Poznámky o niektorých výpočtových aspektoch tradičných testov o pevných a náhodných efektoch v lineárnych zmiešaných modeloch

Viktor WITKOVSKÝ[1]

Ústav merania SAV, Bratislava
witkovsky@savba.sk

18. zimní škola JČMF ROBUST 2014
Jetřichovice, 19. - 24. január 2014

[1] Práca vznikla vďaka podpore grantov APVV-0096-10, SK-AT-0025-12, VEGA 2/0038/12 a VEGA 2/0043/13.

Abstrakt

- Metódy štatistickej inferecie pre testovanie hypotéz a konštrukciu konfidenčných oblastí pre pevné a náhodné efekty v zmiešaných lineárnych modeloch sú obyčajne zaožené na (približných) testoch pomerom vierohodností, alebo na (približných) testoch založených na testovacích štatistikách Waldovho typu.
- Napriek tomu, že tieto metódy sú dobre známe a implementované v algoritmoch známych štatistických balíkov, potreba analyzovať rozsiahle dáta prináša technické problémy ako takéto metódy efektívne implementovať (napr. R: nlme, lme4, SAS: Proc MIXED, Proc HPMIXED, resp. MATLAB fitlme).
- V tomto príspevku stručne popíšeme štandardne používané metódy testovania hypotéz pomocou testovacích štatistík Waldovho typu a metódy aproximácie ich rozdelenia za platnosti nulovej hypotézy (aproximácia pomocou metódy Satterthwaite-Fai-Cornelius a Kenward-Roger) s dôrazom na niektoré výpočtové aspekty implementovania týchto metód, založených na riešení Hendersonových rovníc pre zmiešaný lineárny model (MME).
- Uvedieme zovšeobecnenie metódy Kenwarda-Rogera pre testovanie lineárnych funkcií o pevných efektov aj na testovanie lineárnych funkcií o pevných a náhodných efektoch súčasne.

Proc HPMIXED

The HPMIXED procedure is designed to solve large mixed model problems by using sparse matrix techniques, but with relatively few covariance parameters.

- Parameter estimation (fixed and random effects), inference, and prediction in linear mixed models
- Estimate covariance parameters by restricted maximum likelihood (REML)
- Computation of appropriate standard errors for all specified estimable linear combinations of fixed and random effects, and corresponding t and F tests
- List of features NOT AVAILABLE in the HPMIXED procedure:
  - can model only G-side random effects with variance component structure or an unstructured covariance matrix in a Cholesky parameterization. R-side random effects and direct modeling of their covariance structures are not supported,
  - NO automatic Type III tests of fixed effects. You request tests of fixed effects in the HPMIXED procedure with the TEST statement,
  - NO advanced degree-of-freedom adjustments available by using the DDFM= option,
  - NO maximum likelihood or method-of-moments estimation for the covariance parameters,
  - NO Fisher Information matrix for estimated covariance parameters.

R lmer / lme4

Hi,I am wondering how to conduct Kenward-Roger correction in the linear mixed model using R. Any idea? Thanks a lot, Suyan

Ben Bolker wrote:

- Not really possible, I'm afraid. ... Doug Bates has declined to spend effort implementing K-R because
  1. he's not convinced of the appropriateness of adjusting F-distribution degrees of freedom in this way,
  2. he doesn't think that the K-R algorithm will be feasible for the sorts of large-data problems he's interested in,
  3. he finds the correspondence between K-R's notation and his difficult.

Doug Bates wrote:

- It has been a while since I looked at the Kenward-Roger formulation but my recollection is that it would be difficult to extend the calculations to models with non-nested random effects.
- I am not opposed to the method in principle - it's just that I am not about to have the time to work on it myself in the foreseeable future. If someone else wants to work it out then I say go for it.
  - For models with non-nested random effects it is possible to get an expression for the marginal variance-covariance of the response vector but this is potentially a dense n by n matrix and you really don't want to try working with that for large data sets.

R lmer / lme4

- InstEval Example: University lecture evaluations by students at ETH Zurich
- A data frame with 73421 observations. This is an interesting “medium” sized example of a partially nested mixed effect model.
  - s - a factor with levels 1:2972 denoting individual students.
  - d - a factor with 1128 levels from 1:2160, denoting individual professors or lecturers.
  - service - a binary factor with levels 0 and 1; a lecture is a “service”, if held for a different department than the lecturer’s main one.
  - dept - a factor with 14 levels from 1:15, using a random code for the department of the lecture.
  - y - a numeric vector of ratings of lectures by the students, using the discrete scale 1:5, with meanings of ‘poor’ to ‘very good’.

  Each observation is one student’s rating for a specific lecture (of one lecturer, during one semester in the past).
- The main goal of the survey is to find “the best liked prof”, according to the lectures given. Statistical analysis of such data has been the basis for a (student) jury selecting the final winners.
- fm <- lmer(y $\sim$ dept*service + (1|s) + (1|d), InstEval)

Lineárny zmiešaný model

LMM — zmes pevných a náhodných efektov:

$$y = Xb + Zu + e = \sum_{i=1}^{f} X_i b_i + \sum_{j=1}^{r} Z_j u_j + e$$

pričom

$$X = [X_1 : X_2 : \cdots : X_f], \qquad Z = [Z_1 : Z_2 : \cdots : Z_r]$$
$$b = [b_1; b_2; \ldots; b_f], \qquad u = [u_1; u_2; \ldots; u_r], \qquad e = [e_1; e_2; \ldots; e_t]$$

