Aalto-yliopisto
Perustieteiden
korkeakoulu

M-estimaatit (valmiin työn esittely)

Antti Melén

29.09.2014

Ohjaaja: Pauliina Ilmonen

Valvoja: Pauliina Ilmonen

Systeemianalyysin
laboratorio

Tausta

- Moniulotteisia aineistoja tutkiessa halutaan tietää aineiston lokaatio ja hajonta
- Usein näitä suureita estimoidessa käytetään perinteistä keskiarvovektoria ja kovarianssimatriisia
- Perinteiset estimaatit (keskiarvovektori ja kovarianssimatriisi) estimoivat suureita hyvin vain normaalijakautuneiden aineistojen tapauksissa
  - Ei normaalijakautunut aineisto → ongelma
- Tätä ongelmaa varten kehitetty parempia estimaatteja kuten M-estimaatit

Tavoitteet

- Tutustutaan yleisesti M-estimaatteihin
- Vertaillaan Hettmansperger-Randels M-estimaattia perinteisiin estimaatteihin kolmen eri simuloidun aineiston tapauksessa sekä testataan sen ominaisuuksia

Tutkimusmenetelmät

- Ohjelmoidaan M-estimaatit käyttäen R-ohjelmistoa
- Simuloidaan 3 aineistoa
  - Normaalijakautunut
  - Elliptinen
  - Komponenteiltaan riippumaton

Tulokset: Tunnusluvut (1/2)

	Perinteiset estimaatit	M-estimaatit
Normaalijakautunut	$\hat{\mu} = [0.076887564 \quad -0.005276685 \quad -0.059755290]$ $\hat{\Sigma} = \begin{bmatrix} 1.03370988 & 0.018504381 & -0.057893297 \\ 0.01850438 & 0.977683673 & -0.005987961 \\ -0.05789330 & -0.005987961 & 0.943944236 \end{bmatrix}$	$\hat{T} = [0.06226137 \quad -0.02855955 \quad -0.04184608]$ $\hat{S} = \begin{bmatrix} 0.98426098 & -0.01487838 & -0.02442652 \\ -0.01487838 & 0.97241067 & -0.08355984 \\ -0.02442652 & -0.08355984 & 0.99104062 \end{bmatrix}$
Elliptinen	$\hat{\mu} = [-0.1538575 \quad -0.1046966 \quad 0.1980994]$ $\hat{\Sigma} = \begin{bmatrix} 14.067404 & -2.7959476 & -1.2084514 \\ -2.795948 & 4.8016840 & 0.7569579 \\ -1.208451 & 0.7569579 & 4.9598101 \end{bmatrix}$	$\hat{T} = [-0.002978095 \quad -0.041915712 \quad 0.074566106]$ $\hat{S} = \begin{bmatrix} 5.7421334 & -0.3547696 & -0.2236719 \\ -0.3547696 & 6.3672046 & -0.1189049 \\ -0.2236719 & -0.1189049 & 6.0388496 \end{bmatrix}$
Komponenteittain riippumaton	$\hat{\mu} = [6.060987 \quad 2.713904 \quad 1.461540]$ $\hat{\Sigma} = \begin{bmatrix} 12.1150905 & 0.3223496 & -0.0719036 \\ 0.3223496 & 45.9558258 & 0.3120292 \\ -0.0719036 & 0.3120292 & 2.0103782 \end{bmatrix}$	$\hat{T} = [5.693537 \quad 1.359861 \quad 1.187881]$ $\hat{S} = \begin{bmatrix} 24.9206231 & -0.1018769 & 0.4579609 \\ -0.1018769 & 3.5809983 & 0.1935345 \\ 0.4579609 & 0.1935345 & 3.5081230 \end{bmatrix}$

Tulokset: Tunnusluvut (2/2)

