Aalto-yliopisto
Perustieteiden korkeakoulu
Teknillisen fysiikan ja matematiikan tutkinto-ohjelma

M-estimaatit

Kandidaatintyö
28.9.2014

Antti Melén

AALTO-YLIOPISTO PERUSTIETEIDEN KORKEAKOULU PL 11000, 00076 Aalto http://www.aalto.fi	KANDIDAATINTYÖN TIIVISTELMÄ
Tekijä: Antti Melén	
Työn nimi: M-estimaatit	
Tutkinto-ohjelma: Teknillisen fysiikan ja matematiikan tutkinto-ohjelma	
Pääaine: Systeemitieteet	Pääaineen koodi: F3010
Vastuuopettaja(t): Prof. Pauliina Ilmonen Ohjaaja(t): Prof. Pauliina Ilmonen	

Tiivistelmä:
Tässä työssä tutustutaan tutustutaan yleisesti M-estimaatteihin ja testataan niiden ominaisuuksia eri tavoin jakautuneiden aineistojen tapauksissa. M-estimaatit ovat vaihtoehto tilastollisessa analyysissä usein käytetyille perinteisille keskiarvovektorille ja kovarianssimatriisille lokaatiota ja hajontaa estimoidessa. Perinteisiä estimaatteja käytettäessä aineiston oletetaan usein olevan normaalijakautunut. Ei normaalijakautuneen aineiston tapauksessa on usein parempi käyttää muita estimaatteja, kuten M-estimaatteja.

Työssä tutustutaan erityisesti Hettmansperger-Randels M-estimaattiin ja sitä verrataan perinteisiin estimaatteihin normaalijakautuneen, elliptisesti jakautuneen sekä rippumattomien komponenttien aineiston tapauksissa. Testejä varten aineistot simuloidaan käyttäen R-ohjelmistoa. Lisäksi tarkastellaan estimaattien robustisuutta kun havaintoaineistoon lisätään poikkeavia havaintoja. Tuloksista voidaan päätellä M-estimaattien soveltuvan ei-normaalijakautuneen havaintoaineiston estimointiin perinteisiä estimaatteja paremmin. Niiden huomataan myös olevan robustimpia perinteisiin estimaatteihin verrattuna.

Päivämäärä: 28.9.2014	Kieli: suomi	Sivumäärä: 13 + 3
Avainsanat: M-estimaatti, Hettmansperger-Randels M-estimaatti, R-ohjelmisto, moniulotteinen aineisto, elliptinen jakauma, lokaatio- ja hajontasuureet, robustisuus, affiinisti ekvivariantti		

Sisältö

1 Johdanto **1**

2 Suureiden määrittely **1**

3 Erilaisia lokaatio- ja hajontasuureita **2**
3.1 Keskiarvovektori ja kovarianssimatriisi 2
3.2 M-estimaatit 3
3.3 One-step M-estimaatit 3

4 Simuloituja dataesimerkkejä **4**
4.1 Datan generoiminen 4
4.2 Tunnusluvut 6
4.3 Affiinimuunnos 8
4.4 Robustisuuden testaaminen 9
4.5 Kooste havainnoista 11

5 Yhteenveto **11**

A Liite - R-koodi **14**

1 Johdanto

Moniulotteista aineistoa kuvaamaan tarvitaan eri suureita. Tärkeää on tietää lokaatio, eli missä aineisto sijaitsee, sekä miten aineisto on haujautunut. Näitä suureita estimoidaan usein perinteisillä keskiarvovektorilla, sekä kovarianssimatriisilla. Perinteiset estimaatit sopivat estimointiin normaalijakautuneiden aineistojen tapauksessa. Ne ovat kuitenkin hyvin herkkiä poikkeaville havainnoille.[6] Tätä varten on kehitetty erilaisia robustimpia estimaattoreita kuten M-estimaatit [4, 5], S-estimaatit [1], sekä MCD-estimaatit [7]. Eri estimaatit voivat myös olla laskennallisesti tehokkaampia [5]. Tässä työssä keskitytään M-estimaattien tarkasteluun.
Lokaatio ja hajonta funktionaalit on usein standardoitu siten, että aineiston ollessa standardinormaalijakautunut, lokaatio on nollavektori ja hajonta on identiteettimatriisi. Elliptisellä aineistolla lokaatiofunktionaalit ovat yhtäsuuria ja hajontafunktionaalit vakiolla kertomista vaille samat. [6]
Tässä työssä tutustutaan yleisesti M-estimaatteihin ja valitaan yksi estimaatti, jonka ominaisuuksia testataan eri aineistojen tapauksissa. Luvussa 2 määritellään lokaatio- ja hajontasuureet. Luvussa 3 esitellään niille estimaatit ja erityisesti M-estimaatit. Lisäksi esitellään One-step M-estimaattien määritelmä. Luvussa 4 tutustutaan Hettmansperger-Randels M-estimaattin ominaisuuksiin eri aineistojen tapauksissa ja esitetään saadut tulokset. Lopuksi luvussa 5 esitellään yhteenveto.

