ALKOHOLIPOLIITTISEN
TUTKIMUSLAITOKSEN TUTKIMUSSELOSTE

N:o 14

Tammikuu 1965

TILASTOLLISISTA MONIMUUTTUJAMENETELMISTÄ

Seppo Mustonen

OY ALKOHOLILIIKE AB
Alkoholipoliittinen tutkimuslaitos
Pitkänsillanranta 3 B — Helsinki

4.2. 1965

Seppo Mustonen:

TILASTOLLISISTA MONIMUUTTUJAMENETELMISTÄ

Tilastollisen monimuuttuja-analyysin (Statistical Multivariate Analysis) teoria on kehittynyt rinnan muiden tilastollisten tutkimusmenetelmien kanssa. Useimmat nykyisin käytössä olevista monimuuttujamenetelmistä ovat saaneet alkunsa 1930-luvulla lähinnä R. A. Fisherin, Hotellingin ja Thurstonen ansiosta. Kuitenkin vasta viime vuosina ne ovat tulleet yleisesti tunnetuiksi, sillä etupäässä laskuteknillisistä hankaluuksista johtuen ei ole aikaisemmin ollut mahdollista soveltaa näitä menetelmiä ainakaan laajoissa tutkimuksissa.

1. Multinormaalijakautuma

Puheena olevan teorian eräänä tärkeänä kulmakivenä on multinormaalijakautuma. Keskustelut monimuuttujamenetelmien kelpoisuudesta erilaisissa sovellutuksissa johtavatkin tavallisesti pohdintaan multinormaalisuusoletta-muksen oikeutuksesta. Periaatteessa olisi tietenkin mahdollista rakentaa monimuuttujamalleja muidenkin moniulotteisten todennäköisyysjakautumien pohjalta tai yleisemmässä muodossa, mutta tällaiset yritykset näyttävät olevan ylen harvinaisia. Syitä vallitsevaan tilanteeseen on helppo luetella.

1.1. Samalla tavalla kuin normaalijakautumalla on keskeinen asema jatkuvien yksiulotteisten todennäköisyysjakautumien joukossa todennäköisyyslaskennan keskeisen raja-arvolauseen ansiosta (ks. esim. Elfving //6// s. 75, Cramér //5// s. 231), myös multinormaalijakautumaa voidaan pitää erikoisasemassa jatkuvien moniulotteisten jakautumien joukossa (Cramér //5// s. 316). Käytännön tutkimuksissa on usein "hyvä syy uskoa", että jakautuma on ainakin likipitäen multinormaalinen.

1.2. Muulla kuin "hyvällä uskolla" ei multinormaalisuusolettamusta voi juuri käyttää, sillä on erittäin vaikea osoittaa esim., että annettu otos on kotoisin multinormaalisesta perusjoukosta. Yhteisjakautuman multinormaalisuutta ei nimittäin takaa yksistään kunkin erillisen muuttujan normaalisuus. Onpa jopa niin, ettei edes kahden normaalisesti jakautuneen muuttujan keskinäinen korreloimattomuus ole riittävä ehto multinormaalisuudelle (Anderson //3// s. 37).

1.3. Teoreettisissa tarkasteluissa hallitaan tilanteet paremmin, jos multinormaalisuusolettamus on tehty. Tähän on syynä paitsi multinormaali-

jakautuman siisti analyyttinen rakenne myös sen ainutlaatuinen yksinkertaisuus mm. siinä mielessä, että multinormaalisesti jakautuneiden muuttujien korreloimattomuudesta seuraa aina niiden keskinäinen tilastollinen riippumattomuus. Tämähän pitää yleisessä tapauksessa paikkansa vain päinvastaisessa suunnassa - riippumattomuus implikoi korreloimattomuuden.

1.4. Multinormaalisuustapauksessa tarkastelut ovat lisäksi sikäli helpompia, että jakautuman määrittämiseksi riittää, jos tunnetaan parametreina muuttujien keskiarvot ja kovarianssimatriisi (yhtäpitävästi keskiarvot, keskihajonnat ja korrelaatiot). Tällöin myös jakautuman parametreja estimoitaessa ei tarvitse laskea toista kertalukua korkeampia momentteja.

1.5. Päinvastoin kuin esim. normaalisuuteen perustuvat tilastolliset testit useat monimuuttujamallit eivät liene kovinkaan herkkiä sille, että käytännössä esiintyy suuriakin poikkeavuuksia multinormaalisuudesta. Tämä johtunee lähinnä siitä, että nämä mallit muuttujayhdistelyjen kautta yleensä lähentävät tutkittavaa aineistoa multinormaalijakautuman suuntaan. Niinpä esim. erotteluanalyysia on voitu menestyksellä soveltaa tilanteessa, jossa useimmat muuttujat olivat melkeinpä dikotomisia (Mustonen //9//).

