



# Proc HPMIXED – uusia syvällisiä mahdollisuuksia isojen aineistojen analyyyseihin

Tero Vahlberg  
Biostatistiikka  
Turun yliopisto

# Lineaariset sekamallit

Malli voidaan esittää muodossa

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \gamma_1 z_{i1} + \dots + \gamma_s z_{is} + \varepsilon_i$$

tai vastaavasti matriisimuodossa

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

missä  $\mathbf{y}$  on numeerisen vastemuuttujan havaintovektori

$\mathbf{X}$  on kiinteiden tekijöiden suunnittelumatriisi

$\boldsymbol{\beta}$  on kiinteiden tekijöiden vektori

$\mathbf{Z}$  on satunnaistekijöiden suunnittelumatriisi

$\boldsymbol{\gamma}$  on satunnaistekijöiden vektori

$\boldsymbol{\varepsilon}$  on satunnaisvirheiden vektori

# Mallin oletuksia

- Ehdollinen vastemuuttujan jakauma (ehdollistettuna satunnaistekijöihin) on multinormaalijakauma
- Malli on lineaarinen kiinteiden ja satunnaisten tekijöiden suhteen
- Mallissa satunnaisvirheet ja satunnaistekijät oletetaan normaalisti jakautuneiksi odotusarvonaan  $\mathbf{0}$  keskiarvovektori ja kovarianssimatriiseinaan  $\mathbf{R}$  ja  $\mathbf{G}$
- Lisäksi oletetaan, että satunnaisvirheet ja satunnaistekijät ovat riippumattomia

# Tutkimusasetelmia

- Kokeelliset aineistot
  - Satunnaiset tekijät
- Monikeskusaineistot
- Kaltaistetut aineistot
- Pitkittäisaineistot
  - Toistomittaukset
  - Seuranta-aineistot
  - Paneelaineistot
- Hierarkiset aineistot
  - Monitasomallit

# Proc HPMIXED

- HPMIXED-proseduuri on experimental-muotoinen SAS 9.2:ssa
- HPMIXED-proseduuri on suunniteltu sovittamaan lineaarisia sekamalleja erityisesti seuraavissa tilanteissa
  - Mallin kiinteissä ja/tai satunnaistekijöissä on paljon luokkia
  - Sovittamaan hierarkisia malleja (monitasomalleja), joissa on paljon luokkia kullakin hierarkiatasolla
  - Datassa on paljon havaintoja
- HPMIXED-proseduuri käyttää tähän tarkoitukseen kehitettyjä tehokkaita tekniikoita (high-performance) matriisilaskennassa ja laskenta-algoritmien optimoinnissa

- SAS:issa voidaan sovittaa lineaarisia sekamalleja MIXED, GLIMMIX ja NLMIXED -proseduureilla
  - Kattavat hyvin laajan joukon erilaisia lineaaria sekamalleja, joten estimointitekniikat eivät ole välttämättä optimaalisia suurille datoilte
- HPMIXED -proseduurilla voidaan estimoida vain osa kyseisten proseduurien lineaarista sekamalleista
  - Tämä mahdollistaa tehokkaampien estimointitekniikoiden käytön, mikä pienentää koneen muistikapasiteetin käyttöä ja parantaa laskentanopeutta

# Lineaarinen sekamalliyhtälö

$$\mathbf{C} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \sigma^2\mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}$$

- Lineaaristen sekamallien ratkaisemisessa eli parametrien estimoinnissa **C**-matriisi on keskeisessä roolissa
- **C**-matriisissa on usein paljon 0-alkioita
- HPMIXED-proseduurin estimoinnin tehokkuus perustuu harvojen matriisien tallentamiseen ja matriisilaskentaan kehitettyyn tekniikkaan ja ns. keskiarvoinformaatiomatriisiin käyttöön