- b — p-vektor pevných efektov, $p = \sum_{i=1}^{f} p_i$, vektor b_i reprezentuje úrovne faktora $i = 1, \ldots, f$.
- u — q-vektor náhodných efektov, $q = \sum_{i=1}^{r} q_i$, pričom $u_i \sim N(0, G_i(\theta_i^G))$, $u \sim N(0, G(\theta^G))$, kde $G(\theta) = \oplus_{i=1}^{r} G_i(\theta_i^G)$, pričom $\theta^G = [\theta_1^G; \ldots; \theta_r^G]$
- e — n-vektor náhodných chýb, $n = \sum_{i=1}^{t} n_i$, pričom $e_i \sim N(0, R_i(\theta_i^R))$, $e \sim N(0, R(\theta^R))$, kde $R(\theta^R) = \oplus_{i=1}^{t} R_i(\theta_i^R)$, pričom $\theta^R = [\theta_1^R; \ldots; \theta_t^R]$
- $Var(y) = V(\theta) = ZG(\theta)Z' + R(\theta)$, kde $\theta = [\theta^G; \theta^R]$, pričom $n_V = n_G + n_R$.

Lineárny zmiešaný model

- Častou verziou LMM je model s jednoduchou štruktúrou variančných komponentov (VC):
  - $u_i \sim N(0, G_i(\theta))$, kde $G_i(\theta) = \sigma_i^2 I_{q_i}$, $i = 1, \ldots, r$.
  - $e_i \sim N(0, R_i(\theta))$, kde $R_i(\theta) = \sigma^2 I_{n_i}$, $i = 1, \ldots, t$, teda $e \sim N(0, \sigma^2 I_n)$.
- Potom teda $y \sim N(Xb, V(\theta))$, kde $V(\theta) = ZG(\theta)Z' + R(\theta)$,
  - $G(\theta) = \oplus_{i=1}^{r} G_i(\theta) = diag(\sigma_i^2 I_{q_i})$,
  - $R(\theta) = \oplus_{i=1}^{t} R_i(\theta) = diag(\sigma^2 I_{n_i}) = \sigma^2 I_n$,

  kde $\theta = (\sigma_1^2, \ldots, \sigma_r^2, \sigma_{r+1}^2)$, pričom $\sigma_{r+1}^2 = \sigma^2$.

Hendersonové rovnice pre zmiešaný lineárny model

- Henderson odvodil systém rovníc (MME - Mixed Model Equation), ktoré umožňujú odhad pevných a náhodných efektov: najlepší lineárny nevychýlený odhad (BLUE - Best Linear Unbiased Estimator) vektora Xb (alebo nevychýlene odhadnuteľnej lineárnej kombinácie $K'b$) a najlepší lineárny nevychýlený prediktor (BLUP - Best Linear Unbiased Predictor) vektora u (alebo lineárnej kombinácie $w = K'b + L'u$, ak $K'b$ je odhadnuteľná), za predpokladu, že kovariančná štruktúra LMM je známa.
- MME boli odvodené za predpokladov normality rozdelenia, teda $u \sim N(0, G)$, $e \sim N(0, R)$, pričom $Cov(u, e) = 0$, pre známe variančno-koveriančné matice $G = G(\theta)$ a $R = R(\theta)$.
- Združená hustota (pdf) vektora $(y', u')'$ je teda

$$f(y,u) = f(y|u)f(u)$$
$$\frac{1}{(2\pi)^{n/2}|R|^{1/2}} \exp\left\{-\frac{1}{2}(y - Xb - Zu)'R^{-1}(y - Xb - Zu)\right\}$$
$$\times \frac{1}{(2\pi)^{r/2}|G|^{1/2}} \exp\left\{-\frac{1}{2}u'G^{-1}u\right\}.$$

- Tu predpokladáme, že $G > 0$.

Hendersonové rovnice pre zmiešaný lineárny model

- Riešením rovníc pre modus b a u, t.j.

$$\frac{\partial f(y,u)}{\partial b} = 0, \qquad \frac{\partial f(y,u)}{\partial u} = 0$$

dostávame MME rovnice v tvare

$$\begin{pmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{pmatrix} \begin{pmatrix} \tilde{b} \\ \tilde{u} \end{pmatrix} = \begin{pmatrix} X'R^{-1}y \\ Z'R^{-1}y \end{pmatrix}.$$

- Maticu koeficientov systému rovníc MME budeme označovať ako H (Hendersonová matica). Taktiež budeme používať rozklad $H = H_R + H_G$.

- Inverziu matice H budeme označovať ako C, s blokmi C_{11}, $C_{12} = C_{21}$, C_{22}.

Hendersonové rovnice pre zmiešaný lineárny model

Dôležité vlastnosti riešenia MME rovníc:

