MATEMAATIKA

Mitmemõõtmelistest jaotustest

Tõnu Kollo
Tartu Ülikool

Ülevaade tööde tsüklist *„Mitmemõõtmelised maatrikstehnikal põhinevad statistikamudelid"*, mis sai Eesti Vabariigi 2007. aasta teaduspreemia täppisteaduste alal.

1. Eellugu, tähistused

Matemaatilise statistika eesmärk on välja töötada meetodeid selleks, et teha vaatluste põhjal järeldusi ja otsustusi meid ümbritseva maailma kohta. Vaatlused moodustavad valimi, kusjuures vaatluste arvu n nimetatakse valimimahuks. Eeldame, et uuritavat nähtust kirjeldab p tunnusest koosnev juhuslik vektor $\mathbf{x}$, mille tõenäosusjaotus annab meile informatsiooni vaatluste võimalikust käitumisest: kui suure tõenäosusega paiknevad $\mathbf{x}$ väärtused mingis piirkonnas. Valimi tõenäosuslikku käitumist kirjeldab juhuslik maatriks $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)'$, kus $\mathbf{x}_i$ on sõltumatud vektoriga $\mathbf{x}$ sama jaotusega juhuslikud vektorid, $\mathbf{x}_i \sim \mathbf{x}$. Klassikalise eelduse kohaselt on vektor $\mathbf{x}$ normaaljaotusega: $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, kus parameeter $\boldsymbol{\mu}$ on keskväärtusvektor ja $\boldsymbol{\Sigma}$ dispersioonimaatriks, mis kirjeldab juhusliku vektori $\mathbf{x}$ hajuvust keskväärtuse suhtes.

On mitmeid nähtusi, mida normaaljaotusega tunnusvektor hästi kirjeldab. Aga väga tihti on normaaljaotuse eeldus kunstlik. Peapõhjus tema kasutamiseks on saadava matemaatilise mudeli lihtsus ja elegantsus ning omaduste põhjalik läbitöötatus. Samal ajal

võivad kasutatavad andmed normaaljaotusest tunduvalt erineda. Teatavasti aga valedest eeldustest lähtumine võib viia valede tulemusteni. Viimaste aastakümnete üks peamisi arengutendentse matemaatilises statistikas on suunatud normaaljaotuse eeldusest vabanemisele. Selle saavutamiseks on mitmeid võimalusi, neist kolmele pöörame järgnevalt tähelepanu:

a) jaotuste lähendamine reaksarendustena,

b) mitmemõõtmeliste ebasümmeetriliste jaotusperede kasutamine,

c) koopulate teooria rakendamine.

Rahvusvahelise Tõenäosusteooria ja Matemaatilise Statistika Bernoulli Ühingu president Wilhelm van Zwet sõnastas 1986. aastal ühingu I kongressil kõige olulisema statistika arengusuunana asümptootiliste meetodite arendamise. Millega on tegemist?

Ka normaaljaotusest erineva jaotuse korral on keskväärtusvektor $\boldsymbol{\mu}$ ja dispersioonimaatriks $\boldsymbol{\Sigma}$ informatiivsed: nad iseloomustavad andmete paiknemist ja hajuvust. Valimist leitud nihketa hinnangud neile suurustele on valimikeskmine

$$\bar{\mathbf{x}} = \sum_{i=1}^{n} \mathbf{x}_i$$

ja valimi dispersioonimaatriks

$$\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})',$$

kus $'$ tähistab transponeerimist. Nende hinnangute suurepärane omadus on see, et ka siis, kui vaatlused ei ole normaaljaotusega, on $\bar{\mathbf{x}}$ ja $\mathbf{S}$ ikkagi kirjeldatavad normaaljaotusega, kui valimimaht n on suur. Sama omadus on ka $\bar{\mathbf{x}}$ ja $\mathbf{S}$ praktikas kasutatavatel funktsioonidel. Seetõttu on väga paljude statistikameetodite korral tulemused kirjeldatavad normaaljaotusega suure valimimahu korral. Vastavat jaotust, mis tekib protsessis $n \to \infty$, nimetatakse asümptootiliseks normaaljaotuseks. Valimifunktsioonide asümptootiliste normaaljaotuste leidmisega hakati intensiiv-

