Kordusmõõtmiste andmelünkade käsitlemine koopulate abil

Ene Käärik[1]
Tartu Ülikool

1. Kordusmõõtmistest ja andmelünkadest

Kordusmõõtmistega on tegemist juhul, kui sama objekt/subjekt on mõõdetud korduvalt ajas (ruumis). Kordusmõõtmistele on iseloomulik asjaolu, et samal objektil/subjektil teostatud mõõtmised on omavahel seotud ja seda seost ei tohi ignoreerida. Kordusmõõtmiste analüüsi teostatakse paljudes teadusvaldkondades – biostatistikas, biomeditsiinis, sotsioloogias jm. ning enamasti on probleemiks, et ühel või teisel põhjusel pole võimalik koguda täielikke andmeid.

Puuduvate andmete ehk andmelünkade probleemiga on tegeletud juba pikka aega, süstemaatilise teooria rajajateks võib pidada D. B. Rubinit ja R. J. A. Little'i, kes esitasid ka puudumiste tüpoloogia (Rubin, 1976; Little ja Rubin, 1987). Puuduvate väärtustega andmete analüüsimiseks on välja töötatud terve rida meetodeid, kuid samas pole olemas ühtegi, mida võiks pidada universaalseks ja parimaks.

Lünkliku andmestiku käsitlemisel on põhimõtteliselt kaks ülesannet:
(a) *hindamisülesanne*, kus eesmärgiks on saada lünkliku andmestiku põhjal mudeli parameetritele hinnangud, mis on võimalikult lähedased hinnangutele, mida olnuks võimalik saada siis, kui andmed oleksid olnud täielikud;
(b) *imputeerimisülesanne*, kus eesmärgiks on puuduva väärtuse võimalikult täpne prognoosimine.

Antud töös on vaatluse all teine ülesanne, st lünkade täitmine ehk imputeerimine, mis on eriti oluline praktilistes ülesannetes

[1] Autor kaitses doktoritöö matemaatilise statistika erialal 13. märtsil 2007.

väikeste valimite korral. Prognoosiülesande lahendamiseks võib kasutada näiteks kas ainult vaadeldava tunnuse jaotust või kasutada tunnuse tinglikku keskväärtust, kui teiste tunnuste väärtused on teada. Viimane oleks põhimõtteliselt parim lahendus, mida aga tegelikkuses ei rakendata, sest tihti pole võimalik hinnata tunnuste ühisjaotust ja seega ei saa leida ka tinglikku jaotust. Võimalikuks lahenduseks sel juhul oleks leida tee ühisjaotuse lähendamiseks, seejärel leida puuduva väärtuse tinglik jaotus ja arvutada lähendatud tingliku jaotuse põhjal tinglik keskväärtus (või ka mingi muu tingliku jaotuse karakteristik). Puuduva väärtuse tingliku jaotuse kasutamise eesmärk on maksimaalselt ära kasutada kogu olemasolev informatsioon andmetes:
(1) kasutada mõõtmiste ajalugu (moodustab tingimuse);
(2) kasutada vaatlustulemuste marginaaljaotusi – tingimatuid jaotusi, mida oluliselt täpsustatakse tingimuse abil;
(3) kasutada seostestruktuuri mõõtmiste vahel.

Probleem on selles, et tuntud mitmemõõtmelised jaotused ei pruugi sobida ühisjaotuse kirjeldamiseks ja seepärast on võetud kasutusele koopulad. Uudseks aspektiks antud töös ongi tingliku jaotuse leidmine koopulate abil.

2. Koopulate kasutamine

Koopula on funktsioon, mis ühendab marginaaljaotused ühisjaotuseks. Kasutades koopulat, saame eraldi hinnata marginaaljaotused ja seejärel arvestades seoste struktuuri määrata ühisjaotuse. Põhjalik teoreetiline ülevaade koopulatest on antud H. Joe ja R. B. Nelseni monograafiates (Joe, 1997; Nelsen, 1999).

Koopulad on algselt leidnud rakendust eeskätt kindlustus- ja finantsmatemaatikas, viimasel ajal ka biostatistikas (meteoroloogias), biomeditsiinis ja keskkonnastatistikas.

Kordusmõõtmiste analüüsis on koopulaid kasutanud vaid vähesed autorid. Näiteks, Lindsey ja Lindsey (2002) kirjeldavad Gaussi koopulat kordusmõõtmistega andmete korral, kuid nad ei käsitle lünkadega andmestikku. Lambert ja Vandenhende (2002), Van-

denhende ja Lambert (2002) on rakendanud koopulate lähenemist mudelite leidmisel kordusmõõtmistega lünklike andmestike korral, nad testisid erinevaid marginaaljaotusi (Cauchy, gamma, lognormaalne) ja kasutasid Gaussi koopulat ning Franki koopulat, kirjeldamaks seost uuritava tunnuse ja lünkade vahel.

