Aalto-yliopisto
Perustieteiden korkeakoulu

Ryhmäfaktorianalyysi neurotiedesovelluksissa

(Valmiin työn esittely)

Sami Remes

11.06.2012

Ohjaaja: TkT Arto Klami

Valvoja: Prof. Harri Ehtamo

Systeemianalyysin laboratorio

Ongelman asettelu

Halutaan tutkia datajoukkojen tai muuttujien välisiä riippuvuuksia

Klassisia menetelmiä mm. kanoninen korrelaatio (CCA) ja faktorianalyysi

CCA:ssa etsitään satunnaisvektoreille $\boldsymbol{x}$ ja $\boldsymbol{y}$ lineaariset muunnokset, joiden korrelaatio suurin:

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \operatorname{cor}\left(\boldsymbol{\alpha}^T\boldsymbol{x}, \boldsymbol{\beta}^T\boldsymbol{y}\right)$$

Vektorit $\boldsymbol{\alpha}$ ja $\boldsymbol{\beta}$ antavat kiinnostavaa tietoa $\boldsymbol{x}$:n ja $\boldsymbol{y}$:n muuttujien riippuvuuksista

Ongelman asettelu (cont.)

CCA esimerkiksi aineistolle, jossa datajoukot

1. Oppilaan pituus, älykkyysosamäärä ym. ominaisuuksia
2. Arvosanoja eri aineissa

Saattaisi löytää esimerkiksi korkean painon älykkyysosamäärälle ja menestykselle luonnontieteissä

CCA rajoittuu vain kahden joukon tarkasteluun eikä sovellu hyvin korkeaulotteisille aineistoille

$N \gg D$, missä N oppilaiden/näytteiden lukumäärä, D eri ominaisuuksien määrä

Bayesilainen lähestymistapa

Tavanomaisessa tilastollisessa mallissa parametreilla Θ ei todennäköisyyksiä

Bayesilaisessa mallissa parametreille todennäköisyysjakauma Bayesin kaavan avulla

$$p(\Theta|Y) = \frac{p(Y|\Theta)p(\Theta)}{p(Y)} \propto \mathit{uskottavuus} \times \mathit{priori}$$

Jakauma usein vaikea laskea: approksimaatiot

Ryhmäfaktorianalyysi

Virtanen, Klami, Khan, Kaski, AISTATS 2012

Malli yleistää faktorianalyysin useille datajoukoille muuttujien sijaan

Malli yhdistetylle datalle $\boldsymbol{Y} = [\boldsymbol{Y}_1 \ldots \boldsymbol{Y}_M] \in \mathbb{R}^{N \times D}$

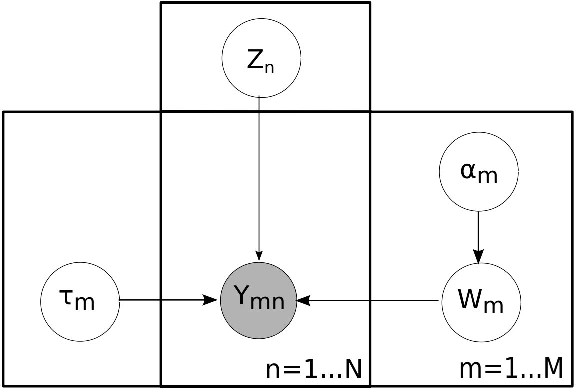
$$\boldsymbol{y}_n = \boldsymbol{W}\boldsymbol{z}_n + \varepsilon_n, \quad n = 1 \ldots N$$

Valitsemalla sopiva rakenne $\boldsymbol{W}$:lle mallinnetaan datajoukkojen väliset riippuvuudet piilomuuttujien avulla

$$\boldsymbol{W} = \begin{pmatrix} \boldsymbol{W}_1 & \boldsymbol{V}_1 & \boldsymbol{0} \\ \boldsymbol{W}_2 & \boldsymbol{0} & \boldsymbol{V}_2 \end{pmatrix}$$

Jos sarake datajoukon kohdalta lähes nolla, vastaava piilomuuttuja ei selitä vaihtelua datajoukossa

Ryhmäfaktorianalyysi (cont.)



Ryhmäfaktorianalyysi (cont.)

