

Helsingin yliopisto  
Matematiikan ja tilastotieteen laitos  
Hierarkkiset mallit

Loppuentti 24.10.2006

VASTAA NELJÄÄN KYSYMYKSEEN.

1. Tarkastellaan satunnaistasomallia

$$y_j = \mathbf{x}'_{ij}\beta + \mathbf{z}'_{ij}\eta_j + \epsilon_{ij},$$

jossa  $\mathbf{x}_{ij}$  on kiinteiden selittäjien vektori havainnolle/klusterille  $j, j = 1, \dots, K$ ,  $\beta$  niitä vastaava regressiokerroinvektori ja  $\mathbf{z}_{ij}$  latentteja muuttujia  $\eta_j$  vastaava kiinteä design-muuttujien vektori. Oletetaan, että  $\eta_j \sim N(0, \psi)$ ,  $\epsilon_{ij} \sim N(0, \theta)$ .

a) Kirjoita satunnaistasomallin (random intercept model) hierarkkinen esitys ja siihen liittyvät oletukset havaintojen ja latenttien muuttujien keskinäisestä ja välisestä riippuvuudesta.

b) Mitkä ovat mallin redusoidun muodon parametrit ja mikä on niiden tulkinta?

2. Lineaarinen sekamalli on yleistys kohdan 1 mallista ja muotoa

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\eta + \epsilon, \quad \epsilon \sim N(0, \Theta), \quad \eta \sim N(0, \Psi).$$

a) Esitä latenttien muuttujien posterioriodotusarvo  $E(\eta|\mathbf{y})$ . Miten tulkitset tulosta?

b) Satunnaistasomallissa (vrt. tehtävä 1) design-vektorin  $\mathbf{z}_{ij}$  alkio saavat vain arvoja 0 ja 1. Mikä on tässä tapauksessa latentin muuttujan  $\eta_j$  odotusarvo?

Opastus: käytä multinormaalijakauman ehdollisia jakaumia koskevia tuloksia: Olk. yleisesti  $(X, Z) \sim N(\mu, \Sigma)$ . Tällöin  $X|Z \sim N(\mu_{X|Z}, \Sigma_{X|Z})$ , jossa  $\mu_{X|Z} = \mu_x + \Sigma_{XZ}\Sigma_{ZZ}^{-1}(Z - \mu_z)$  ja  $\Sigma_{X|Z} = \Sigma_{XX} - \Sigma_{XZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX}$ .

3. Johda moniulotteisen faktorimallin  $\mathbf{y} = \Lambda\eta + \epsilon$  mittausten  $\mathbf{y}$  kovarianssimatriisi  $cov(\mathbf{y})$ , kun mallin oletukset ovat  $\eta \sim N(0, \Psi)$ ,  $\epsilon \sim N(0, \Theta)$ ,  $cov(\eta, \epsilon) = 0$ .

4. Latenttien rakenteiden merkitys ja käyttö mallintamisessa

5. Allaolevassa taulukossa on luennolla läpikäydyn aineiston 'Attitudes to abortion' tulostus parametriestimaattien ja suluissa niiden keskivirheet. Tutkimuksessa kysyttiin vastaajan mielipidettä abortin hyväksyttävyydestä 7 kysymyksen kuvaamissa tilanteissa: nainen päättää, pari päättää, aviottomuus, taloudelliset syyt, lapsen sairaus, äidin riski, raiskaus.

a) Mistä tilastollisista malleista on kyse? Kirjoita niiden analyttinen muoto ja oletukset. Mitkä ovat näiden mallien parametrien tulkinnat ja miten mallit eroavat toisistaan?

b) Tulkitse aborttimallin tulokset ja vertaile niiden tulkintaa.