2 Suureiden määrittely

Tässä luvussa määritellään affiinisti ekvivariantit lokaatio- ja hajontasuureet. Tämä luku perustuu kirjaan H. Oja: Multivariate Nonparametric Methods with R, [6].
Olkoon y p-ulotteinen satunnaismuuttuja, jonka kertymäfunktio on F_y. Lokaatiovektori on p-ulotteinen affiinisti ekvivariantti vektori siten, että

$$T(F_{Ay+b}) = AT(F_y) + b \tag{1}$$

kaikille vektoreille y, kaikille täyden asteen $p \times p$ matriiseille A sekä kaikille p-ulotteisille vektoreille b.
Hajontamatriisi on $p \times p$ positiivisesti definiitti matriisi, joka on affiinisti ekvivariantti siten, että

$$S(F_{Ay+b}) = AS(F_y)A^T \tag{2}$$

kaikille vektoreille y, kaikille täyden asteen $p \times p$ matriiseille A sekä kaikille p-ulotteisille vektoreille b.
Eli lokaatio ja hajonta mukautuvat käytettyyn koordinaattisysteemiin. Affiinisti ekvivarianttius on tärkeä ominaisuus tilastollisessa analyysissä siirryttäessä koordinaatistosta toiseen. Tämä tarkoittaa sitä, että jos esimerkiksi mittaamme painoa ja pituutta kilogrammoilla ja senttimetreillä tai paunoilla ja tuumilla, niin analyysin tulokset eivät riipu käytetyistä yksiköistä.

3 Erilaisia lokaatio- ja hajontasuureita

Tässä luvussa tarkastellaan erilaisia lokaatio- ja hajontasuureita. Tämä luku perustuu pääosin kirjaan H. Oja: Multivariate Nonparametric Methods with R, [6].

3.1 Keskiarvovektori ja kovarianssimatriisi

Olkoon x p-ulotteinen satunnaismuuttuja, jonka kertymäfunktio on F_x. Olkoon $(x_1, \ldots, x_n)$ otos jakaumasta F_x. Odotusarvoa $E[x] = \mu$ estimoidaan yleensä keskiarvovektorilla

$$\frac{1}{n}\sum_{i=1}^{n} x_i = \hat{\mu} \tag{3}$$

Satunnaismuuttujan x teoreettinen kovarianssamatriisi on

$$E[(x - E[x])(x - E[x])^T] = \Sigma \tag{4}$$

jonka estimaatti on

$$\frac{1}{n}\sum_{i=1}^{n}[(x_i - \hat{\mu})(x_i - \hat{\mu})^T] = \hat{\Sigma} \tag{5}$$

Yleensäkin teoreettisia suureita estimoidaan usein siten, että odotusarvot korvataan datasta lasketuilla vastaavilla keskiarvoilla.

3.2 M-estimaatit

M-funktionaalit määritellään seuraavalla tavalla

Lokaatio:

$$T(F_x) = E[[w_1(r)^{-1}] \cdot E[w_1(r)x]] \tag{6}$$

Hajonta:

$$S(F_x) = E[w_2(r)(x - T(F_x))(x - T(F_x))^T] \tag{7}$$

missä $w_1(r)$ ja $w_2(r)$ ovat ei negatiivisia funktioita ja

$$r = \|S(F_x)^{-\frac{1}{2}}(x - T(F_x))\|, \tag{8}$$

missä $\|\cdot\|$ on L_2 normi.