selt tegelema 1970-ndatel aastatel, ka Tartus tekkis see uurimissuund 1970-ndate teisel poolel. Paraku ei anna asümptootiline normaaljaotus kaugeltki alati praktikas rakendatavat tulemust. Väga ebasümmeetrilise jaotusega vaatluste korral on koondumine normaaljaotuseks aeglane, samuti sõltub koonduvuskiirus funktsioonist, mille jaotust leitakse. Ka tuhandetesse ulatuv vaatluste arv ei ole mõnikord piisav praktikas kasutatava lähendi saamiseks. Samas on vaja rakendada statistilisi meetodeid ka olukorras, kus valimimaht on mõnikümmend. Tekib küsimus, kas on võimalik korrigeerida asümptootilist normaaljaotust selliselt, et võttes arvesse andmetes leiduva informatsiooni paiknemise, hajuvuse, ebasümmeetria ja keskväärtusest kaugel paiknevate vaatluste osakaalu kohta, saame tunduvalt parema lähendi, kui seda on asümptootiline normaaljaotus ise. Neid omadusi saab kirjeldada ühe tunnuse korral juhusliku suuruse esimeste momentidega. Esimest järku moment võrdub keskväärtusega ja kirjeldab paiknemist:

$$m_1(X) = EX,$$

teist järku tsentraalne moment ehk dispersioon hajuvust:

$$\overline{m}_2(X) = E(X - EX)^2,$$

kolmas ja neljas tsentraalne moment aga iseloomustavad vastavalt asümmeetriat ja järsakust:

$$\overline{m}_3(X) = E(X - EX)^3, \qquad \overline{m}_4(X) = E(X - EX)^4.$$

Osutub, et see idee lähendjaotuse korrigeerimisest on realiseeritav. Juba 1937. aastal konstrueeriti seos kahe juhusliku suuruse tõenäosustiheduste ja jaotusfunktsioonide vahel (Cornish & Fisher, 1937). Tundmatu keeruka jaotusega juhusliku suuruse Y tõenäosustihedus $f_Y(x)$ avaldub tuntud jaotusega juhusliku suuruse X tõenäosustiheduse $f_X(x)$ kaudu reaksarendusena järgmiselt:

$$f_Y(x) = f_X(x) + a_1 f_X(x)^{(1)} + a_2 f_X(x)^{(2)} + a_3 f_X(x)^{(3)} + \cdots, \quad (1)$$

kus $f_X(x)^{(k)}$ tähistab tiheduse $f_X(x)$ k-ndat järku tuletist ja kordaja a_k avaldis sisaldab kuni k-ndat järku momentide vahesid.

See tähendab, et a_1 sisaldab keskväärtuste vahet $EY - EX$, a_2 võtab arvesse dispersioonide erinevuse $DY - DX$, a_3 võtab arvesse ebasümmeetriat sisaldades kolmandat järku tsentraalsete momentide vahet jne.

See, kas saadud tiheduse $f_Y(x)$ esitust $f_X(x)$ kaudu saab ka lähendina kasutada, sõltub konkreetsest juhuslikust suurusest Y. Juhul kui Y rollis on juhuslik suurus, mis on esitatav sõltumatute juhuslike suuruste aritmeetilise keskmisena (näiteks valimimomendid, kaasa arvatud valimikeskmine ja valimidispersioon) või selle funktsioonina ja X jaotusena kasutame asümptootilist normaaljaotust, kahanevad liidetavad reaksarenduses $n^{-1/2}$ astmetena ja esituse (1) esimesed liikmed annavad meile $f_Y(x)$ jaoks teatud järku lähendi.