# Harva matriisi (Sparse matrix)

- Yleensä matriisilaskenta tehdään koko **C**-matriisille, jolloin matriisilaskenta vaatii paljon muistia ja laskenta-aikaa suurissa aineistoissa
- Jos **C**-matriisissa on paljon 0-alkioita, niin matriisilaskenta saadaan tehokkaammaksi käyttämällä harvojen matriisien tekniikkaa
- Ideana on tallentaa koko **C**-matriisista 0-alkioista poikkeavat alkiot ja niiden sijainti (rivi ja sarake)
- Harvoja matriiseja muodostuu yleensä, kun kiinteissä tekijöissä tai satunnaistekijöissä on paljon luokkia

- Ratkaistaessa lineaarisia sekamalliyhtälöjä perinteisellä tavalla symmetrisen **C**-matriisin ( $N \times N$ ) koon vaikutus tallennukseen (=tarvittavan muistin määrään) on suhteellinen verrattuna lukuun  $N^2$  ja matriisilaskennassa tarvittavaan laskentaoperaatioiden määrään suhteessa  $N^3$
- Harvojen matriisien tekniikkaa käytettäessä **C**-matriisin koon vaikutus tallennukseen on suhteellinen verrattuna lukuun  $n$  (=nollasta poikkeavien alkioiden lukumäärä matriisissa) ja matriisilaskennassa tarvittavaan laskentaoperaatioiden määrään suhteessa  $N \times n$
- Jos  $n$  on paljon pienempi kuin  $N^2$ , niin silloin harvojen matriisien tekniikkaa kannattaa soveltaa

- Esimerkki harvojen matriisien tekniikasta.  
Symmetrinen 5×5 **C**-matriisi

$$\mathbf{C} = \begin{bmatrix} 8 & 0 & 0 & 2 & 0 \\ 0 & 4 & 3 & 0 & 0 \\ 0 & 3 & 5 & 0 & 0 \\ 2 & 0 & 0 & 7 & 0 \\ 0 & 0 & 0 & 0 & 9 \end{bmatrix}$$

- Yläkolmiomatriisissa on 15 alkioita, joista kahdeksan on 0-alkioita. Nollasta poikkeavista alkioista tallennetaan arvot ja sijainnit:

i	1	1	2	2	3	4	5
j	1	4	2	3	3	4	5
C <sub>ij</sub>	8	2	4	3	5	7	9

# Keskiarvoinformaatiomatriisi (Average Information Matrix)

- REML-menetelmää käytetään HPMIXED-proseduurissa kovarianssiparametrien estimoimiseen
- Newton-Raphson ja Fisher scoring ovat yleensä vaihtoehtoisia iteratiivisia menetelmiä uskottavuusfunktioiden maksimoimiseen
- HPMIXED-proseduuri käyttää informaatiomatriisin laskennassa molempien menetelmien yhdistelmää
- Ideana on, että AI-algoritmi korvaa Hessian-matriisiin havaitun (Newton-Raphson) ja odotetun (Fisher scoring) informaatiomatriisin keskiarvolla, mikä tekee laskennasta yksinkertaisempaa ja iteroinnista tehokkaampaa suurissa aineistoissa

# Proc HPMIXED: syntaksi

```
PROC HPMIXED ;  
  BY variables ;  
  CLASS variables ;  
  ID variables ;  
  MODEL dependent = fixed-effects ;  
  RANDOM random-effects ;  
  PARMS ;  
  TEST fixed-effects;  
  CONTRAST 'label' contrast-specification ;  
  ESTIMATE 'label' contrast-specification ;  
  LSMEANS fixed-effects ;  
  NLOPTIONS ;  
  OUTPUT ;  
  WEIGHT variable ;
```