1. V triede lineárnych nevychýlených prediktorov, BLUP maximalizuje koreláciu medzi u a $\tilde{u}$.
2. $K'\tilde{b}$ je (jednoznačne daný) BLUE vektora $K'b$, ak $K'b$ je odhadnuteľná funkcia.
3. $E(u \mid \tilde{u}) = \tilde{u}$.
4. BLUP $\tilde{u}$ je daný jednoznačne.
5. $K'\tilde{b} + L'\tilde{u}$ je (jednoznačne daný) BLUP funkcie $K'b + L'u$ ak $K'b$ je odhadnuteľná funkcia.
6. $Var\left(K'\tilde{b}\right) = K'C_{11}K = K'\left(X'V^{-1}X\right)^{-}K.$
7. $Var\left(K'\tilde{b} + L'\tilde{u}\right) = K'C_{11}K + L'(G - C_{22})L.$
8. $MSE\left(K'\tilde{b} + L'\tilde{u}\right) = Var\left(\left(K'\tilde{b} + L'\tilde{u}\right) - (K'b + L'u)\right) = (K', L')C(K', L')'.$
9. $Cov\left(K'\tilde{b}, \tilde{u}'\right) = 0.$
10. $Cov\left(K'\tilde{b}, u'\right) = -K'C_{12}.$
11. $Cov\left(K'\tilde{b}, u' - \tilde{u}'\right) = -K'C_{12}.$
12. $Var(\tilde{u}) = Cov(\tilde{u}, u') = G - C_{22}.$
13. $MSE(\tilde{u}) = Var(\tilde{u} - u) = C_{22}.$

Metódy štatistickej inferencie pre pevné a náhodné efekty

- Ak sú variančné-kovariančné komponenty θ neznáme, potom musia byť odhadnuté z pozorovaných dát rozumnou odhadovacou procedúrou - napríklad metódou ML (Maximum Likelihood) alebo REML (Restricted Maximum Likelihood).
- Existuje niekoľko implementácií na odhadovanie variančných-kovariančných komponentov vo všeobecnom LMM.
- Takéto metódy sú implementované v mnohých počítačových balíkoch určených na analýzu dát pomocou LMM, napr. SAS: PROC MIXED, PROC HPMIXED, PROC GLIMMIX, R: lme4, nlme, MATLAB: fitlme.
- *Avšak potreba analyzovať rozsiahle dáta prináša technické problémy ako takéto metódy efektívne implementovať, a predovšetkým ako implementovať metódy pre štatistickú inferenciu v týchto modeloch. SAS: PROC HPMIXED $\rightarrow$?, R: lme4 $\rightarrow$? (lmerTest, pbkrtest, ...), julia: MixedModels $\rightarrow$?.*
- Nech $\hat{\theta}$ je výsledok takejto numerického riešenia (v prípade jednoduchého LMM $\hat{\theta} = (\hat{\sigma}^2_1, \ldots, \hat{\sigma}^2_{r+1})$). Potom *riešenie MME* budeme označovať ako $\hat{b}$, $\hat{u}$. Podobne $\hat{G} = G(\hat{\theta})$, $\hat{R} = R(\hat{\theta})$, $\hat{H} = H(\hat{\theta})$ a $\hat{C} = C(\hat{\theta})$ budú označovať odhadnuté verzie matíc H, G, R, a C.
- Spolu s odhadom vektora variančných-kovariančných komponentov $\hat{\theta}$ je potrebný aj odhad kovariančnej matice Σ takéhoto odhadu - napr. pomocou odhadu inverzie Fisherovej informačnej matice, resp. inverzie Hessiánu. $\hat{\Sigma} = \left(I_{ML}(\hat{\theta})\right)^{-1}$ resp. $\hat{\Sigma} = \left(I_{REML}(\hat{\theta})\right)^{-1}$.

Elementy štatistickej inferencie pre pevné a náhodné efekty

- H_{fac}: Cholesky-factor matice H, pričom $H = H'_{fac}H_{fac}$, a $C_{fac} = H_{fac}^{-1}$, že $C = C_{fac}C'_{fac}$.
- Riešenie MME rovníc

$$[\tilde{b}; \tilde{u}] = Cr, \text{ kde } r = [X'R^{-1}y; Z'R^{-1}y].$$

- REML log-likelihood funkcia:

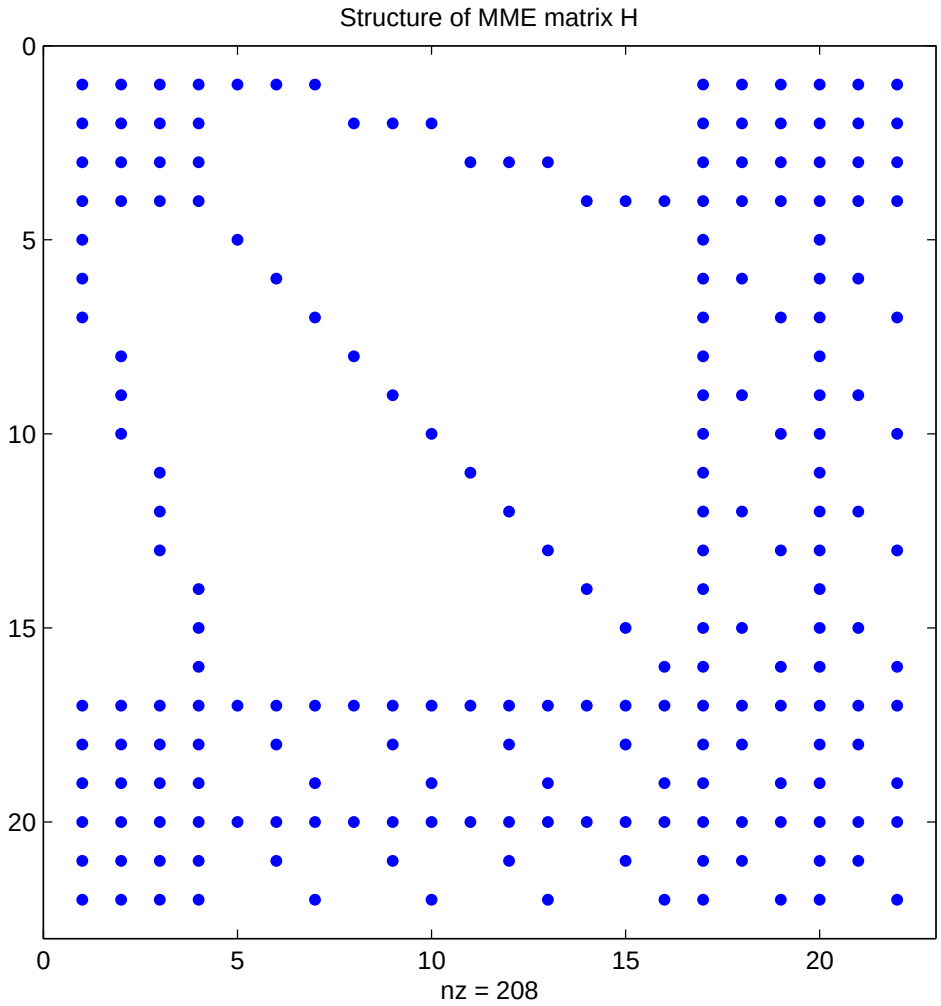
$$\begin{aligned} \log(L_R) &= -\frac{1}{2}\left[\text{const} + \log|V| + log|X'VX| + y'Py\right] \\ &= -\frac{1}{2}\left[\text{const} + \log|R| + \log|G| + \log|H| + y'Py\right], \end{aligned}$$