Tartus hakkas asümptootiline statistika arenema 1970-ndate teisel poolel just mitmemõõtmeliste valimifunktsioonide koonduvuse uurimisega (T. Kollo, A.-M. Parring, E. Saar, E.-M. Tiit). Seega sattusime aktuaalsele uurimistemaatikale kümmekond aastat enne selle väljatoomist ühe statistika arengu põhisuunana. See andis teatud edumaa ja äratas huvi ka mujal maailmas. Siiski õnnestus alles 40 aastat pärast artiklit Cornish & Fisher (1937) leida analoogiline tihedustevaheline seos mitmemõõtmelisel juhul (Traat, 1986). See sai võimalikuks tänu maatriksalgebra vahendite kasutamisele, millega 1970-ndatel tuletati Tartu rühma poolt erinevate $\bar{\mathbf{x}}$ ja $\mathbf{S}$ funktsioonide asümptootilisi normaaljaotusi. Kolm olulist mõistet, mille süstemaatilisel kasutamisel baseerub tänapäeva mitmemõõtmelise analüüsi esitus, on olemuselt lihtsad:

a) vec-operaator,

b) otsekorrutis,

c) maatrikstuletis.

Vec-*operaator* teisendab $p \times q$-maatriksi $\mathbf{A}$ veergude $\mathbf{a}_i$ üksteise alla paigutamise teel pq-vektoriks $\mathrm{vec}\,\mathbf{A}$.

Maatriksite $\mathbf{A} : p \times q$ ja $\mathbf{B} : r \times s$ *otsekorrutis* $\mathbf{A} \otimes \mathbf{B}$ on plokkmaatriks, mis koosneb $r \times s$ plokkidest

$$\mathbf{A} \otimes \mathbf{B} = [a_{ij}\mathbf{B}], \qquad i = 1, \ldots, p; \quad j = 1, \ldots, q.$$

Maatrikstuletis $r \times s$-maatriksist $\mathbf{Y}$ $p \times q$-maatriksi $\mathbf{X}$ järgi on osatuletistest $\frac{\partial y_{ij}}{\partial x_{kl}}$ koosnev maatriks, kus osatuletised on järjestatud vec-operaatori ja otsekorrutise abil:

$$\frac{d\mathbf{Y}}{d\mathbf{X}} = \frac{1}{\partial \mathrm{vec}\,'\mathbf{X}} \otimes \mathrm{vec}\,\mathbf{Y}.$$

Maatrikstuletis kujutab endast Frechet' tuletise maatriksesitust.

Need kolm mõistet on omavahel orgaaniliselt seotud ja maatrikstehnika abil õnnestuski võrdus (1) üle kanda mitmemõõtmelisele juhule. Saadud võrduses asendusid tavalised tuletised maatrikstuletistega, tavaline korrutamine otsekorrutisega ja vec-operaator teisendas maatriksid vektoriteks

$$\begin{aligned} & f_{\mathbf{y}}(\mathbf{x}) = f_{\mathbf{x}}(\mathbf{x}) + -(E\mathbf{y} - E\mathbf{x})'\mathrm{vec}\, f_{\mathbf{x}}^{(1)}(\mathbf{x}) \\ + \;& \frac{1}{2}\mathrm{vec}\,'\{D(\mathbf{y}) - D(\mathbf{x}) + (E\mathbf{y} - E(\mathbf{x}))(E\mathbf{y} - E(\mathbf{x}))'\}\mathrm{vec}\, f_{\mathbf{x}}^{(2)}(\mathbf{x}) \\ - \;& \frac{1}{6}\{\mathrm{vec}\,'[\overline{m}_3(\mathbf{y}) - \overline{m}_3(\mathbf{x}) + 3\mathrm{vec}\,'(D(\mathbf{y}) - D(\mathbf{x})) \otimes (E(\mathbf{y}) - E(\mathbf{x}))] \\ + \;& (E(\mathbf{y}) - E(\mathbf{x}))'^{\otimes 3}\}\mathrm{vec}\, f_{\mathbf{x}}^{(2)}(\mathbf{x}) \cdots . \end{aligned}$$