	Model 1	Model 2
Intercepts		
1 [Woman]	-0.54 (0.15)	-0.41 (0.14)
2 [Couple]	0.20 (0.15)	0.48 (0.25)
3 [Marriage]	-0.14 (0.15)	-0.04 (0.19)
4 [Financial]	0.27 (0.15)	0.36 (0.16)
5 [Defect]	3.71 (0.21)	3.32 (0.31)
6 [Risk]	4.77 (0.26)	4.72 (0.53)
7 [Rape]	4.45 (0.24)	3.38 (0.29)
Factor loadings		
1 [Woman]	1 -	1 -
2 [Couple]	1 -	1.66 (0.26)
3 [Marriage]	1 -	1.42 (0.21)
4 [Financial]	1 -	1.15 (0.17)
5 [Defect]	1 -	0.91 (0.17)
6 [Risk]	1 -	1.08 (0.22)
7 [Rape]	1 -	0.64 (0.14)
Structural model	6.43 (0.59)	4.57 (0.92)
Log lik.	-1821.5	-1805.9

Helsingin yliopisto  
Matematiikan ja tilastotieteen laitos  
Hierarkkiset mallit  
Uusintatentti 20.12.2006

### VASTAA KOLMEEN KYSYMYKSEEN

1. Tarkastellaan yleistä lineaarista satunnaistermimallia

$$y_j = \mathbf{x}'_{ij}\beta + \mathbf{z}'_{ij}\eta_j + \epsilon_{ij},$$

jossa  $\mathbf{x}_{ij}$  on kiinteiden selittäjien vektori havainnolle/klusterille  $j, j = 1, \dots, K$ ,  $\beta$  niitä vastaava regressiokerroinvektori ja  $\mathbf{z}_{ij}$  latentteja muuttujia  $\eta_j$  vastaava kiinteä design-muuttujien vektori. Oletetaan, että  $\eta_j \sim N(0, \psi)$ ,  $\epsilon_{ij} \sim N(0, \theta)$ .

Kirjoita satunnaiskerroinmallin (random coefficient model) hierarkkinen esitys ja siihen liittyvät oletukset havaintojen ja latenttien muuttujien keskinäisestä ja välistä riippuvuudesta. Mikä on hierarkkisen mallin parametrien tulkinta?

2. Tarkastellaan latentin muuttujan  $\eta_j$  faktorimallia

$$y_{ij} = \beta_i + \lambda_i \eta_j, \quad i = 1, \dots, n, \quad j = 1, \dots, k, \quad \eta_j \sim N(\gamma, \psi), \quad \epsilon_{ij} \sim N(0, \theta_{ii}).$$

Tehdään lineaarinen muunnos  $f_j = a\eta_j + c$ . Kirjoita muunnetun mallin redusoitu muoto ja faktorin  $f_j$  jakauma. Osoita, että malli ei ole identifioituva ilman rajoitusehtoja. Millä faktorilla  $\eta_j$  koskevilla vaihtoehtoisilla rajoitusehdoilla malli on identifioituva?

3. Oletetaan, että tehtävässä 1 vektorin  $\mathbf{z}_{ij}$  alkioit saavat arvoja 0 ja 1, jolloin satunnaiskerroin  $\beta_j$  kuvaa aineiston ryhmittelyä (ts. kyseessä satunnaistasomalli).

a) Mitä tarkoittaa tällaisessa mallissa lokaalinen riippumattomuus (local independence)? Esitä se havaintojen ehdollisen odotusarvon avulla.

b) Mitä tarkoittaa samassa mallissa sisäkorrelaatio (intraclass correlation)? Esitä se havaintojen ehdollisen odotusarvon avulla.

c) Mikä on sisäkorrelaatio latenttien vasteiden mallissa, jossa  $y_j^* \sim N(0, 1)$ , mutta havaitut vasteet  $y$  saavat arvoja 0 ja 1.

4. a) Latenttien luokkien mallien rakenne ja päättely.

b) Sovellus latenttien luokkien mallista: Oletetaan, että havainnot  $\mathbf{y}$  ovat potilaiden testitulosten mittauksia. Oletetaan lisäksi tietyn sairauden  $c$  todennäköisyyden (eli prevalenssin) väestössä olevan 0.1. Kun testin sensitivisyys taudin suhteen on  $P(y_j = 1 | \eta = c) = 0.9$  ja spesifisyys  $P(y_j = 1 | \eta \neq c) = 0.8$  (kaikille  $j$ ) niin mikä on todennäköisyys, että mielivaltainen potilas, jolla havaitaan positiivinen testitulos, todellakin sairastaa kyseistä tautia  $c$ ?