kde $P = V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}$, pričom

  - $P = PVP = P(ZGZ' + R)P$,
  - $Py = V^{-1}(y - X\tilde{b}) = R^{-1}(y - X\tilde{b} - Z\tilde{u}) = R^{-1}\tilde{e}$,
  - $y'Py = y'R^{-1}(y - X\tilde{b} - Z\tilde{u}) = (y - X\tilde{b} - Z\tilde{u})'R^{-1}(y - X\tilde{b} - Z\tilde{u})$,
  - $\tilde{u} = GZ'V^{-1}(y - X\tilde{b})$.

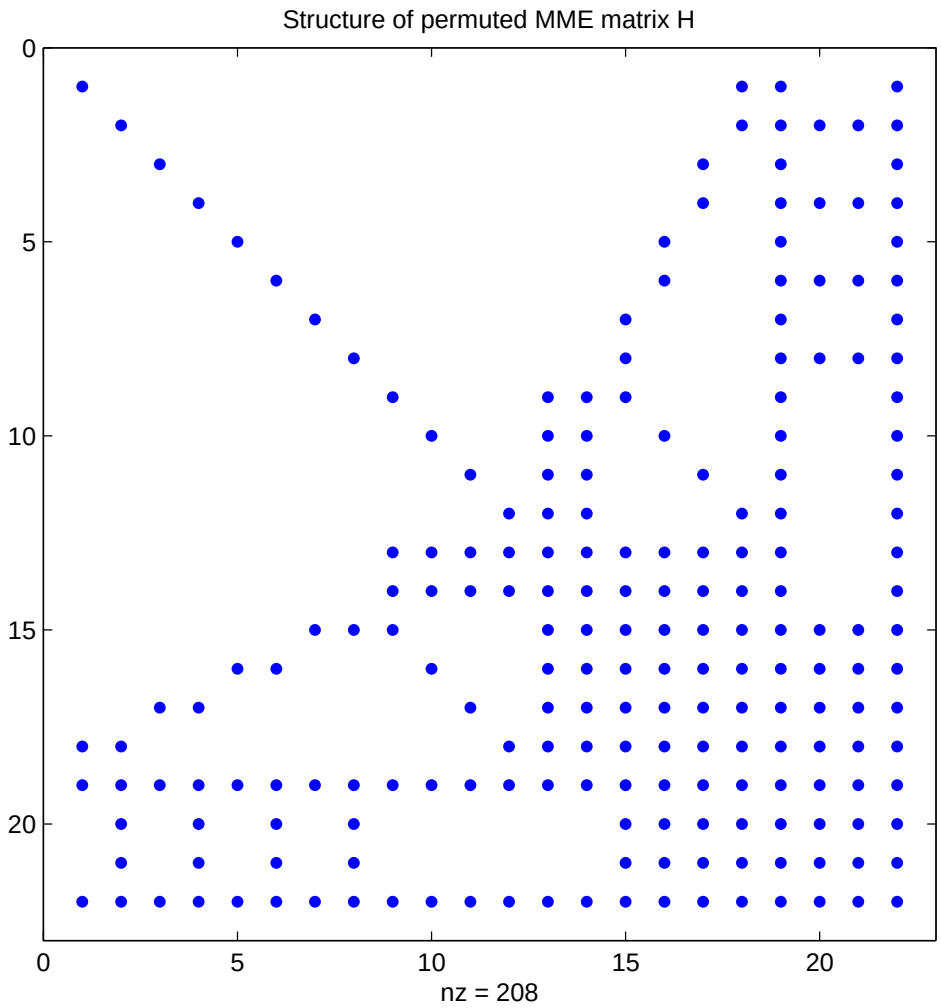
- $\frac{\partial y'Py}{\partial \theta_i^G} = \frac{\partial y'P(ZGZ'+R)Py}{\partial \theta_i^G} = \tilde{u}'\left(G^{-1}\frac{\partial G}{\partial \theta_i^G}G^{-1}\right)\tilde{u} = \tilde{u}_i'\left(G_i^{-1}\frac{\partial G}{\partial \theta_i^G}G_i^{-1}\right)\tilde{u}_i$,
- $\frac{\partial y'Py}{\partial \theta_i^R} = \frac{\partial y'P(ZGZ'+R)Py}{\partial \theta_i^R} = y'P\left(R^{-1}\frac{\partial R}{\partial \theta_i^R}R^{-1}\right)Py = \tilde{e}'\left(R_i^{-1}\frac{\partial R}{\partial \theta_i^R}R_i^{-1}\right)\tilde{e}$,
- $I_{REML} = \frac{\partial^2(\log|R|+\log|G|+\log|H|)}{\partial\theta\partial\theta'}$.