Reaksarenduses esinevad tsentraalsed momendid $\overline{m}_3(\mathbf{x})$ on defineeritud otsekorrutise abil:

$$\overline{m}_3(\mathbf{x}) = E[(\mathbf{x} - \boldsymbol{\mu}) \otimes (\mathbf{x} - \boldsymbol{\mu})' \otimes (\mathbf{x} - \boldsymbol{\mu})]$$

ja

$$D\mathbf{x} = E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})'].$$

Kahjuks ei lahenda paljusid probleeme ka võimalus lähendada huvipakkuvat keerulist p-mõõtmelist jaotust lihtsama p-mõõtmelise jaotusega. Statistiliste otsustuste puhul on vaja hinnata selliste suuruste nagu valimi dispersioonimaatriksi $\mathbf{S}$ ja valimi korrelatsioonimaatriksi $\mathbf{R}$ jälje, determinandi, omaväärtuste ja omavektorite jaotust. Need on reeglina juhuslikud suurused või väiksemamõõtmelised juhuslikud vektorid, kui seda on tunnusvektor $\mathbf{x}$ või maatriks $\mathbf{S}$. Oleks hea, kui saaksime nende juhuslike suuruste või vektorite

jaotuse kirjeldamisel ära kasutada kogu valimis oleva informatsiooni. Artiklis Kollo & von Rosen (1998) on leitud üldine seos, mille abil saab väiksemamõõtmelist tundmatut tõenäosustihedust esitada reaksarendusena suuremadimensionaalse kaudu. Matemaatiliselt avaldub see seos ülaltoodud $f_{\mathbf{y}}(\mathbf{x})$ reaksarenduse analoogina. Saadud esituses lisandus ka võimalus lähendada tihedust $f_{\mathbf{Y}}(\mathbf{y})$ tiheduse $f_{\mathbf{X}}(\mathbf{x})$ kaudu nii, et tiheduste argumendid on erinevad.

2. Lähendamine reaksarendusena

Esimene teemadering käsitleb tundmatu jaotuse lähendamist reaksarendusena teise lihtsama jaotuse kaudu. Ülalkirjeldatud seosest kahe eridimensionaalse tõenäosustiheduse vahel kasvas välja monograafia Kollo & von Rosen (2005), mis on vaadeldavas tööde tsüklis kesksel kohal. Raamatu esimeses peatükis on esitatud põhjalik ülevaade kasutatavast maatriksaparatuurist ja võreteooriast, kus autoritepoolsed tulemused on seotud eeskätt maatrikstuletise ja kujundmaatriksitega. Viimaste puhul on tegemist maatriksitega, kust osa elemente on välja jäetud. Saadud mõiste võimaldab eraldada maatriksitest korduvad ja konstantsed elemendid, mis on oluline maatrikstuletise rakendamisel sümmeetrilistele ja korrelatsioonimaatriksi tüüpi maatriksitele. Viimased sisaldavad ka konstante lisaks sümmeetriale. Ka raamatu teine peatükk on lähenduste seisukohalt ettevalmistava iseloomuga. Selles on esitatud olulisemad mitmemõõtmelised ja maatriksjaotused, mida kasutatakse töö kolmandas ja neljandas peatükis. Pearõhk on siin maatriks-normaaljaotuse ja Wisharti jaotuse omaduste kirjeldamisel, esitatud on ka ülevaade elliptilistest jaotustest, mis üldistavad normaaljaotust. Jaotuste kohta on saadud mitmeid uusi tulemusi, leitud on kõigi uuritud jaotuste esimeste momentide avaldised ja tuletatud on edaspidi arendustes kasutatavad Hermite maatrikspolünoomid.