- RANDOM-lauseella määritellään satunnaistekijät
  - Varianssistrukturiksi voidaan valita VC (variance components) tai CHOL (Unstructured Choleskyn parametrisoinnilla)
  - Mallissa voi olla useampia RANDOM-lauseita
- PARMS-lauseella voidaan antaa kovarianssiparametreille alkuarvot tai useita vaihtoehtoja alkuarvoiksi
- NLOPTIONS-lause sisältää optimoinnin onnistumiseen liittyviä määrittämiä
- TEST-lauseella saadaan laskettua kiinteiden tekijöiden tyyppiä 3 testit
- CONTRAST- ja ESTIMATE-lauseilla voidaan testata erilaisia hypoteeseja

# HPMIXED ja MIXED-proseduurien vertailu

- HPMIXED:ssä ei ole REPEATED-lausetta
- MIXED:ssä on tarjolla enemmän kovarianssistruktuureja RANDOM-lauseessa
- Kiinteiden tekijöiden tyyppin 3 testit ovat MIXED:ssä oletuksena, HPMIXED:ssä ne saa pyydettäessä
- ODS Statistical Graphics:ia ei ole saatavilla HPMIXED:ssä
- MIXED:ssä tarjolla erilaisia vaihtoehtoja kiinteiden tekijöiden vapausastekorjauksiin, HPMIXED käyttää Residual-menetelmää
- MIXED:ssä valittavana ML- ja REML-estimointi sekä momenttimenetelmä, HPMIXED:ssä REML-estimointi
- MIXED:ssä PRIOR-lause (Bayesiläinen analyysi)

# Esimerkki sekamallin soveltamisesta lehmien arvioituun jalostusarvoon (EBV)

- Simuloitu aineisto (60 000 havaintoa), joka sisältää 15 karjatilaa ja jokaista tilaa kohti noin 100 lehmää 5:stä eri lajista
  - Yhtä lehmää kohti noin 40 mittausta maidontuotannosta
  - Simuloinnissa satunnaistekijöiden varianssina oli 4 ja mallin virhevarienssina 8
  - Vastemuuttujana on maidontuotanto (paunoina) päivää kohti, joka simuloitiin kaavalla:

$$\text{Yield} = 1 + \text{Species} + \text{Farm} + \text{BV}\{\text{Animal}\} + \varepsilon$$

- Mallissa laji ja karjatila ovat kiinteitä tekijöitä ja lehmä satunnaistekijä
- Esimerkissä siis oletetaan, että 33.3% ( $ICC=4/(4+8)$ ) maidontuotannon vaihtelusta selittyy satunnaistekijällä
- Tallennetaan mallin ennusteet kunkin lehmän satunnaistekijän kertoimelle (=EBV).

ods listing close;

```
proc hpmixed data=Sim;
```

```
class Species Farm Animal;
```

```
model Yield = Species Farm*Species;
```

```
random Animal / cl;
```

```
ods output SolutionR=EBV;
```

```
run;
```

ods listing;