Koopulate kasutamisel on terve rida eeliseid klassikaliste meetodite ees. Klassikalised mudelid baseeruvad mitmemõõtmelisel normaaljaotusel või mõnel teisel mitmemõõtmelisel jaotusel, mis seavad teatud nõudmised ka marginaaljaotuste kohta. Koopulamudel on paindlikum, ta lubab kombineerida erinevaid marginaaljaotusi ja rakendada nende sidumiseks erinevaid seostestruktuure. Saadud koopulamudeli sobivuse kontrollimiseks võib kasutada klassikalisi sobivuse teste (χ^2, AIC, BC ja nende modifikatsioone).

Koopulal baseeruv imputeerimise eeskiri sisaldab üldjuhul järgmisi etappe:

1. Hinnata marginaaljaotused.
2. Valida sobiv koopula ja hinnata valimi põhjal koopula parameetrid. Tulemuseks on mõõtmiste ühisjaotus, avaldada mõõtmiste ühistihedusfunktsioon.
3. Leida puuduvat väärtust sisaldava mõõtmise tinglik tihedusfunktsioon, kus tingimuse määravad olemasolevad mõõtmised.
4. Hinnata tinglik keskväärtus või mediaan ja kasutada seda puuduva väärtuse imputeerimiseks (asendamiseks).

3. Gaussi koopula

Võrreldes omavahel koopulaid, mis võimaldavad siduda suvalise arvu marginaaljaotusi mitmemõõtmeliseks jaotuseks, on kõige loomulikum alustada normaal- ehk Gaussi koopulaga (Clemen ja Reilly, 1999; Song, 2000; Lambert ja Vandenhende). Gaussi koopula puhul on võimalik hinnata ja arvestada kõiki k-mõõtmelise tunnusvektori $\frac{k(k-1)}{2}$ paarisseose kordajaid hindamaks seoste struktuuri ning lihtsamate seosestruktuuride korral on võimalik tingliku

keskväärtuse (st asendusväärtuse) ja standardhälbe jaoks tuletada lihtsad valemid.

Mitmemõõtmeline Gaussi koopula on defineeritud järgmiselt.

Definitsioon. *Olgu R positiivselt määratud maatriks, mille korral $diag(R) = (1, 1, \cdots, 1)^T$ ja olgu $\Phi_{(k)}$ k-mõõtmelise standardnormaaljaotuse jaotusfunktsioon korrelatsioonimaatriksiga R, siis mitmemõõtmeline Gaussi koopula avaldub kujul*

$$C(u_1, \cdots, u_k; R) = \Phi_{(k)}(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_k)),$$

kus $u_j \in (0, 1),\ j = 1, \ldots, k$ ja Φ^{-1} on ühemõõtmelise standardnormaaljaotuse jaotusfunktsiooni pöördfunktsioon.

Tähistame $X = (X_1, \ldots, X_k)$ korduvate mõõtmiste vektori, kus mõõtmised on tehtud ajahetkedel $1, \ldots, k$. Mõõtmise X_j jaotus- ja tihedusfunktsioonid olgu vastavalt F_j ja f_j. Oletame, et ajahetkeni $k - 1$ on mõõtmised täielikud ning alates ajahetkest k esineb vähemalt üks puuduv väärtus.

Mõõtmiste $X_1, \ldots, X_k$ ühistihedusfunktsioon avaldub marginaaltiheduste f_j ja koopula tiheduse c korrutisena

$$f_X(x_1, \ldots, x_k; R) = f_1(x_1) \cdot \ldots \cdot f_k(x_k)\ c[F_1(x_1), \ldots, F_k(x_k); R].$$

Kasutades normaalkoopula tiheduse avaldist jõuame ühistihedusfunktsioonini kujul

$$f_X(x_1, \ldots, x_k; R) = f_1(x_1) \cdot \ldots \cdot f_k(x_k) \frac{\exp\{-\frac{1}{2}[\mathbf{q}^T(R^{-1} - I)\mathbf{q}]\}}{|R|^{\frac{1}{2}}},$$

kus $\mathbf{q} = (\Phi^{-1}[F_1(x_1)], \ldots, \Phi^{-1}[F_k(x_k)])^T$.