Malli toimii myös korkeaulotteiselle aineistolle, jolle $D \gg N$

Tarvittava määrä komponentteja valitaan automaattisesti, jos maksimimäärä riittävän suuri

Erikoistapauksina saadaan (bayesilaiset versiot)

- Pääkomponenttianalyysi ($M = 1$)
- Kanoninen korrelaatioanalyysi ($M = 2$)
- Faktorianalyysi (yksiulotteiset datajoukot)

Variaatioapproksimaatio

Posteriorin laskemiseksi approksimaatio q olettaa ylimääräisiä riippumattomuuksia muuttujien välillä

$$p(\Theta|\boldsymbol{Y}) \approx q(\Theta) = \prod_{j=1}^{K} q(\theta_j)$$

Ryhmäfaktorianalyysille

$$q(\Theta) = \left[\prod_{m=1}^{M} q(\boldsymbol{\alpha}_m) q(\tau_m)\right] \left[\prod_{d=1}^{D} q(\boldsymbol{w}_d)\right] \prod_{n=1}^{N} q(\boldsymbol{z}_n)$$

Variaatioapproksimaatio (cont.)

Minimoidaan Kullback-Leibler -divergenssi approksimaation ja posteriorin välillä

Ratkaisuna saadaan jakaumat, joiden parametrit riippuvat toistensa odotusarvoista

Yleensä riippuvuudet odotusarvoista liian vaikeita; ei analyyttista ratkaisua yhtälöryhmälle

Alustamalla jakaumien parametrit esim. satunnaisesti ja laskemalla parametrit uudestaan, saadaan algoritmi, joka konvergoi lokaaliin optimiin

Vertailua CCA:han

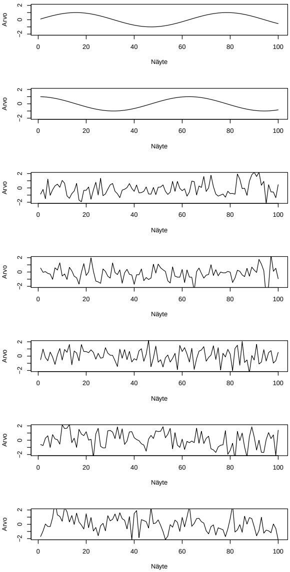
Keinotekoinen aineisto

- 7 piilomuuttujaa; 3 jaettua, 2+2 vain yhdelle datajoukolle
- Piilomuuttujista rakennetaan 15 ja 20 ulotteiset datajoukot satunnaisella $\boldsymbol{W}$:lla

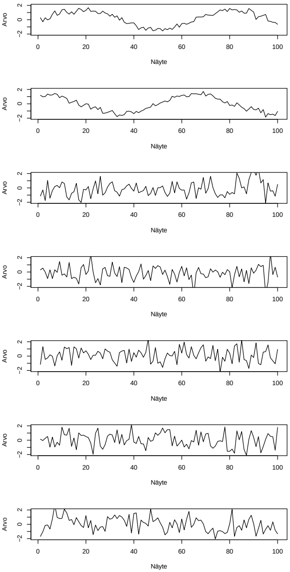
Vertaillaan ryhmäfaktorianalyysia ja tavallista CCA:ta; tavoitteena näyttää mallin toimivan oikein tässä erikoistapauksessa

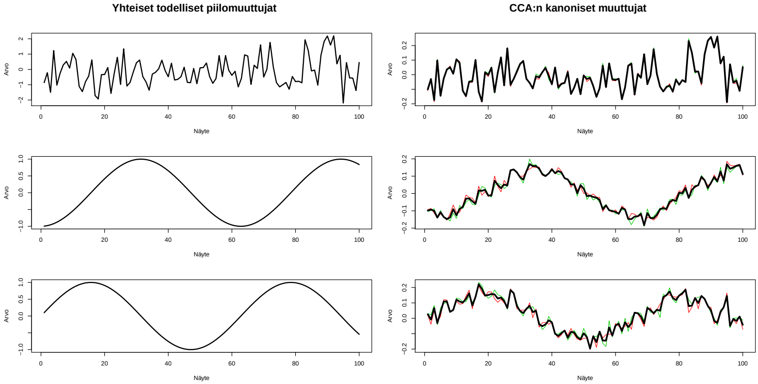
Molemmat menetelmät löytävät oikeat piilomuuttujat (CCA etsii vain jaetut)