Jaotuste lähendamisega tegeleb kolmas peatükk. Siin on tuletatud asümptootilised normaaljaotused põhiliste mitmemõõtmeliste valimifunktsioonide, sealhulgas hinnangute $\bar{\mathbf{x}}$ ja $\mathbf{S}$ funktsioonide jaoks, esitatud on üldine seos kahe mitmemõõtmelise tihe-

dusfunktsiooni vahel ja rakendatud seda juhul, kui lähendavaks jaotuseks on nii normaaljaotus kui ka Wisharti jaotus. Samuti on leitud lähendid uuritavale tihedusele kahe mitmemõõtmelise normaaljaotuse segu kaudu. Valimi dispersioonimaatriksi kõrval on valimi korrelatsiooni-maatriks $\mathbf{R}$ teine väga oluline vaatluste funktsioon, millel baseeruvad paljud mitmemõõtmelise analüüsi meetodid. Korrelatsioonimaatriksi jaotusele on leitud lähendid normaaljaotuse ja Wisharti jaotuse kaudu. Need tulemused on avaldatud artiklis Kollo & Ruul (2003).

Üks statistikameetodite eesmärke on tunnustevaheliste seoste kindlakstegemine ja andmetes leiduva informatsiooni kokkusurumine mitmesuguste mudelite konstrueerimise teel. Kõige levinumaks statistikamudeliks on lineaarne mudel, mis püüab kirjeldada andmeid lineaarse funktsiooni abil. See on klassikaline mudel juhul, kui vaatlused on sõltumatud. Statistilisi mudeleid on vaja konstrueerida aga ka olukorras, kus on tegemist sõltuvate vaatlustega. Üheks tüüpolukorraks on juhtum, kus sama objekti on ajas korduvalt mõõdetud. Sel juhul on tegemist nn. üldiste lineaarsete mudelitega, mis erijuhul on tuntud kui kasvukõvera mudelid. Kasvukõvera mudeli korral esitatakse vaatluste $p \times n$-maatriks järgmise võrdusega:

$$\mathbf{X} = \mathbf{ABC} + \mathbf{\Sigma}^{1/2}\mathbf{E},$$

kus $\mathbf{A} : p \times q$ ja $\mathbf{C} : k \times n$ on teadaolevad konstantsed maatriksid ning $\mathbf{B} : q \times k$ ja $\mathbf{\Sigma} : p \times p$ tundmatud parameetermaatriksid, mida tuleb hinnata. Maatriks $\mathbf{E} : p \times n$ on normaaljaotusega juhuslike vigade maatriks. Lineaarse mudeli korral on üks maatriksitest, $\mathbf{A}$ või $\mathbf{C}$, võrdne ühikmaatriksiga. Kui lineaarsete mudelite omadused on hästi teada, siis teise konstantse nn. disainimaatriksi lisamine mudelisse komplitseerib olukorda märgatavalt. Kirjandusest on teada maatriksite $\mathbf{B}$ ja $\mathbf{\Sigma}$ suurima tõepära hinnangud (Khatri, 1966), kuid nende hinnangute tõenäosuslik käitumine vajas uurimist. Raamatu Kollo & von Rosen (2005) neljas peatükk on pühendatud üldiste lineaarsete mudelite kirjeldamisele. Siin on põhjalikult esitatud nii klassikaline kasvukõvera mudel kui ka teised üldised lineaarsed mudelid. Ühtlasi on selle peatüki tulemuste näol tegemist ka

eelmistes osades esitatud teooria mittetriviaalsete rakendustega. Leitud on parameetermaatriksi $\mathbf{A}$ hinnangu tõenäosustiheduse lähend normaaljaotuse kaudu, mis osutus üllatavalt täpseks (järku n^{-2}). Lähem analüüs näitas (Kollo & Roos (2005), Kollo, Roos & von Rosen (2007)) et saadud lähendi näol on tegemist nn. elliptiliste jaotuste klassi kuuluva kahe jaotuse: normaaljaotuse ja Kotzi jaotuse seguga. Samuti on leitud teise parameetermaatriksi $\mathbf{\Sigma}$ hinnangu tõenäosustihedusele lähend reaksarendusena Wisharti jaotuse kaudu.