- Normaalijakautuneen aineiston tapauksesessa perinteiset sekä M-estimaatit lähes samat
- Elliptisessä aineistossa M-estimaatin hajontamatiriisi selvästi lähempänä estimoitavaa jakauman shape-matriisia (tässä identiteettimatriisi)
- Kolmannelle aineistolle hajontaestimaatit ovat kohtuulllisen lähellä diagonaalimatriisia. Erot estimaattien välillä kuitenkin suuria

Affinisti ekvivariantti

- Tehdään aineistolla affiinimuunnos $x' = Ax + b$
- Lasketaan estimaatit muunnetusta datasta ja huomataan niiden olevan samat kuin, jos affiinimuunnos olisi tehty suoraan estimaateille
  - –→M-estimaatit ovat affiinisti ekvivariantteja eli ne mukautuvat käytettävissä olevaan koordinaatistoon

Robustisuuden testaaminen (1/2)

	Perinteiset estimaatit	M-estimaatit
Normaalijakautunut	$\hat{\mu} = [0.3655995 \quad 0.2422765 \quad 0.1822014]$ $\hat{\Sigma} = \begin{bmatrix} 2.386117 & 1.245660 & 1.209738 \\ 1.245660 & 2.106753 & 1.134744 \\ 1.209738 & 1.134744 & 2.137444 \end{bmatrix}$	$\hat{T} = [0.14446803 \quad 0.03208035 \quad 0.01360665]$ $\hat{S} = \begin{bmatrix} 1.5815596 & 0.10755086 & 0.11747315 \\ 0.1075509 & 1.47993225 & 0.01120652 \\ 0.1174731 & 0.01120652 & 1.54736701 \end{bmatrix}$
Elliptinen	$\hat{\mu} = [0.06538494 \quad 0.15666198 \quad 0.39328526]$ $\hat{\Sigma} = \begin{bmatrix} 15.069643 & -1.635493 & -0.383272 \\ -1.635493 & 5.893987 & 1.791010 \\ -0.383272 & 1.791010 & 5.356110 \end{bmatrix}$	$\hat{T} = [0.06141576 \quad 0.03462676 \quad 0.12191922]$ $\hat{S} = \begin{bmatrix} 6.5239180 & 0.4043855 & 0.1577146 \\ 0.4043855 & 7.1415466 & 0.7633116 \\ 0.1577146 & 0.7633116 & 6.8304244 \end{bmatrix}$
Komponenteittain riippumaton	$\hat{\mu} = [6.537060 \quad 3.092498 \quad 1.910397]$ $\hat{\Sigma} = \begin{bmatrix} 15.583747 & 3.616374 & 3.670815 \\ 3.616374 & 47.707907 & 3.218542 \\ 3.670815 & 3.218542 & 5.482011 \end{bmatrix}$	$\hat{T} = [5.794452 \quad 1.448990 \quad 1.300048]$ $\hat{S} = \begin{bmatrix} 31.6630812 & 0.6990734 & 1.590197 \\ 0.6990734 & 5.6016470 & 1.109351 \\ 1.5901967 & 1.1093514 & 5.360370 \end{bmatrix}$

Robustisuuden testaaminen (2/2)

- Korvataan 5% datapisteistä poikkeavilla havainnoilla
- Kaikkien kolmen aineiston kohdalla Hettmansperger-Randels M-estimaatti antaa tulokset jotka ovat lähempänä alkuperäisestä datasta saatuja estimaatteja
- Testien perusteella voidaan todeta M-estimaattien olevan robustimpia

Johtopäätökset

- Perinteiset lokaatio ja hajonta estimaatit soveltuvat normaalijakautuneen aineiston kuvaamiseen
- Muullatavoin jakautuneiden tai poikkeavia havaintoja sisältävien jakaumien estimoinnissa käytettävä niihin paremmin sopivia estimaatteja kuten M-estimaatit

Lähteitä

- Multivariate Nonparametric Mehtods with R
  - Oja, 2010
- Robust Statistics: Theory and Methods
  - R. A. Maronna, R. D. Mardin, V. J. Yohai, 2006