_tp)
    }
    AIC=2*(length(alpha)+length(r)+1)-2*log_tp
    return(list(alpha=alpha,r=r,theta=theta,log_toepara=log_tp,
    toeparad=toeparad,AIC=AIC))
}

#KUJUPARAMEETRITE KOHANDAMINE:
kujuparameetrid=function(x,r,alpha,theta,epsilon){
    hinnang=EM(x,r,alpha,theta,epsilon)
    r=hinnang$r
    alpha=hinnang$alpha
    theta=hinnang$theta
    #Logtõepärad while tsükli jaoks ehk 1 kohendamise kontrolliks
    log_toepara_vana=-Inf #Vajalik while tsükli esimeseks sammuks
    log_toepara_uus=hinnang$log_toepara
    #Logtõepära vajalik komponendi muutmise kontrolliks (for tsüklites)
    log_tp=log_toepara_uus
    M=length(r)
    while(log_toepara_uus-log_toepara_vana>epsilon){
        log_toepara_vana=log_toepara_uus
        for (i in M:1){
            #tõeväärtus, mis näitab, kas tõepära paranes või mitte:
            tp_paranes=TRUE
            while((tp_paranes==TRUE) && (((r[i+1]-r[i])>1) || (i==M) )
            &&          (alpha[i]>0.0001)){
                r_uus=r
                r_uus[i]=r_uus[i]+1
                hinnang=EM(x,r_uus,alpha,theta,epsilon)
                log_tp_uus=hinnang$log_toepara
                if (log_tp_uus-log_tp>epsilon){
```

```
                    log_tp=log_tp_uus
                    r=r_uus
                    alpha=hinnang$alpha
                    theta=hinnang$theta
                }
                else {tp_paranes=FALSE}
            }
        }
        for (i in 1:M){
            tp_paranes=TRUE
            while((tp_paranes==TRUE) && (r[i]>1) && (((r[i]-r[i-1])>1)
            || (i==1)) && (alpha[i]>0.0001)){
                r_uus=r
                r_uus[i]=r_uus[i]-1
                hinnang=EM(x,r_uus,alpha,theta,epsilon)
                log_tp_uus=hinnang$log_toepara
                if (log_tp_uus-log_tp>epsilon){
                    log_tp=log_tp_uus
                    r=r_uus
                    alpha=hinnang$alpha
                    theta=hinnang$theta
                }
                else {tp_paranes=FALSE}
            }
        }
        log_toepara_uus=log_tp
    }
    AIC=2*(length(alpha)+length(r)+1)-2*log_toepara_uus
    BIC=(length(alpha)+length(r)+1)*log(length(x))-2*log_toepara_uus

return(list(r=r,alpha=alpha,theta=theta,log_toepara=log_toepara_uus,AI
C=AIC,BIC=BIC))
}

#ERLANGITE ARVU VALIK:
M_valik=function(x,r,alpha,theta,epsilon){
    hinnang=kujuparameetrid(x,r,alpha,theta,epsilon)
    BIC=hinnang$BIC
    BIC_d=BIC
    r=hinnang$r
    alpha=hinnang$alpha
    theta=hinnang$theta
    M=length(r)
    BIC_paranes=TRUE #vajalik tsükli alustamiseks
    while((BIC_paranes==TRUE) && M>1){
        eemaldame=which(alpha==min(alpha))#leiame            minimaalse
kaalu/kaalud
        #eemaldame vastava(d)kaalu(d) ja kujuparameetri(d)
        alpha_uus=alpha[-eemaldame]
        r_uus=r[-eemaldame]
        alpha_uus=alpha_uus/sum(alpha_uus)#skaleerime kaalud
        #uued hinnangud
        hinnang=kujuparameetrid(x,r_uus,alpha_uus,theta,epsilon)
        BIC_uus=hinnang$BIC
        if (BIC_uus<BIC){
            r=hinnang$r
            alpha=hinnang$alpha
            theta=hinnang$theta
            BIC=BIC_uus
```