Príklady



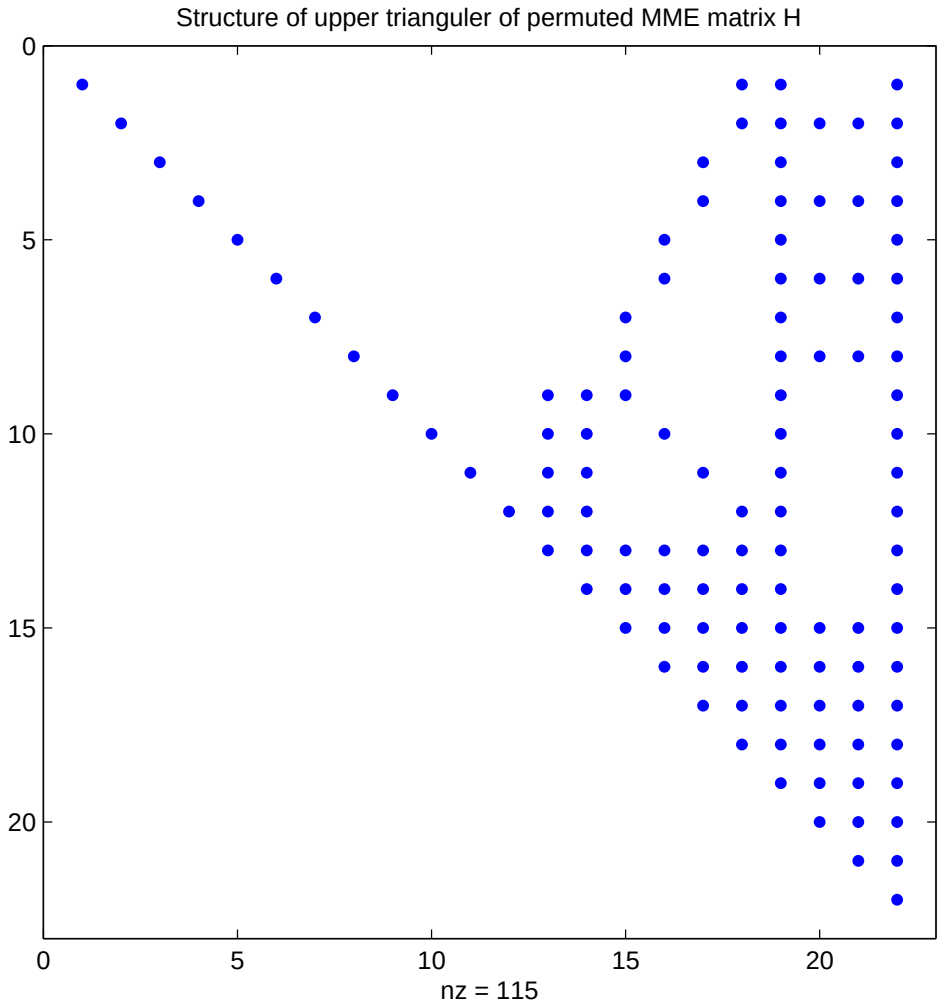
Obr. : Ilustračný príklad (Split Plot Data): Štruktúra matice MME koeficientov H

Príklady



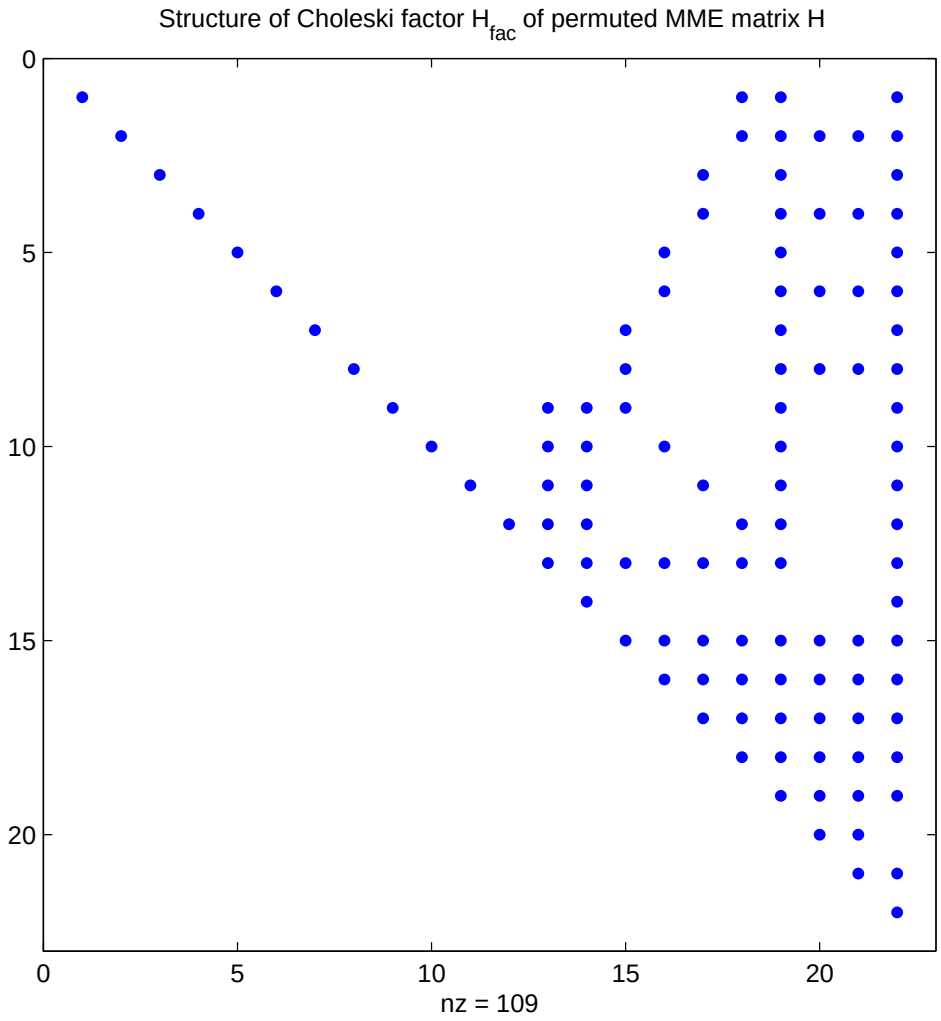
Obr. : Ilustračný príklad (Split Plot Data): Štruktúra permutovanej matice MME koeficientov *H*