4. Üldine imputeerimiseeskiri ja selle rakendused

Meid huvitab puuduvat väärtust sisaldava tunnuse tinglik jaotus täielike mõõtmiste suhtes. Täielikud mõõtmised moodustavad mõõtmiste ajaloo $H_{(k-1)} = (X_1, \ldots, X_{k-1})$.

Olgu $R = \{r_{ij}\},\ (i, j = 1, \ldots, k),$ andmete korrelatsioonimaatriks.

Arvestades mõõtmiste ajalugu saame teostada korrelatsioonimaatriksi lahutuse

$$R = \begin{pmatrix} R_{(k-1)} & \mathbf{r}_{(k-1)} \\ \mathbf{r}_{(k-1)}^T & 1 \end{pmatrix}, \tag{1}$$

kus $R_{(k-1)}$ on ajaloo korrelatsioonimaatriks ja

$$\mathbf{r}_{(k-1)} = (r_{1,k}, \ldots, r_{k-1,k})^T$$

on korrelatsioonide vektor ajaloo ja k-nda mõõtmispunkti vahel.

Rakendame normaliseerivat teisendust

$$Y_j = \Phi^{-1}[F_j(X_j)], \quad j = 1, \ldots, k,$$

ja saame standardnormaaljaotusega tunnused $Y_1, \ldots, Y_k$.

Eelnevat arvesse võttes saame tingliku tihedusfunktsioonini järgmisel kujul:

$$f(y_k|H_{(k-1)}; R) = \frac{1}{\sqrt{2\pi(1 - \mathbf{r}_{(k-1)}^T R_{(k-1)}^{-1} \mathbf{r}_{(k-1)})}} \exp\{-\frac{(y_k - \mathbf{r}_{(k-1)}^T R_{(k-1)}^{-1} \mathbf{y}_{(k-1)})^2}{2(1 - \mathbf{r}_{(k-1)}^T R_{(k-1)}^{-1} \mathbf{r}_{(k-1)})}\},$$

kus $\mathbf{y}_{(k-1)} = (y_1, \ldots, y_{k-1})^T$.

Imputeerimiseks ehk puuduva väärtuse asendamiseks tuleks leida väärtus, mis maksimiseerib tingliku tihedusfunktsiooni, st leida *argmax* tinglikust tihedusfunktsioonist. Tehniliselt on see protseduur samaväärne tingliku keskväärtuse suurima tõepära hinnangu leidmisega.

Lihtsuse mõttes kasutame logaritmilist tihedusfunktsiooni

$$\ln f(y_k|H_{(k-1)}; R).$$

maksimiseerides seda y_k suhtes, saame

$$\frac{\partial \ln f(y_k|H_{(k-1)}; R)}{\partial y_k} = \frac{-y_k + \mathbf{r}_{(k-1)}^T R_{(k-1)}^{-1} \mathbf{y}_{(k-1)}}{1 - \mathbf{r}_{(k-1)}^T R_{(k-1)}^{-1} \mathbf{r}_{(k-1)}} = 0.$$

Seega saame *üldise tinglikul keskväärtusel baseeruva imputeerimiseeskirja* järgmisel kujul (Käärik, 2005):

$$\hat{y}_k = \mathbf{r}_{(k-1)}^T \cdot R_{(k-1)}^{-1} \mathbf{y}_{(k-1)}, \tag{2}$$

kus $\mathbf{r}_{(k-1)}$ on korrelatsioonide vektor ajaloo ja k-nda mõõtmise vahel, $R_{(k-1)}^{-1}$ on ajaloo korrelatsioonimaatriksi pöördmaatriks ja $\mathbf{y}_{(k-1)} = (y_1, \ldots, y_{k-1})^T$ on vaatluste vektor, kus k-ndal ajahetkel mõõtmist ei toimunud.

On lihtne kontrollida, et

$$\frac{\partial^2 \ln f(y_k | H_{(k-1)}; R)}{\partial^2 y_k} = \frac{-1}{1 - \mathbf{r}_{(k-1)}^T R_{(k-1)}^{-1} \mathbf{r}_{(k-1)}} < 0,$$

sest nimetajas on mitmese korrelatsioonikordaja ruut (Rao, 1965, p 223), mis ei saa olla ühest suurem. Seega hinnang, mille saime seosega (2) maksimiseerib tõepoolest tingliku tihedusfunktsiooni.

Tulemuse saame sõnastada järgmiselt.