- Arvioidut jalostusarvot (Estimate) ja niiden luottamusvälit: TOP 10 lehmät

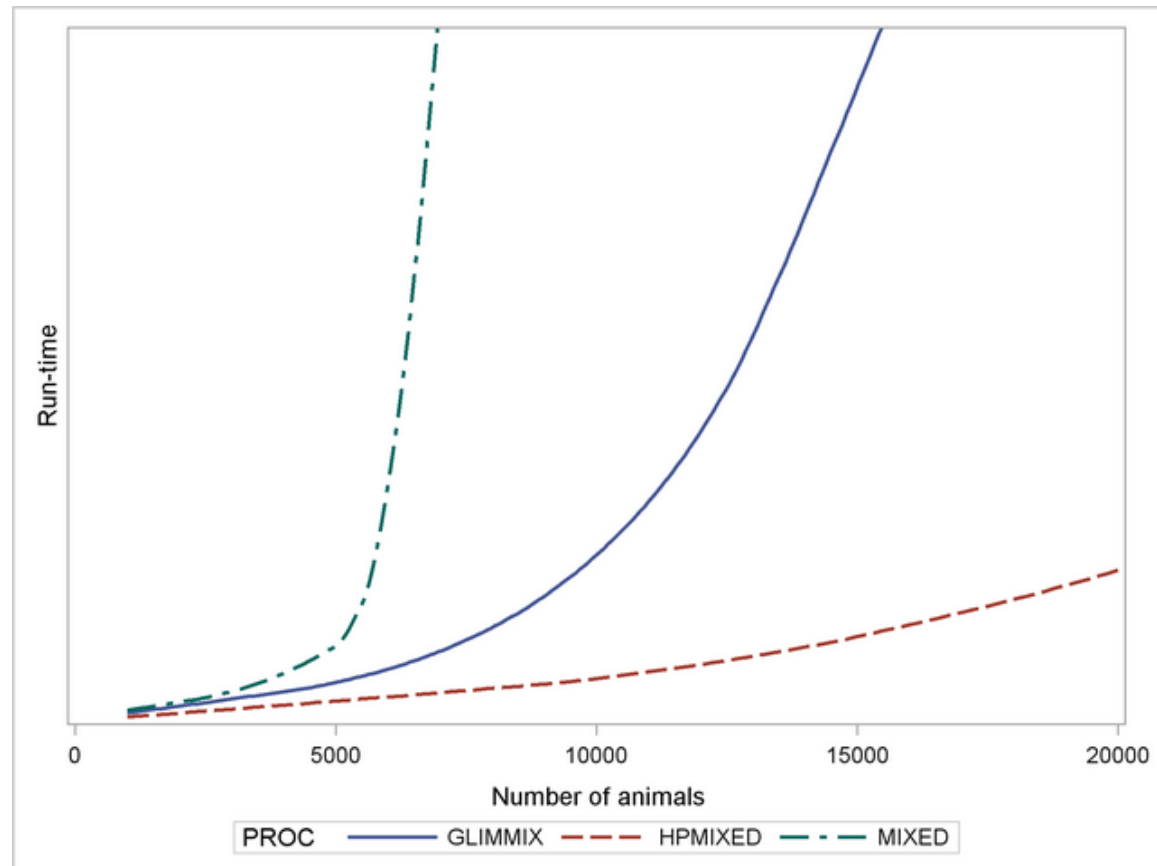
Animal	Estimate	Pred	Lower	Upper
1294	5.9703	0.6317	4.7321	7.2085
1219	5.0081	0.6396	3.7544	6.2618
1054	4.9452	0.5874	3.7939	6.0966
758	4.9340	0.6196	3.7195	6.1485
986	4.9329	0.5767	3.8025	6.0633
1150	4.7444	0.5806	3.6064	5.8824
962	4.6651	0.5794	3.5294	5.8008
225	4.5294	0.6137	3.3266	5.7322
1252	4.5012	0.5686	3.3868	5.6157
1033	4.4971	0.6080	3.3054	5.6889