Príklady



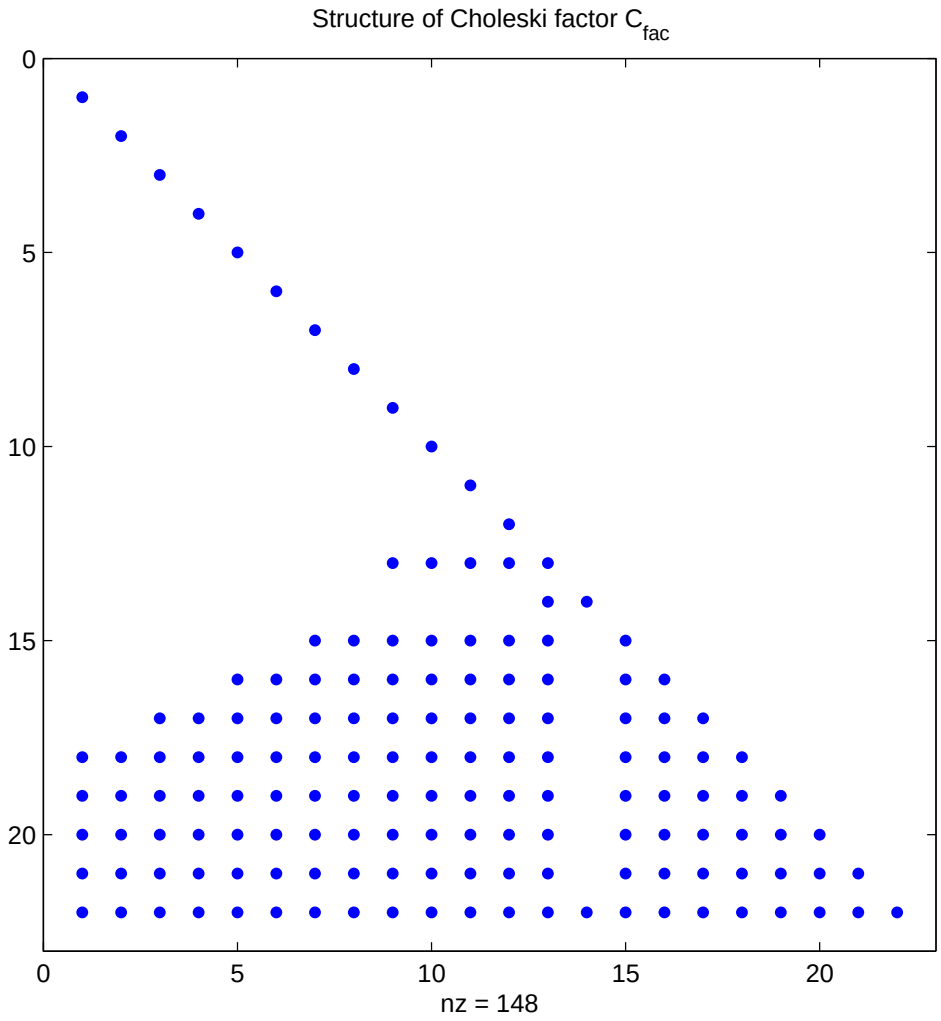
Obr. : Ilustračný príklad (Split Plot Data): Štruktúra hornej trojuholníkovej časti permutovanej matice MME koeficientov *H*