3. Mitmemõõtmelised ebasümmeetrilised jaotused

Teine tee ebasümmeetrilise mitmemõõtmeliste andmete kirjeldamiseks on ebasümmeetriliste jaotuste perede kasutamine. Kuni 1990-ndate keskpaigani selline võimalus praktiliselt puudus – peale mitmemõõtmelise normaaljaotuse ja selle üldistuse elliptiliste jaotuste näol teisi võimalusi juhusliku vektori jaotuse kirjeldamiseks polnudki, kui oli vaja jaotuses arvesse võtta ka tunnustevahelisi sõltuvusi. Viimase kümnekonna aasta jooksul on selles valdkonnas toimunud tõsine murrang. Kasutusele on võetud mitmed uued jaotuste pered, mis võimaldavad modelleerida ka ebasümmeetrilisi andmeid. Neist esimene oli *asümmeetriline normaaljaotus*, mis saadi mitmemõõtmelisest normaaljaotusest deformeerimise teel (Azzalini & Dalla Valle, 1996). Tema tõenäosustihedus on kujul

$$f(\mathbf{x}) = 2f_{N_p(\mathbf{0},\mathbf{\Sigma})}(\mathbf{x})\Phi(\boldsymbol{\alpha}'\mathbf{x}),$$

kus $f_{N_p(\mathbf{0},\mathbf{\Sigma})}(\mathbf{x})$ on p-mõõtmelise normaaljaotuse tihedus ja $\Phi(\cdot)$ standardse normaaljaotuse $N(0,1)$ jaotusfunktsioon. Jaotusel on kaks parameetrit: hajuvust kirjeldav maatriks $\mathbf{\Sigma}$ ja kujuparameeter vektor $\boldsymbol{\alpha}$. Asümmeetrilise normaaljaotusega juhusliku vektori tähistame $\mathbf{x} \sim SN_p(\mathbf{\Sigma}, \boldsymbol{\alpha})$. Tööde tsükli kolmes artiklis on arendatud teooriat asümmeetrilise normaaljaotuse kohta. Selle jaotuse kasutamise muudavad komplitseerituks raskused parameetrite hindamisel. Momentide meetod annab nihkega hinnangud ja suurima tõepära meetod võib viia valedele tulemustele. Artiklis

Dunajeva, Kollo & Traat (2003) on leitud parandusliige kujuparameetri momentide meetodi hinnangule ja esitatud uus, senisest lihtsam simuleerimiseeskiri asümmeetrilisest normaaljaotusest väärtuste genereerimiseks. Artiklis Gupta & Kollo (2003) on leitud asümmeetrilise normaaljaotuse esimesed momendid ja konstrueeritud lähendusvalem tundmatu tõenäosustiheduse $f_{\mathbf{y}}(\mathbf{y})$ lähendamiseks reaksarendusena asümmeetrilise normaaljaotuse baasil, töös Kollo & Selart (2004) aga rakendatud eelmise artikli tulemusi korrelatsioonikordaja ja korrelatsioonimaatriksi omaväärtuste jaotuste lähendamiseks. Paljude õnnestunud asümmeetrilise normaaljaotuse rakenduste kõrvale tekkisid varsti ka ülesanded, kus see jaotus ei andnud kooskõlalist mudelit ebasümmeetriliste andmete korral. Peamine põhjus oli siin keskväärtusest kaugel asetsevate väärtuste liiga väike osakaal võrreldes andmetega.

Sellest aspektist on tunduvalt paremate omadustega teine uus jaotuste klass – mitmemõõtmelised asümmeetrilised Laplace jaotused. Need jaotused tõusid statistikute tähelepanu orbiiti Tomasz J. Kozubowski ja kaasautorite töödega 1990-ndate teisel poolel (vt. näiteks Kozubowski, 1997). Raamat Kotz, Kozubowski & Podgorski (2001) andis ülevaate tulemustest 2001 aasta seisuga. Selle jaotuse väärtus rakenduste jaoks peitub eelkõige võimaluses võtta ebasümmeetrilise jaotuse korral arvesse keskväärtusest kaugel asetsevaid väärtusi suurema tõenäosusega kui normaaljaotuse korral. See teeb jaotuse eriti väärtuslikuks finantsandmete analüüsimisel. Nagu asümmeetriline normaaljaotus, nii ka mitmemõõtmeline asümmeetriline Laplace jaotus moodustab kaheparameetrilise jaotuste pere. Kahjuks ei saa selle jaotuse tõenäosustihedust esitada lihtsa avaldisena, küll aga saab jaotust hõlpsasti kirjeldada karakteristliku funktsiooni abil.