# Proseduurien välistä laskenta-aikojen vertailua

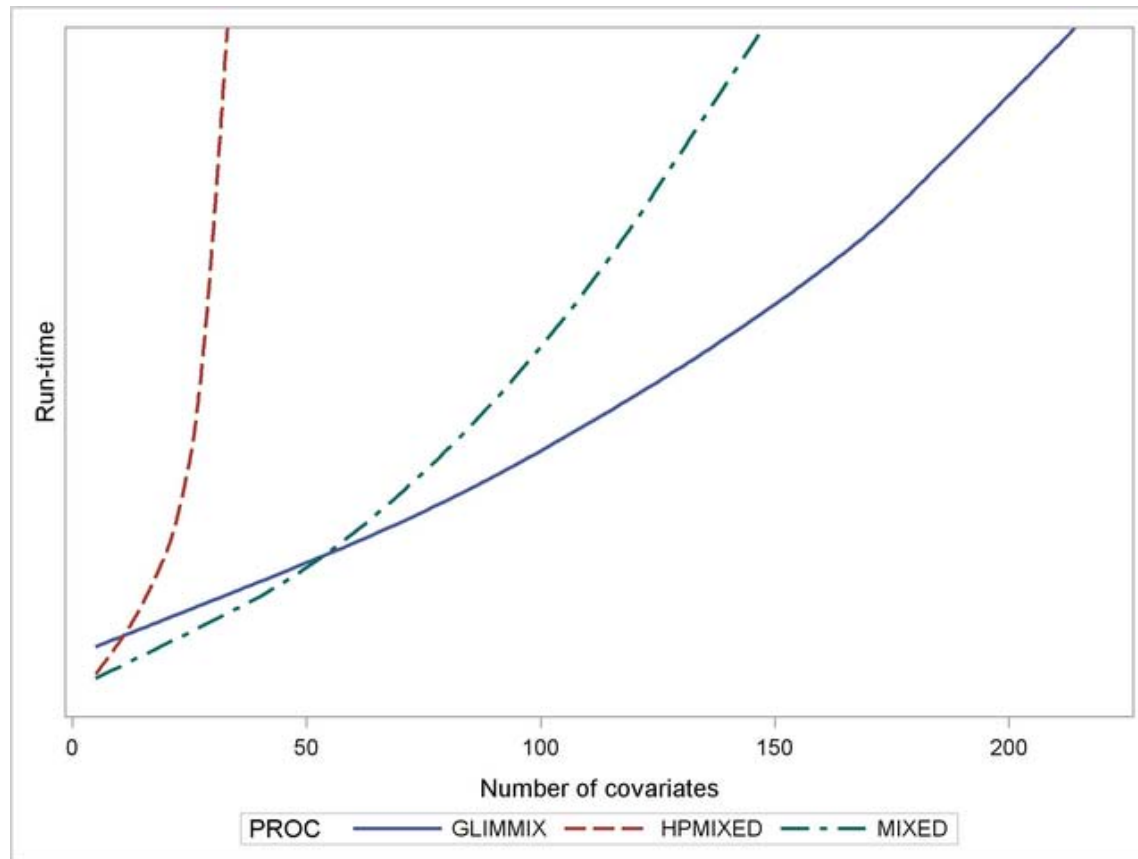
- Esimerkkimallin **X**-matriisissa 81 saraketta ja **Z**-matriisissa 1500 saraketta eli **C**-matriisissa 1581-saraketta ja **C**-matriisin alkioista n. 99% on 0-alkioita
- Kolmen mallin vaatima laskenta-aika eri proseduureilla (koneessa 3.5 GB RAM-muistia)
  - Malli 1 (1500 lehmää), Malli 2 (3000) ja Malli 3 (5000)

CPU aika (min:sec)	HPMIXED	MIXED	GLIMMIX
Malli 1	0:0.85	0:31.56	1:13.03
Malli 2	0:2.71	4:19.34	10:06.25
Malli 3	0:5.37	24:28.65	out of memory

- Kuvaajan taustalla malli, jossa yhden karjatilän lisäys tuo simuloituun aineistoon 500 lehmää lisää
- Mallin vaatima laskenta-aika eri proseduureilla, kun aineiston lehmien määrä kasvaa



- Muuten sama malli kuin edellisessä kuvaajassa, mutta mallissa on lisäksi numeerisia kovariaatteja
- HPMIXED ei ole kovariaattien määrän kasvaessa yhtä tehokas kuin MIXED ja GLIMMIX



# Lähteet

- SAS Help and Documentation: SAS/STAT User's Guide
- Wang T and Tobias T: All the Cows in Canada: Massive Mixed Modeling with the HPMIXED Procedure in SAS® 9.2. Paper 256-2009, SAS Institute Inc., Cary NC, USA.

# Kysymyksiä?

Yhteystiedot:

Tero Vahlberg

Biostatistiikka

Turun yliopisto

email: [tervah@utu.fi](mailto:tervah@utu.fi)

puh: 02-333 8543