Príklady



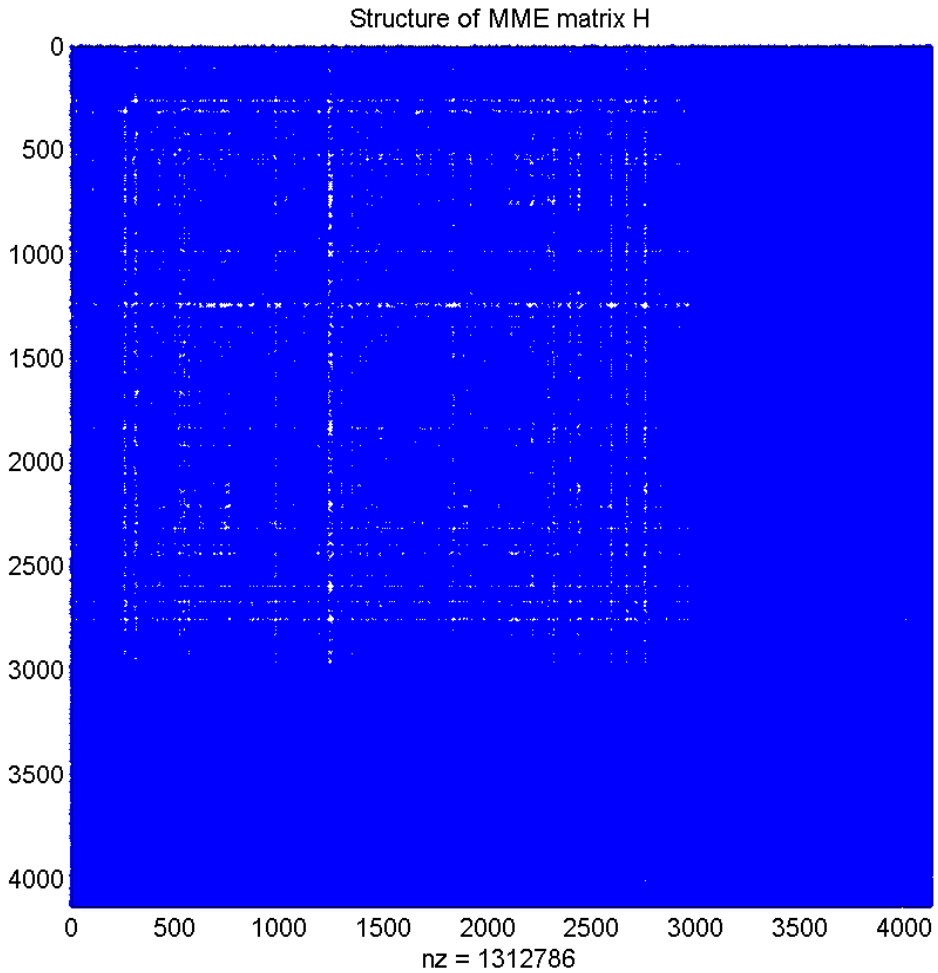
Obr. : Ilustračný príklad (Split Plot Data): Štruktúra choleského faktora H_{fac} permutovanej matice MME koeficientov H

Príklady



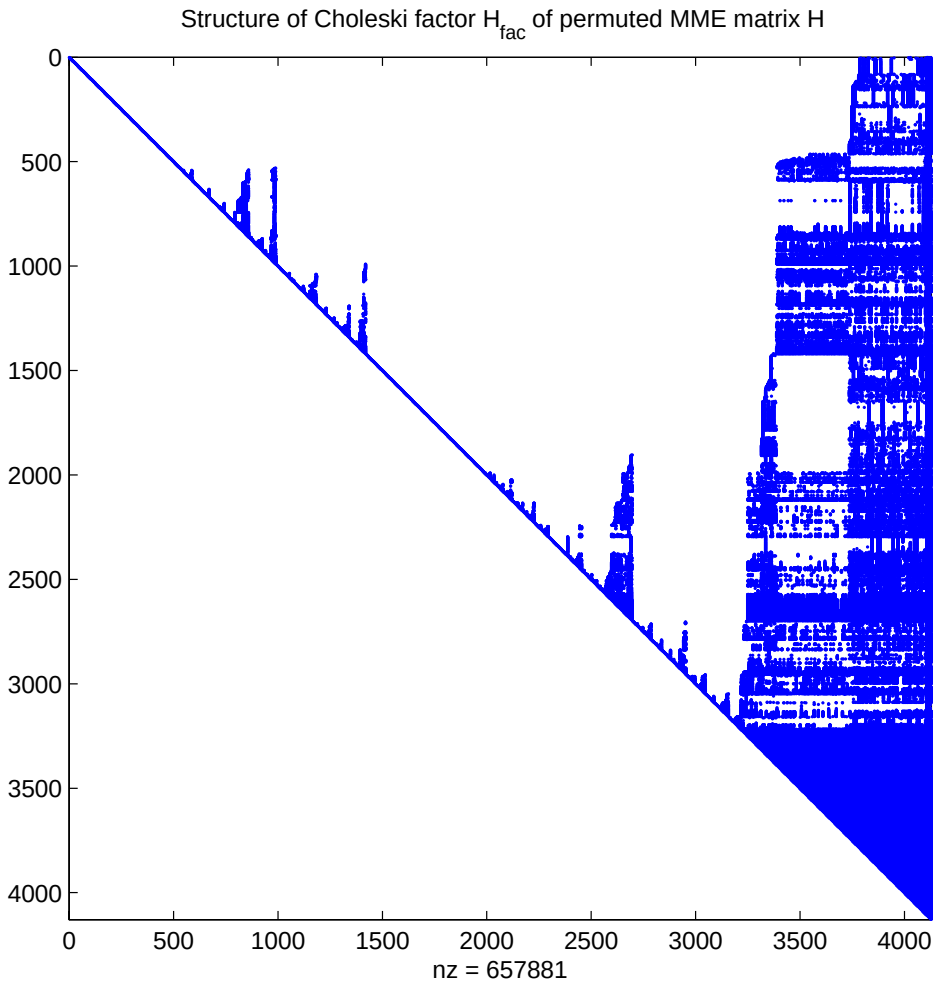
Obr. : Ilustračný príklad (Split Plot Data): Štruktúra inverzie choleského faktora $inv(H_{fac})$