M-estimaatit saadaan korvaamalla yllä olevissa kaavoissa olevat odotusarvot niitä vastaavilla otoskeskiarvoilla.
M-estimaatit ovat siis painotettuja versioita odotusarvosta ja kovarianssimatriisista. Perinteiset keskiarvovektori ja kovarianssimatriisi saadaan M-estimaateista kun asetetaan $w_1 = 1$ ja $w_2 = 1$.
Tässä työssä esimerkkinä käytetään Hettmansperger-Randels M-estimaattia [2], joka määritellään $w_1(r) = 1/r$ ja $w_2(r) = p \cdot 1/r^2$, jossa $p = 3$ on skaalaustekijä (3 ulotteinen aineisto). Näiden lisäksi on olemassa useita muita eri funktionaaleja, joissa käytetään eri painofunktioita.

3.3 One-step M-estimaatit

Normaalien M-estimaattien lisäksi on olemassa One-step M-estimaatit. Ne lähtevät alkuarvoista T_1 ja S_1 (= esimerkiksi perinteinen keskiarvovektori, sekä kovarianssimatriisi), joille suoritetaan vain yksi laskuiteraatio. Näin ollen ne ovat laskennallisesti kevyempiä kuin normaalit M-estimaatit. Estimaatit määritellään seuraavasti

Lokaatio:

$$T_2(F_x) = E[[w_1(r_1)^{-1}] \cdot E[w_1(r_1)x]] \tag{9}$$

Hajonta:

$$S_2(F_x) = E[w_2(r_1)(x - T(F_x))(x - T(F_x))^T] \tag{10}$$

missä $w_1(r)$ ja $w_2(r)$ ovat ei negatiivisia funktioita ja

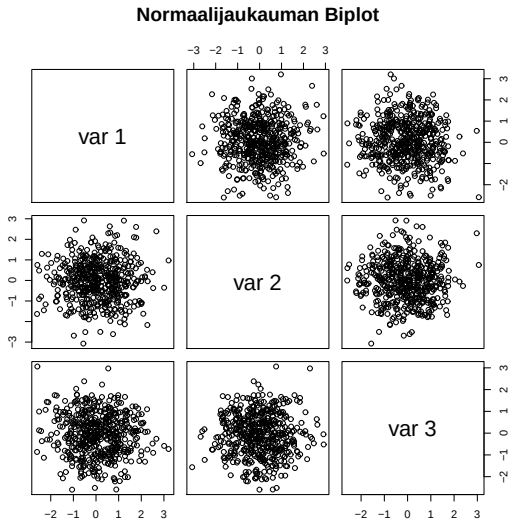
$$r_1 = \|S_1(F_x)^{-\frac{1}{2}}(x - T_1(F_x))\|. \tag{11}$$

4 Simuloituja dataesimerkkejä

Tässä osiossa tutustutaan aiemmin määritellyn Hettmansperger-Randels M-estimaattiin vertaamalla sen ominaisuuksia perinteiseen keskiarvovektoriin ja kovarianssimatriisiin kolmen erilaisen data-aineiston tapauksessa. Kaikki tulokset lasketaan käyttäen R-ohjelmistoa.

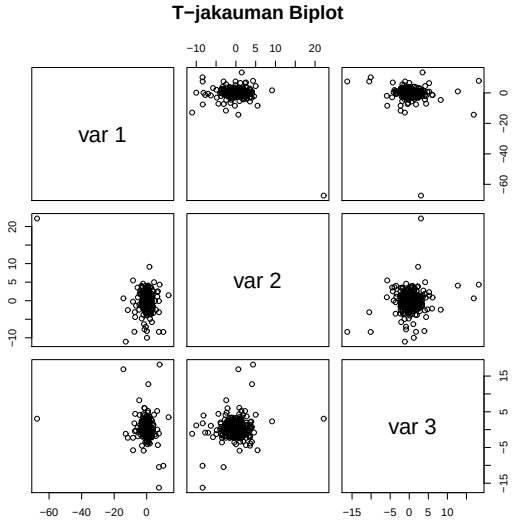
4.1 Datan generoiminen

M-estimaattien tutkimista varten simuloidaan kolme kolmiulotteista esimerkkiaineistoa, joista jokaisessa on 500 datapistettä. Ensimmäinen esimerkki on standardinormaalijakaumasta generoitu aineisto (kuva 1).