Todelliset piilomuuttujat



Estimoidut aktiiviset piilomuuttujat





Yhteiset todelliset piilomuuttujat
CCA:n kanoniset muuttujat
Arvo
Näyte

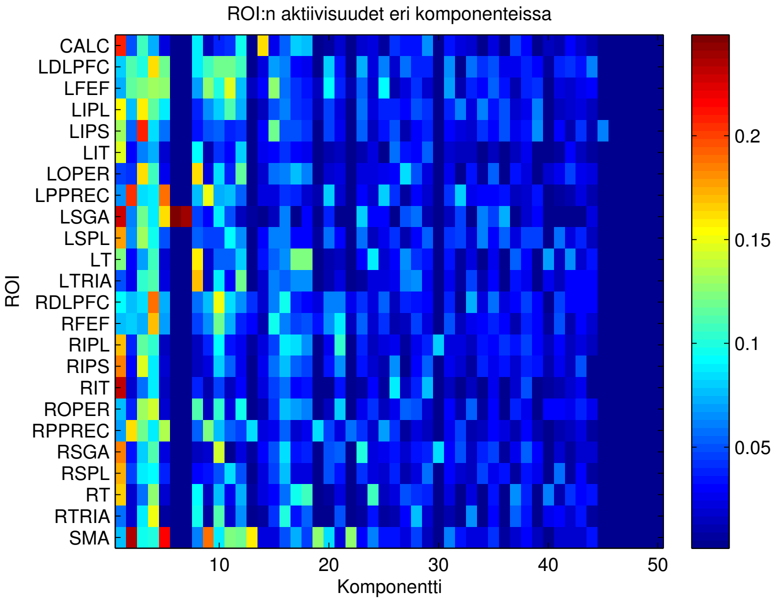
fMRI-datan analyysi

Dataa kokeesta, jossa koehenkilölle näytettiin kuvia ja tekstiä

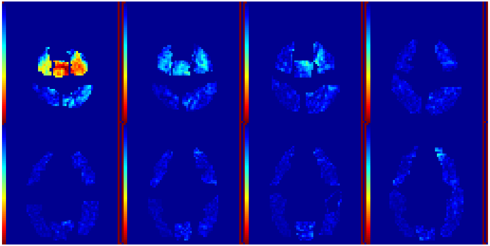
Voidaan jakaa 25 anatomisesti eroteltuun aivoalueeseen

Tavoitteena tutkia alueiden välisiä riippuvuuksia ts. mitkä alueet toimivat yhdessä

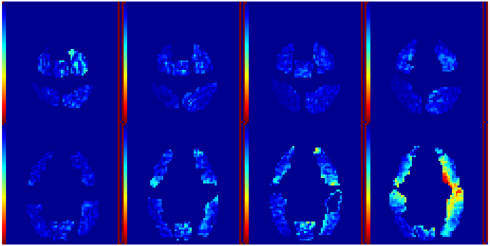
Malli löytää joitain potentiaalisesti mielenkiintoisia komponentteja


ROI:n aktiivisuudet eri komponenteissa
CALC
LDLPFC
LFEF
LIPL
LIPS
LIT
LOPER
LPPREC
LSGA
LSPL
LT
LTRIA
RDLPFC
RFEF
RIPL
RIPS
RIT
ROPER
RPPREC
RSGA
RSPL
RT
RTRIA
SMA
ROI
10
20
30
40
50
Komponentti
0.2
0.15
0.1
0.05

Komponentti 2



Komponentti 8



Yhteenveto

Tarkasteltiin uutta, mielenkiintoista mallia useiden datajoukkojen välisten riippuvuuksien tutkimiseen

Sopii erityisesti myös tilanteisiin, joissa vähän näytteitä suhteessa datan dimensioon

Tehokas toteutus bayesilaisessa viitekehyksessä

Kiinnostava sovelluskohde aivosignaalianalyysissa, jossa potentiaalia saada hyviä tuloksia

Lähteitä

C. M. Bishop. **Pattern recognition and machine learning**. Springer, 2006. ISBN 9780387310732.

S. Virtanen, A. Klami, and S. Kaski. Bayesian CCA via group sparsity. In L. Getoor and T. Scheffer, editors, **Proceedings of the 28th International Conference on Machine Learning (ICML-11)**, ICML '11, pages 457–464, New York, NY, 2011. ACM. ISBN 978-1-4503-0619-5. Implementation in R available at `http://research.ics.tkk.fi/mi/software/gsCCA/`.

S. Virtanen, A. Klami, S. A. Khan, and S. Kaski. Bayesian group factor analysis. In N. Lawrence and M. Girolami, editors, **Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2012)**, JMLR: Workshop and Conference Proceedings, pages 1269–1277, 2012.