2. Monimuuttujamenetelmät

Oletetaan, että tarkasteltavana on p satunnaismuuttujaa $x_1, x_2, \ldots, x_p$, jotka yhteisesti noudattavat multinormaalijakautumaa parametrein μ, Σ. (μ on $p \times 1$ vektori, joka muodostuu muuttujien keskiarvoista ja Σ $p \times p$ matriisi, jonka alkioina ovat muuttujien varianssit ja kovarianssit.)

Tavallisesti jakautuman parametrit joko kokonaan tai osittain ovat tuntemattomia ja ne joudutaan estimoimaan otoksen avulla, mikä merkitsee empiiristen keskiarvojen, keskihajontojen ja korrelaatioiden laskemista.

Monimuuttujamenetelmien tunnusomaisimmaksi piirteeksi voidaan lukea, että tarkastelu kohdistuu paitsi kunkin muuttujan omaan vaihteluun myös muuttujien keskinäiseen vaihteluun, joka multinormaalitilanteessa kuvastuu tyhjentyvästi muuttujien välisissä korrelaatioissa.

Monimuuttujamallien teoriaa on hankala esittää täysin yhtenäisessä muodossa niin, että kaikkien eri menetelmien ominaisuudet voitaisiin ottaa esille teorian erikoistapauksina. Ainakin yksi yritys, lähinnä informaatioteoriaan nojautuen, on tehty (Kullback //8//). Tässä yhteydessä tyydyn kuitenkin käyttämään varsin alkeellisia keinoja menetelmien kuvaamiseksi, jolloin yleinen selostus vain hyvin vajanaisesti valaisee kunkin menetelmän taustaa ja menetelmien välisiä yhteyksiä.

Tilastolliset monimuuttujamenetelmät jaetaan yleensä kahteen pääryhmään:

I Yhden muuttujan menetelmien yleistykset

II Varsinaiset monimuuttujamallit

I ryhmän menetelmillä on kullakin oma vastineensa yhden muuttujan teoriassa. Erikoisesti normaalijakautumaan perustuvat tilastolliset testit ovat helposti yleistettävissä useita muuttujia koskeviksi. Niinpä esim. t-testi, jolla perusmuodossaan testataan normaalisen muuttujan keskiarvoa μ koskeva hypoteesi $\mu = \mu_0$, yleistyy multinormaalisuustilanteessa Hotellingin T^2-testiksi, jossa testattava hypoteesi on muodollisesti sama, mutta koskee nyt keskiarvovektoria μ (testataan siis hypoteesi: kaikkien muuttujien keskiarvot ovat annettuja lukuja) (Anderson //3//, Chapter 5).

Myös useilla tunnetuilla yhden satunnaismuuttujan tilastomatemaattisilla malleilla, kuten varianssi- ja regressioanalyysilla, on omat vektorimuuttujia koskevat yleistyksensä (Anderson //3//, Chapter 8).

II ryhmän menetelmät, jotka ovat vailla vastinetta yhden muuttujan teoriassa, voidaan tulkita helpoimmin muuttujakuvauksiksi, joissa kuvaus valitaan siten, että mielenkiinnon kohteena olevat ominaisuudet tutkittavassa aineistossa saadaan parhaiten esille. Monimuuttujamalli on yleensä viivallinen, jolloin malli on muotoa

$$y_1 = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1p}x_p$$

$$y_2 = a_{21}x_1 + a_{22}x_2 + \ldots + a_{2p}x_p$$

$$\vdots$$

$$y_m = a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mp}x_p$$

eli matriisimuodossa $y = Ax$. Tässä muuttujat $x_1, x_2, \ldots, x_p$ (muuttujavektori x) ovat havaittavia muuttujia, muuttujat $y_1, y_2, \ldots, y_m$ (muuttujavektori y) uusia ns. "piilomuuttujia" ja kertoimet a_{ij} (kerroinmatriisi A) mallin parametrit, "lataukset", jotka täsmentävät kuvauksen havaittavilta muuttujilta piilomuuttujille.

Joskus (esim. faktorianalyysissa) on parempi muodostaa malli toisinpäin, siis kuvauksena piilomuuttujilta havaittaville muuttujille, jolloin tulee mahdolliseksi ikäänkuin lajitella piilomuuttujia jo mallia rakennettaessa (faktorianalyysissa jako: common factors - unique factors).

Mallia muodostettaessa on luonnollisesti tärkeintä kuvauksen A valinta. Yleisesti voidaan sanoa, että valinta perustuu kahdelle päävaatimukselle

1° $m \leq p$,

2° y "parempi kuin" x .

Ensimmäinen vaatimus on järkevä jo kuvauksen ekonomian kannalta. Lisäksi on ilmeistä, että m:n ollessa huomattavasti pienempi kuin p (esim. p = 50, m = 3) multinormaalisuusolettamus piilomuuttujien kohdalla on oikeutetumpi kuin havaittujen muuttujien kohdalla.