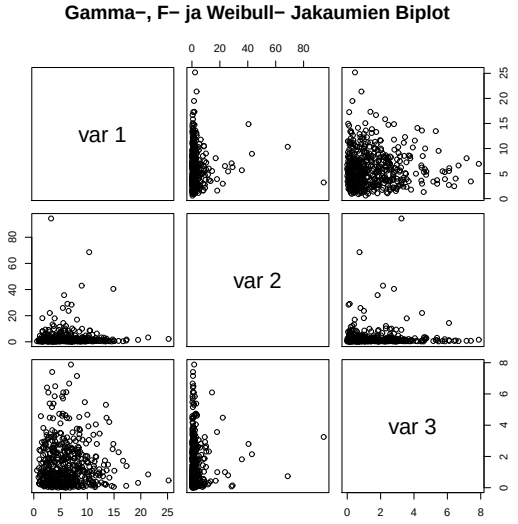
Kuva 1: Normaalijakautunut aineisto

Toisena esimerkkinä on elliptinen aineisto Studentin t-jakaumasta vapausasteella 2 (Shape-matriisina on identiteettimatriisi) (kuva 2).



Kuva 2: Elliptinen aineisto

Viimeisenä komponenteiltaan riippumaton aineisto, jonka generoimisessa on käytetty Gamma-jakaumaa (parametreillä 3,2), F-jakaumaa (parametreillä 20,2) sekä Weibull jakaumaa (parametreillä 1,1.5).



Kuva 3: Riippumattomien komponenttien aineisto

4.2 Tunnusluvut

Lasketaan jokaisesta aineistoista niiden keskiarvovektori, kovarianssimatriisi sekä Hettmansperger-Randels M-estimaatti. Ensimmäiselle aineistolle saadaan seuraavat tulokset:

Keskiarvo:

$$\hat{\mu} = \begin{bmatrix} 0.076887564 & -0.005276685 & -0.059755290 \end{bmatrix}$$

Kovarianssi:

$$\hat{\Sigma} = \begin{bmatrix} 1.03370988 & 0.018504381 & -0.057893297 \\ 0.01850438 & 0.977683673 & -0.005987961 \\ -0.05789330 & -0.005987961 & 0.943944236 \end{bmatrix}$$

Vastaavaksi M-estimaatiksi saadaan

$$\hat{T} = \begin{bmatrix} 0.06226137 & -0.02855955 & -0.04184608 \end{bmatrix}$$

$$\hat{S} = \begin{bmatrix} 0.98426098 & -0.01487838 & -0.02442652 \\ -0.01487838 & 0.97241067 & -0.08355984 \\ -0.02442652 & -0.08355984 & 0.99104062 \end{bmatrix}$$

Nähdään, että M-estimaatti ei ero kovinkaan paljoa perinteisistä estimaateista normaalijakautuneen aineiston tapauksessa. Kun aineisto on normaalijakautunut, molemmat estimaatit estimoivat samaa populaatiosuuretta [6].

Toiselle aineistolle saadaan seuraavat tulokset:

Keskiarvo:

$$\hat{\mu} = \begin{bmatrix} -0.1538575 & -0.1046966 & 0.1980994 \end{bmatrix}$$

Kovarianssi:

$$\hat{\Sigma} = \begin{bmatrix} 14.067404 & -2.7959476 & -1.2084514 \\ -2.795948 & 4.8016840 & 0.7569579 \\ -1.208451 & 0.7569579 & 4.9598101 \end{bmatrix}$$

Vastaavaksi M-estimaatiksi saadaan

$$\hat{T} = \begin{bmatrix} -0.002978095 & -0.041915712 & 0.074566106 \end{bmatrix}$$

$$\hat{S} = \begin{bmatrix} 5.7421334 & -0.3547696 & -0.2236719 \\ -0.3547696 & 6.3672046 & -0.1189049 \\ -0.2236719 & -0.1189049 & 6.0388496 \end{bmatrix}$$