Juhuslik p-vektor $\mathbf{x}$ on *asümmeetrilise Laplace jaotusega* parameetritega $\mathbf{\Sigma}$ ja $\boldsymbol{\theta}$, kui tema karakteristlik funktsioon on kujul

$$\varphi_{\mathbf{x}}(\mathbf{t}) = \frac{1}{1 + i\mathbf{t}'\boldsymbol{\theta} + \frac{1}{2}\mathbf{t}'\mathbf{\Sigma}\mathbf{t}}.$$

Ka siin on $p \times p$-maatriks $\mathbf{\Sigma}$ hajuvusparameeter, p-vektor $\boldsymbol{\theta}$ on kujuparameeter. Mida suuremate väärtustega on $\boldsymbol{\theta}$ koordinaadid,

seda ebasümmeetrilisem on jaotus. Samas on parameetrite hindamine komplitseeritud. Artiklis Kollo & Srivastava (2004) kasutusele võetud uus parametrisatsioon Laplace jaotuse jaoks võimaldas hindamisülesande lahendada. Samuti leiti selles artiklis Laplace jaotuse esimeste momentide avaldised ja rakendati saadud tulemusi hüpoteeside kontrollimiseks jaotuse parameetrite kohta.

4. Statistilised koopulamudelid

Kolmas võimalus tundmatu mitmemõõtmelise jaotuse kirjeldamiseks on koopulate teooria abil. See teooria sai alguse artiklist Sklar (1959) ja arenes mõõduteooria osana aastakümneid ilma rakendusteta kuni 1990-ndate lõpul avastati tema väärtus finants- ja kindlustusandmete analüüsimiseks. Ülevaate koopulate teooriast leiab asjahuviline näiteks raamatust Nelsen (1999). Kui senivaadeldud jaotuste perede korral juhusliku vektori koordinaadid on kõik sama tüüpi (normaaljaotusega, Laplace jaotusega jne.), siis koopulate teooria annab võimaluse mitmemõõtmelise jaotuse konstrueerimiseks ka juhul, kui koordinaatide jaotused on eri tüüpi. Seejuures on võimalik arvesse võtta ka tunnustevaheline sõltuvus. Mis on koopula? Üks võimalus on seda defineerida ühtlase jaotusega juhuslike suuruste ühisjaotuse jaotusfunktsioonina. Kahemõõtmelisel juhul saab selgitada koopula seost suvalise kahe juhusliku suurusega järgmiselt. Olgu U ja V standardse ühtlase jaotusega juhuslikud suurused: $U, V \sim U(0,1)$ ning X ja Y rangelt monotoonselt kasvavate pidevate jaotusfunktsioonidega $F_X(x)$ ja $F_Y(y)$. Kasutades tõsiasja, et $F_X(X)$ ja $F_Y(Y)$ on ühtlase jaotusega $U(0,1)$, saame X ja Y ühisjaotusfunktsiooni $F_{X,Y}(x,y)$ esitada koopula $C(u,v)$ kui U ja V jaotusfunktsiooni kaudu järgmiselt:

$$\begin{aligned} C(u,v) &= P(U \le u, V \le v) = P(F_X(X) \le u, F_Y(Y) \le v) \\ &= P(X \le F_X^{-1}(u), Y \le F_Y^{-1}(v)) = F_{X,Y}(x,y), \end{aligned}$$

kus $x = F_X^{-1}(u)$ ja $y = F_Y^{-1}(v)$ See tähendab, et juhuslike suuruste X ja Y ühisjaotus on esitatav koopulana, kusjuures monotoonset sõltuvust kirjeldavad astakkorrelatsioonikordajad X