Toisen vaatimuksen ilmaisemaa paremmuutta voi edustaa esim.

a) jokin tilastomatemaattinen ominaisuus
   esim. pääkomponenttianalyysissa kuvaus A valitaan siten, että uudet muuttujat y (pääkomponentit) kukin vuorollaan sitovat itseensä mahdollisimman suuren osan havaittujen muuttujien x kokonaisvaihtelusta ja ovat lisäksi korreloimattomia.

b) tulkittavuus
   esim. faktorianalyysin rotaatiovaiheessa pyrkiminen "simple structure" -periaatteella ilmeikkäisiin faktoreihin (Thurstone //11//).

On kuitenkin huomattava, että silloinkin, kun kuvauksen valinnan ensiarvoisena perusteena ei ole paremmuus tilastomatemaattisessa mielessä, pyritään paremmuuden tavoittelu palauttamaan tähän (esim. analyyttiset rotaatiomenetelmät faktorianalyysissa), mikä vähentää tulkittavuus-periaatteeseen helposti liittyvää subjektiivisuuden vaaraa. Vaatimus 2^{o} johtaakin yleensä siihen, että kuvaus A joudutaan määräämään matemaattisen optimointitehtävän kautta, joka yleensä on jokin kovarianssimatriiseihin liittyvä ominaisarvotehtävä ja täten laskennallisesti suuritöinen.

3. Esimerkkejä

Pääkomponenttianalyysin periaate on selostettu edellä. Tehtävänä on siis siirtää muuttujien x kokonaisvaihtelu voimakkuusjärjestyksessä uusille muuttujille, pääkomponenteille y (Anderson //3//, Chapter 11).

Faktorianalyysi on itse asiassa usean monimuuttujamallin yhdistelmä. Sen faktorointivaihe tulee lähelle pääkomponenttianalyysia. Erona kuitenkin on, että faktoreille y ei siirretä muuttujien x kokonaisvaihtelua, vaan ainoastaan se osa tästä vaihtelusta (keskinäinen vaihtelu), joka on yhteistä kahdelle tai useammalle x-muuttujalle. Voidaan sanoa, että pääkomponenttianalyysi on menetelmänä varianssiorientoitu, kun taas faktorianalyysi on kovarianssiorientoitu.

Rotaatio on faktorointivaiheen pohjalle rakentuva monimuuttujakuvaus, jossa, kuten edellä todettiin, analyysin tulokset pyritään saattamaan tulkinnan kannalta edulliseen muotoon (Harman //7//, Thurstone //11//).

Faktorianalyysiin läheisesti liittyvistä monimuuttujamenetelmistä mainittakoon vielä faktorianalyysitulosten vertailumenetelmä, transformaatioanalyysi (Ahmavaara //1//, //2//).

Kanonisessa analyysissa tarkastellaan yhtaikaa kahta muuttujaryhmää ja päämääränä on löytää kummastakin ryhmästä sellaiset yhdistetyt muuttujat, että aina pareittain näiden yhdistettyjen muuttujien, kanonisten muuttujien, väliset riippuvuudet ovat korrelaation mielessä mahdollisimman suuret mutta muuten niiden väliset korrelaatiokertoimet ovat nollia. Kanonisessa analyysissa yritetään siis löytää "paras yhteys" kahden muuttujaryhmän välillä (Anderson //3//, Chapter 12).

Erotteluanalyysissa (diskriminaatioanalyysissa) on vuorostaan kaksi tai useampia perusjoukkoja (ryhmiä) vertailtavana ja pyritään löytämään sellaiset yhdistetyt muuttujat, diskriminaattorit, jotka parhaiten kuvaavat ryhmien välisiä eroja. Erotteluanalyysiin liittyvät läheisesti luokittelumenetelmät, joiden tehtävänä on osoittaa kullekin tutkittavalle yksilölle, esim. aikaisemmin suoritetun erotteluanalyysin pohjalla, ryhmä, johon yksilö lähinnä kuuluu (Rao //10//, Cooley-Lohnes //4//, Chapters 6, 7).

LÄHDELUETTELO

//1// Ahmavaara, Y., Transformation Analysis of Factorial Data. Ann. Acad. Sci. Fenn. Ser. B. Tom. 88, 2, 1954

//2// Ahmavaara, Y., On the Mathematical Theory of Transformation Analysis, Alkoholipoliittisen tutkimuslaitoksen julkaisuja n:o 1, 1963

//3// Anderson, T. W., An Introduction to Multivariate Statistical Analysis, Wiley, 1958

//4// Cooley-Lohnes, Multivariate Procedures for the Behavioral Sciences, Wiley, 1962

//5// Cramér, H., Mathematical Methods of Statistics, Princeton, 1959

//6// Elfving, G., Todennäköisyyslaskenta, Otava, 1956

//7// Harman, H. H., Modern Factor Analysis, University of Chicago Press, 1960

//8// Kullback, S., Information Theory and Statistics, Wiley, 1959

//9// Mustonen, S., Multiple Discriminant Analysis in Linguistic Problems, NordSAM 64, 1964

//10// Rao, C. R., Advanced Statistical Methods in Biometric Research, Wiley, 1952

//11// Thurstone, L. L., Multiple-Factor Analysis, University of Chicago Press, 1947