Näemme, että estimaatit ovat aika kaukana toisistaan. Koska kyseessä on elliptinen jakauma, molemmat estimaatit estimoivat jakauman shape-matriisia - vakiota vaille [5, 6]. Tässä esimerkissä shape-matriisi oli identiteettimatriisi. Huomaamme, että M-estimaatti on lähempänä diagonaalimatriisia, jolla on samat diagonaalialkiot. Tässä esimerkissä M-estimaatti näyttäisi siis toimivan paremmin. Tämä johtunee siitä, että Hettmansperger-Randels M-estimaatti on robustimpi ja toimii siksi paremmin kun jakauma on paksuhäntäinen.

Kolmannelle aineistolle saadaan seuraavat tulokset:

Keskiarvo:

$$\hat{\mu} = \begin{bmatrix} 6.060987 & 2.713904 & 1.461540 \end{bmatrix}$$

Kovarianssi:

$$\hat{\Sigma} = \begin{bmatrix} 12.1150905 & 0.3223496 & -0.0719036 \\ 0.3223496 & 45.9558258 & 0.3120292 \\ -0.0719036 & 0.3120292 & 2.0103782 \end{bmatrix}$$

Vastaavaksi M-estimaatiksi saadaan

$$\hat{T} = \begin{bmatrix} 5.693537 & 1.359861 & 1.187881 \end{bmatrix}$$

$$\hat{S} = \begin{bmatrix} 24.9206231 & -0.1018769 & 0.4579609 \\ -0.1018769 & 3.5809983 & 0.1935345 \\ 0.4579609 & 0.1935345 & 3.5081230 \end{bmatrix}$$

Kolmannen, riippumattoman aineiston tapauksessa hajontaestimaattien pitäisi olla lähellä diagonaalimatriisia ja tässä esimerkissä molemmat hajontaestimaatit ovat kohtuulllisen lähellä diagonaalimatriisia. Sekä lokaatio-, että hajontaestimaatit ovat kuitenkin aika kaukana toisistaan. Tämä johtuu siitä, että estimaatit estimoivat eri populaatiosuureita. Hajontamatriisit estimoivat yleisesti samaa populaatiosuuretta vain silloin kun jakauma on elliptinen tai normaalijakauma [6]. Lokaatioestimaatit estimoivat yleisesti samaa populaatiosuuretta vain silloin kun jakauma on symmetrinen [6]. Tässä kolmannessa esimerkissä jakaumat eivät ole elliptisesti jakautuneita ja ne ovat vinoja. Sen vuoksi näemme selvän eron estimaattien välillä.

4.3 Affiinimuunnos

Tehdään aiemmin generoiduille aineistoille affiinimuunnos ja lasketaan uudestaan keskiarvovektori, kovarianssimatriisi sekä Hubertin M-estimaatti. Valitaan esimerkkiä varten (satunnaiset) matriisi A ja vektori b ja määritellään affiinimuunnos $x' = Ax + b$, jossa

$$A = \begin{bmatrix} 1 & -2 & -2 \\ 0 & -1 & 1 \\ 4 & 5 & -5 \end{bmatrix}$$

$$b = \begin{bmatrix} 0.08 & -0.20 & 1.00 \end{bmatrix}$$

Ensimmäiselle datajoukolle saadaan affiinimuunnoksen jälkeen seuraavat keskiarvovektori ja kovarianssimatriisi

$$\hat{\mu} = \begin{bmatrix} -0.0821336 & -0.6472749 & 1.1397246 \end{bmatrix}$$

$$\hat{\Sigma} = \begin{bmatrix} 15.67367 & 16.99059 & -20.19914 \\ 16.99059 & 30.00289 & -20.50133 \\ -20.19914 & -20.50133 & 27.53913 \end{bmatrix}$$

sekä M-estimaatit:

$$\hat{T} = \begin{bmatrix} -0.02512294 & -0.50519359 & 1.05614809 \end{bmatrix}$$

$$\hat{S} = \begin{bmatrix} 16.64550 & 18.27469 & -21.82091 \\ 18.27469 & 30.95009 & -22.64698 \\ -21.82091 & -22.64698 & 30.09205 \end{bmatrix}$$