ja Y vahel ning U ja V vahel on võrdsed. Keerukaks teeb koopulate kasutamise andmeanalüüsis see, et erinevaid koopulate peresid on väga palju ja nende hulgast sobivaima leidmine ei ole lihtne. Kahjuks ei ole seni leitud üldist lähenemisviisi parima koopula väljavalimiseks ja mudelite konstrueerimine seisneb paljude erinevate perede sobitamises andmetele ja nende hulgast parima valimises. Vaadeldavas tööde tsüklis modelleeritakse artiklis Kollo & Pettere (2006) koopulate abil kindlustusfirma vajalikke reserve liikluskindlustuse korral toimunud, aga veel teatamata kahjude jaoks.

Kirjandus

[1] Azzalini, A.; Dalla Valle, A. (1996), *The multivariate skew-normal distribution.* Biometrika, **83**, 715–726.

[2] Cornish, E. A.; Fisher, R. A. (1937), *Moments and cumulants in the specification of of distributions.* International Statistical Review, **5**, 307–322.

[3] Dunajeva,Õ.; Kollo, T.; Traat, I. (2003), *Bias correction for the shape parameter of the skew normal distribution.* Tatra Mountains Mathematical Publications, PROBASTAT'02, Part II, **26**, 281–289.

[4] Gupta, A. K.; Kollo, T. (2003), *Density expansions based on the multivariate skew normal distribution.* Sankhya, **65**, 821–835.

[5] Khatri, C. G. (1966), *A note on manova model applied to problems in growth curve.* Annals of Institute of Statistical Mathematics, **18**, 75–86.

[6] Kollo, T.; Pettere, G. (2006), *Copula models for estimating outstanding claim provisions.* In: Festschrift for Tarmo Pukkila on His 60th Birthday. Eds. Liski, E. P., Isotalo, J., Niemelä, J., Puntanen,Š., Styan, G. P. H., University of Tampere, Tampere, 115–125.

[7] Kollo, T.; Roos, A. (2005), *On Kotz-type elliptical distributions.* In: Contemporary Multivariate Analysis and Design of Experiments, Eds. Fan, J.; Li, G., World Scientific, New Jersey, 159–170.

[8] Kollo, T.; Roos, A.; von Rosen, D. (2007), *Approximation of the distribution of the location parameter in the Growth Curve model.* Scandinavian Journal of Statistics, **34**, 499–510.

[9] Kollo, T.; von Rosen, D. (1998), *A unified approach to to the approximation of multivariate densities.* Scandinavian Journal of Statistics, **25**, 93–109.

[10] Kollo, T.; von Rosen, D. (2005), *Advanced Multivariate Statistics with Matrices.* Springer, Dordrecht.

[11] Kollo, T.; Ruul, K. (2003), *An approximation to the distribution of the sample correlation matrix.* Journal of Multivariate Analysis **85**, 318–334.

[12] Kollo, T.; Selart, A. (2004). *Density expansions for correlations and eigenvalues of the covariance matrix.* Acta et Commentationes Universitatis Tartuensis de Mathematica, **8**, 155–168.

[13] Kollo, T.; Srivastava, M. S. (2004), *Estimation and testing of parameters in multivariate Laplace distribution.* Communications in Statistics. Theory and Methods, **33**, 2363–2387.

[14] Kozubowski, T. J. (1997), *Characterization of multivariate geometric stable distributions.* Statistics & Decisions, **15**, 397–416.

[15] Kotz,Š.; Kozubowski, T. J.; Podgorski, K. (2001), *The Laplace Distribution and Generalizations. A Revisit with Applications to Communications, Economics, Engeneering and Finance.* Birkhäuser, Boston.

[16] Nelsen, R. B. (1999), *An Introduction to Copulas.* Springer, New York.

[17] Sklar, A. (1959), *Fonctions de répartition à n dimensions et leurs marges.* Publications de l'Institut de Statistique de l'Universit de Paris, **8**, 229–231.

[18] Traat, I. (1986), *Matrix calculus for multivariate statistics.* Acta et Commentationes Universitatis Tartuensis, **733**, 64–84.