```
            BIC_d=c(BIC_d,BIC)
            M=length(r)
        }
        else{BIC_paranes=FALSE}
    }
    AIC=BIC+(length(alpha)+length(r)+1)*(2-log(length(x)))

return(list(r=r,alpha=alpha,theta=theta,BIC=BIC,AIC=AIC,BIC_d=BIC_d,M=
M))
}


#ALGVÄÄRTUSTE KOHANDAMINE
algvaartuste_kohandamine=function(x,M_alg,max_k,epsilon){
    #max_k - maksimaalne kattefaktor
    k=1 #kattefaktor
    alg=algvaartustamine(x,M_alg,k)
    hinnang=M_valik(x,alg$r,alg$alpha,alg$theta,epsilon)
    BIC=hinnang$BIC
    #Hoiame meeles vähima BIC väärtusega parameetrid
    #Esialgu võtame nendeks kattefaktoriga 1 leitud,
    #kui BIC paraneb, siis muudame
    M=hinnang$M
    r=hinnang$r
    alpha=hinnang$alpha
    theta=hinnang$theta
    for (i in 2:max_k){
        print(i)
        alg=algvaartustamine(x,M_alg,i)
        hinnang=M_valik(x,alg$r,alg$alpha,alg$theta,epsilon)
        BIC_uus=hinnang$BIC
        if (BIC_uus<BIC){ #->paranes, muudame parameetreid
            print("PARANES")
            k=i
            M=hinnang$M
            r=hinnang$r
            alpha=hinnang$alpha
            theta=hinnang$theta
            BIC=BIC_uus
        }
    }
    AIC=BIC+(length(alpha)+length(r)+1)*(2-log(length(x)))

return(list(M=M,r=r,alpha=alpha,theta=theta,kFak=k,BIC=BIC,AIC=AIC))
}
```

# Lisa 2 Parameetrite hinnangud genereeritud andmete korral

**Tabel 14.** Lognormaalsest jaotusest andmetele sobitatud Erlangi jaotuste segu parameetrite hinnangud

| Parameetrite hinnangud | | |
|---|---|---|
| $r_j$ | $\alpha_j$ | $\theta$ |
| 3 | 0.70020 | 0.298 |
| 8 | 0.24436 | |
| 18 | 0.05067 | |
| 36 | 0.00377 | |
| 74 | 0.00100 | |

**Tabel 15.** Weibulli jaotusest andmetele sobitatud Erlangi jaotuste segu parameetrite hinnangud

| Parameetrite hinnangud | | |
|---|---|---|
| $r_j$ | $\alpha_j$ | $\theta$ |
| 13 | 0.00957 | 0.022 |
| 22 | 0.08965 | |
| 31 | 0.29756 | |
| 42 | 0.60322 | |

# Lisa 3 Kvantiilide leidmise programmikood

```
#Lahendab võrrandeid F(x)-p=0:
F_inv=function(p,r,alpha,theta,br=c(0,10**8))
{
   G=function(x){return(Jaotus_erlang(x,r,alpha,theta)-p)}
   return(uniroot(G,br)$root)
}

#Järgnev osa lahendab võrrandi punktides ja
#interpoleerib, et saada ka vahepealsed väärtused
#tagastab sisuliselt sarnase funktsiooni nagu qnorm,
#st saab suvalise argumendi anda

#Loodud otseselt q-q jooniste tegemiseks, võrrandid lahendab sobivates
punktides
QuantileErlang=function(data,r,alpha,theta){
    p=(1:length(data))/(length(data)+1)
    kvantiilid=rep(NA,length(p))
    for(i in 1:length(kvantiilid)){
        kvantiilid[i]=F_inv(p[i],r,alpha,theta,br=c(0,10**8))
    }
    Q=approxfun(x=p,y=kvantiilid)
    return(Q)
}
#Kutsume välja, tulemuseks saame funktsiooni, mis sisendina tahab vaid
#vastavat tõenäosust ja annab meile kvantiili
Q=QuantileErlang(kahjud_kogu,r=r,alpha=alpha,theta=theta)
```

# Lisa 4 Testide tulemused vaadeldud sõidukiliikide korral

**Tabel 16.** Kolmogorov-Smirnovi ja hii-ruut testi tulemused sõidukiliikide sõiduautod ning bussid, trollid ja trammid korral

| Sõidukiliik | Kolmogorov-Smirnovi test | | Hii-ruut test | |
|---|---|---|---|---|
| | Teststatistik | Olulisustõenäosus | Statistiku väärtus | Kriitiline väärtus |
| Sõiduautod | $D = 0.0354$ | $p < 0.0001$ | $\chi^2 = 814.88$ | 7.81 (df=3) |
| Bussid, trollid, trammid | $D = 0.0385$ | $p = 0.2575$ | $\chi^2 = 61.60$ | 16.92 (df=9) |

**Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks**

Mina, Kristjan Kokorev ,
(*autori nimi*)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose

Erlangi jaotuste segude sobitamine kindlustuskahjudele ,
(*lõputöö pealkiri*)

mille juhendaja on Meelis Käärik ,
(*juhendaja nimi*)

1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;
1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace´i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.

3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus, **13.05.2015**