Toiselle aineistolle:

$$\hat{\mu} = \begin{bmatrix} 0.7185399 & 1.2029084 & 0.2125216 \end{bmatrix}$$

$$\hat{\Sigma} = \begin{bmatrix} 83.75675 & 74.45486 & -111.38926 \\ 74.45486 & 190.48221 & -64.95774 \\ -111.38926 & -64.95774 & 164.51173 \end{bmatrix}$$

M-estimaatit:

$$\hat{T} = \begin{bmatrix} 0.3752863 & 0.2207024 & 0.5912099 \end{bmatrix}$$

$$\hat{S} = \begin{bmatrix} 100.5744 & 110.7941 & -130.1839 \\ 110.7941 & 184.5504 & -135.5590 \\ -130.1839 & -135.5590 & 178.4417 \end{bmatrix}$$

Kolmannelle aineistoille:

$$\hat{\mu} = \begin{bmatrix} 11.987148 & -7.728176 & -15.715772 \end{bmatrix}$$

$$\hat{\Sigma} = \begin{bmatrix} 43.70591 & 14.62263 & -61.93253 \\ 14.62263 & 144.28282 & -44.63463 \\ -61.93253 & -44.63463 & 138.82788 \end{bmatrix}$$

M-estimaatit:

$$\hat{T} = \begin{bmatrix} 10.525062 & -7.007529 & -14.966619 \end{bmatrix}$$

$$\hat{S} = \begin{bmatrix} 84.71428 & 18.27507 & -125.28494 \\ 18.27507 & 179.46450 & 10.33376 \\ -125.28494 & 10.33376 & 198.59795 \end{bmatrix}$$

Tällä tavoin saadut tulokset ovat kaikissa kolmessa tapauksessa samat kuin jos affiinimuunnos olisi tehty alkuperäisistä aineistoista lasketuille estimaateille. Estimaatit ovat siis affiinisti ekvivariantteja ja siirtyminen eri koordinaatistojen välillä voidaan tehdä.

4.4 Robustisuuden testaaminen

Testataan M-estimaattien robustisuutta korvaamalla generoiduista aineistoista 25 datapistettä selkeästi poikkeavilla havainnoilla ja vertaamalla tuloksia alkuperäisestä datasta laskettuihin estimaatteihin. Ensimmäisellä aineistolla saadaan seuraavat tulokset:

$$\hat{\mu} = \begin{bmatrix} 0.3655995 & 0.2422765 & 0.1822014 \end{bmatrix}$$

$$\hat{\Sigma} = \begin{bmatrix} 2.386117 & 1.245660 & 1.209738 \\ 1.245660 & 2.106753 & 1.134744 \\ 1.209738 & 1.134744 & 2.137444 \end{bmatrix}$$

Ja vastaavat M-estimaatit:

$$\hat{T} = \begin{bmatrix} 0.14446803 & 0.03208035 & 0.01360665 \end{bmatrix}$$

$$\hat{S} = \begin{bmatrix} 1.5815596 & 0.10755086 & 0.11747315 \\ 0.1075509 & 1.47993225 & 0.01120652 \\ 0.1174731 & 0.01120652 & 1.54736701 \end{bmatrix}$$

Ensimmäisen aineiston tulosten perusteella muutos M-estimaateissa on pienempi kuin perinteisissä estimaateissa. Näin ollen M-estimaattien voidaan sanoa olevan robustimpia kuin perinteiset estimaatit.

Toiselle aineistolle:

$$\hat{\mu} = \begin{bmatrix} 0.06538494 & 0.15666198 & 0.39328526 \end{bmatrix}$$

$$\hat{\Sigma} = \begin{bmatrix} 15.069643 & -1.635493 & -0.383272 \\ -1.635493 & 5.893987 & 1.791010 \\ -0.383272 & 1.791010 & 5.356110 \end{bmatrix}$$

Ja vastaavat M-estimaatit:

$$\hat{T} = \begin{bmatrix} 0.06141576 & 0.03462676 & 0.12191922 \end{bmatrix}$$

$$\hat{S} = \begin{bmatrix} 6.5239180 & 0.4043855 & 0.1577146 \\ 0.4043855 & 7.1415466 & 0.7633116 \\ 0.1577146 & 0.7633116 & 6.8304244 \end{bmatrix}$$

Myös toisen aineiston tapauksessa M-estimaatti antaa robustimman arvion. Lokaatiovektori on lähempänä alkuperäistä ja myös muutokset hajontamatriisissa ovat pienempiä.