Lause 1. *Olgu $Y_1, \ldots, Y_k$, korduvad mõõtmised standardnormaaljaotusega. Olgu andmete korrelatsioonimaatriks lahutatud kujul* (1) *ja olgu ajahetkel k vähemalt üks puuduv mõõtmine, nii et ajalugu ajahetkeni $k-1$ on täielik. Sel juhul puuduva väärtuse imputeerimiseks ajahetkel k on kasutatav eeskiri* (2).

Esitatud tulemust üldistab järgmine järeldus.

Järeldus 1. *Olgu $X_1, \ldots, X_k$, korduvad mõõtmised suvaliste marginaalidega $F_1, \ldots, F_k$. Sel juhul puuduva väärtuse imputeerimiseks ajahetkel k on kasutatav eeskiri* (2).

Suvaliste marginaalide korral kasutatakse järgmist kolmesammulist protseduuri:

1. Rakendada normaliseerivat teisendust $Y_j = \Phi^{-1}(F_j(X_j))$, $j = 1, \ldots, k$.
2. Imputeerida puuduv väärtus kasutades eeskirja (2).
3. Rakendada pöördteisendust $X_k = F_k^{-1}[\Phi(Y_k)]$ saamaks väärtust imputeerimiseks esialgsel kujul.

Üldise imputeerimiseeskirja (2) rakendamisel on põhiliseks probleemiks andmete korrelatsioonimaatriksi struktuuri hindamine ja korrelatsioonimaatriksi pöördmaatriksi leidmine. Näitena on vaatluse all kolm lihtsamat korrelatsioonistruktuuri: (a) konstantne korrelatsioonistruktuur, kus sõltuvus kõikide mõõtmispunktide vahel on ühesugune; (b) esimest järku autoregressiivne struktuur, kus sõltuvus väheneb kui mõõtmiste vaheline aeg kasvab; (c) tõkestatud Toeplitzi struktuur, kus ainult kaks järjestikust mõõtmist on sõltuvad.

Vaadeldud korrelatsioonistruktuuride korral on leitud vastavad korrelatsioonimaatriksite pöördmaatriksid ja tuletatatud nende korral imputeerimise valemid (vt Käärik, 2006a; Käärik 2006b).

Teostatud on simulatsioonieksperimendid võrdlemaks erinevaid imputeerimismeetodeid ja selgitamaks välja koopula abil imputeerimise plusse. Simulatsioonid näitavad, et esitatud metoodika on sobiv rakendamiseks andmelünkade täitmiseks väikese valimi korral.

Kirjandus

[1] Clemen, R. T.; Reilly, T. (1999), *Correlations and copulas for decision and risk analysis.* Fuqua School of Business, Duke University.

[2] Joe, H. (1997), *Multivariate Models and Dependence Concepts.* London: Chapman and Hall.

[3] Käärik, E. (2005), *Handling dropouts by copulas.* In: WSEAS Transactions on Biology and Biomedicine, Ed. N. Mastorakis. Vol **1**, (2), 93–97.

[4] Käärik, E. (2006a), *Imputation algorithm using copulas.* Advances in Methodology and Statistics, Ed. A. Ferligoj. Vol **3** (1), 109–120.

[5] Käärik, E. (2006b), *Imputation by conditional distribution using Gaussian copula.* In: Proceedings in Computational Statistics. Compstat'06, Ed. A. Rizzi and M. Vichi. Physica-Verlag, Springer, 1447–1454.

[6] Lambert, P.; Vandenhende, F. (2002), *A copula-based model for multivariate non-normal longitudinal data: analysis of a dose titration safety study on a new antidepressant.* Statistics in Medicine, **21**, 3197–3217.

[7] Lindsey, J. K.; Lindsey, P. J. (2002), *Multivariate distributions with correlation matrices for nonlinear repeated measurements* (available from www.luc.ac.be/~jlindsey : Nov, 2005).

[8] Little, J. A.; Rubin, D. B. (1987),. *Statistical Analysis with Missing Data*. New York: Wiley.

[9] Nelsen, R. B. (1999), *An introduction to copulas*. Lecture Notes in Statistics, **139**, New York: Springer Verlag.

[10] Rao, C. R. (1965), *Linear statistical inference and its applications*. New York: Wiley.

[11] Rubin, D. B. (1976), Inference and Missing Data. Biometrika, **63**, 581–592.

[12] Song, P. X. K. (2000), *Multivariate dispersion models generated from Gaussian Copula.* Scandinavian Journal of Statistics, **27**, 305–320.

[13] Vandenhende, F.; Lambert, P. (2002), *On the joint analysis of longitudinal responses and early discontinuation in randomized trials.* Journal of Biopharmaceutical Statistics, **12** (4), 425–440.