Kolmannelle aineistolle:

$$\hat{\mu} = \begin{bmatrix} 6.537060 & 3.092498 & 1.910397 \end{bmatrix}$$

$$\hat{\Sigma} = \begin{bmatrix} 15.583747 & 3.616374 & 3.670815 \\ 3.616374 & 47.707907 & 3.218542 \\ 3.670815 & 3.218542 & 5.482011 \end{bmatrix}$$

Ja vastaavat M-estimaatit

$$\hat{T} = \begin{bmatrix} 5.794452 & 1.448990 & 1.300048 \end{bmatrix}$$

$$\hat{S} = \begin{bmatrix} 31.6630812 & 0.6990734 & 1.590197 \\ 0.6990734 & 5.6016470 & 1.109351 \\ 1.5901967 & 1.1093514 & 5.360370 \end{bmatrix}$$

Myös kolmannella aineistolla havaitaan samat tulokset kuin kahdella edellisesä aineistolla. Kaikista kolmesta aineistosta nähdään käytetyn Hettmansperger-Randels M-estimaattien olevan robustimpia kuin perinteiset estimaattorit.

4.5 Kooste havainnoista

Edellä tehtyjen havaintojen perusteella voidaan todeta M-estimaateilla saatujen tulosten eroavan perinteisistä estimaateista kun havaintoaineisto ei ole normaalijakautunut. Lisäksi M-estimaatit ovat perinteisten estimaattien tavoin affiinisti ekvivariantteja, eli ne mukautuvat käytettävään kordinaatistoon. Esimerkkinä käytetty Hettmansperger-Randels M-estimaatti on robustimpi kuin perinteiset estimaatit. Parametreja w_1 ja w_2 muuttamalla voidaan valita eri tavoin jakautuneiden aineistojen vertailuun parhaiten sopiva M-estimaatti.

5 Yhteenveto

Moniulotteisessa data-analyysissä tarvitaan erilaisia lokaatio- ja hajontafunktionaaleja, sekä niitä vastaavia estimaatteja, jo ihan senkin vuoksi, että erilaiset estimaatit kuvaavat eri tavoin jakaumien ominaisuuksia. Kovarianssimatriisi ja keskiarvovektori kuvaavat hyvin normaalijakautuneen aineiston

hajontaa ja paikkaa. Monien jakaumien kuvaamiseen soveltuvat kuitenkin paremmin jotkin muut hajonta- ja lokaatiosuureet. Tämän lisäksi on olemassa sellaisia menetelmiä moniulotteisten aineistojen analysoimiseen, joissa käytetään samanaikaisesti useampia lokaatio- ja/tai hajontasuureita. Tälläisiä menetelmiä käytetään esimerkiksi riippumattomien komponenttien analyysissä ja invarianttien koordinaattien valinnassa [3].

Viitteet

[1] L. Davies: Asymptotic behavior of S-estimates of multivariate location parameters and dispersion matrices, Annals of Statistics 15, 1269–1292, 1987.

[2] T. P. Hettmansperger and R.H. Randles: A Practical Affine Equivariant Multivariate Median, Biometrica 89, p. 851–860, 2002.

[3] P. Ilmonen, H. Oja and R. Serfling: On Invariant Coordinate System (ICS) Functionals, International Statistical Review 80(1), p.93–110, 2012.

[4] R. A. Maronna: Robust M-estimators of multivariate location and scatter, Annals of Statistics 4, p. 51–67, 1976.

[5] R. A. Maronna, R. D. Mardin, V. J. Yohai: Robust Statistics: Theory and Methods, John Wiley and Sons, Chichester, 2006.

[6] H. Oja: Multivariate Nonparametric Methods with R, Springer-Verlag, New York, 2010.

[7] P. J. Rousseeuw: Multivariate estimation with high breakdown point, Mathematical Statistics and Applications 8 (W. Grossmann, G. Pug, I. Vincze, W. Wertz, eds.), p. 283–297, 1985.

A Liite - R-koodi

```
#####################
#Datan generointi esimerkkiä varten
#Generoidaan n datapistettä.Ja plotataan datat

n <- 500

###Normaalijakauma
library(MASS)
set.seed(999)
A <-matrix(c(1,0,0,0,1,0,0,0,1), ncol=3)
Narvot <- mvrnorm(n, c(0,0,0), A)
#pairs(Narvot, main = "Normaalijaukauman Biplot")

###Studentin t-jakauma
install.packages("mvtnorm")
library(mvtnorm)
#Studentin t-jakauma vapausasteella 2
set.seed(50)
A <-matrix(c(1,0,0,0,1,0,0,0,1), ncol=3)
Narvot <- rmvt(n, A,df=2)
#pairs(Narvot, main = "T-jakauman Biplot")

###Kolmas jakauman (Gamma, F, Weibull)
set.seed(500)
Narvot <- matrix(,n,3)
#Gamma-jakauma parametreilla shape 3 scale 2
Narvot[,1] <- rgamma(n,3,,2)
#F-jakauma vapausasteilla 20, 2
Narvot[,2] <- rf(n,20,3)
#Weibull jakauma scale=1, shape = 1.5
Narvot[,3] <- rweibull(n,1,1.5)
#pairs(Narvot, main = "Gamma-, F- ja Weibull- Jakaumien Biplot")


##################
#Affiinimuunnos datalle
D <-matrix(c(1,-2,-2,0,-1,1,4,5,-5), ncol=3)
```

```
b <- c(0.08, -0.2, 1)
for (i in 1:n) {
Narvot[i,] = t(D %*% Narvot[i,]) + b
}

#Estimaattien muuntaminen uuteen koordinaatistoon
Lokaatio <- t(D %*% u) + b
Hajonta <- D %*% Sigma %*% t(D)
Affiinit <- list(loc=Lokaatio, scat = Hajonta)
Affiinit

######################
#Datapisteiden korvaaminen poikkeavilla
#arvoilla robustisuuden testaamista varten
#Jakauma 1
Narvot[476:500,] <- mvrnorm(25, c(5,5,5), A)
#Jakauma 2
Narvot[476:500,] <- mvrnorm(25, c(5,5,5),A)
#Jakauma 3
Narvot[476:500,] <- mvrnorm(25, c(15,10,10),A)


######################
#Initial values ennen looppia
#Keskiarvovektorin muodostaminen u
u <- colMeans(Narvot)

#kovarianssimatriisi Sigma (harhaton n-1 jakajana)
B <- matrix(, nrow= n, ncol = 3)
B[,1] <- Narvot[,1] - u[1]
B[,2] <- Narvot[,2] - u[2]
B[,3] <- Narvot[,3] - u[3]
Sigma <- (t(B) %*% B)/(n-1)

INI <- list(loc=u, scat = Sigma)


#############################
#Tästä alkaa looppi
#muodon vuoksi 100 kierrosta
for (k in 1:100) {
```

```
#Lasketaan mahaloboniksen mitta vektori r
r = vector(mode = "numeric", length = n)
for (i in 1:n) {
r[i] = sqrt( t(B[i,]) %*% solve(Sigma) %*% B[i,] )
}

##Lasketaan wyksi ja wkaksi
wyksi = vector(mode = "numeric", length = n)
for (i in 1:n) {
wyksi[i] = (1/r[i])
}

wkaksi = vector(mode = "numeric", length = n)
for (i in 1:n) {
wkaksi[i] =  (3/(r[i])^2)
}

#Päivitetään uudet estimaatit

#lokaatio
pwyksi <- mean(wyksi)
apu <- wyksi * Narvot
apu2 <- c(colMeans(apu))
u <- pwyksi^(-1) * apu2

#hajonta
#(y-M) matriisi
B <- matrix(, nrow= n, ncol = 3)
B[,1] <- Narvot[,1] - u[1]
B[,2] <- Narvot[,2] - u[2]
B[,3] <- Narvot[,3] - u[3]

#hajontamatriisi
apu3 = wkaksi * B
Sigma = (t(apu3) %*% B)/n

#Tallennetaan tulokset
RES <- list(loc=u, scat = Sigma)

}
```