Príklady



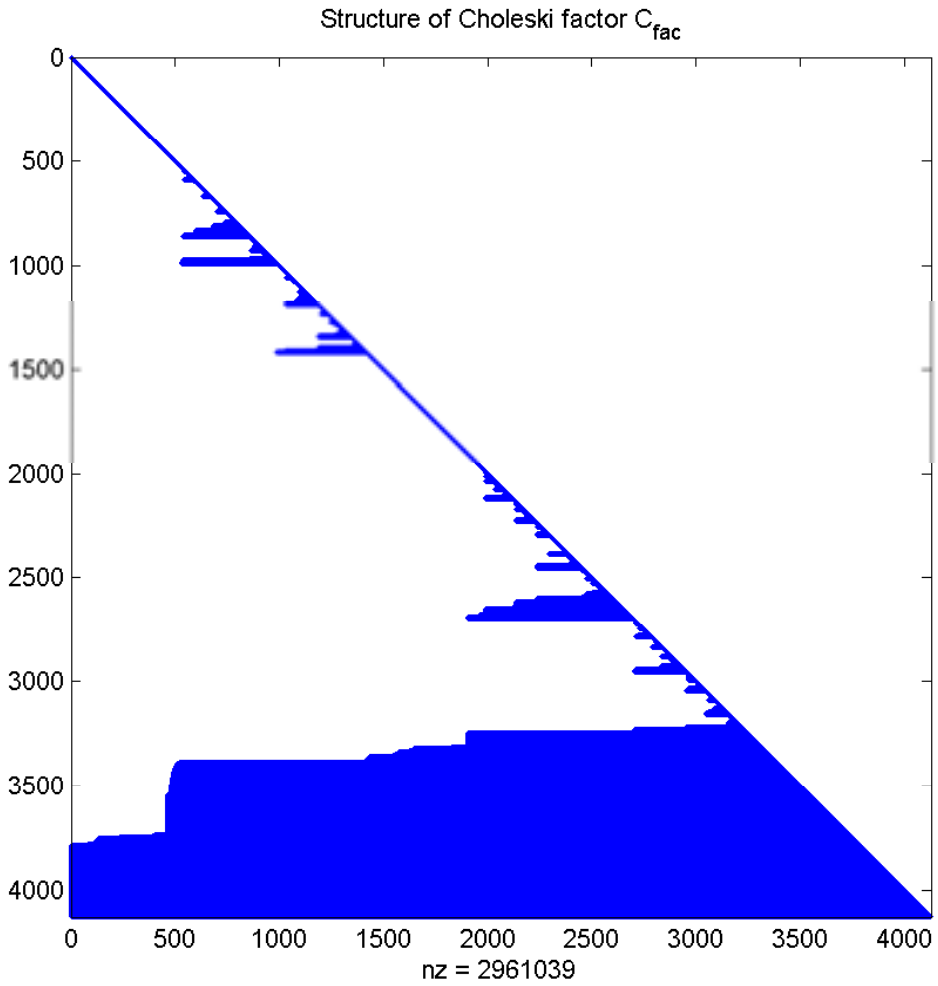
Obr. : Ilustračný príklad: Exaktné konfidenčné oblasti pre variančné komponenety normálneho lineárneho modelu s dvomi variančnými komponentami

Príklady



Obr. : Ilustračný príklad: Exaktné konfidenčné oblasti pre variančné komponenety normálneho lineárneho modelu s dvomi variančnými komponentami

Príklady



Obr. : Ilustračný príklad: Exaktné konfidenčné oblasti pre variančné komponenety normálneho lineárneho modelu s dvomi variančnými komponentami

Štandardné metódy štatistickej inferencie pre pevné a náhodné efekty

- V ďalšom budeme uvažovať metódy štatistickej inferencie o q lineárnych funkciách pevných efektov b a náhodných efektov u, teda o funkcii $w = \Lambda' \left(b', u'\right)' = K'b + L'u$, kde Λ je $((p + r) \times q)$-rozmerná matica plnej hodnosti, pričom $K'b$ je odhadnuteľná (t.j. $K = X'A$ pre nejakú maticu A).

- Nech $\tilde{b}$ a $\tilde{u}$ sú riešenia MME rovníc, teda $\tilde{w} = \Lambda' \left(\tilde{b}', \tilde{u}'\right)' = K'\tilde{b} + L'\tilde{u}$ je BLUP funkcie $w = K'b + L'u$.

- Na základe známych vlastností riešenia $\tilde{b}$ a $\tilde{u}$ dostávame

$$MSE\left(\tilde{w}\right) = E\left(\left(\tilde{w} - w\right)\left(\tilde{w} - w\right)'\right) = Var\left(\tilde{w} - w\right) = \Lambda' C \Lambda = M_{\tilde{w}}.$$

- MSE matica $M_{\tilde{w}}$ náhodného vektora $\tilde{w}$ funkčne závisí od variančných-kovariančných komponentov θ, špeciálne $\theta = \left(\sigma_1^2, \ldots, \sigma_r^2, \sigma_{r+1}^2\right)'$.

Štandardné metódy štatistickej inferencie pre pevné a náhodné efekty

- Ak by bol vektor variančných-kovariančných komponentov θ známy, potom (na základe prepokladov modelu) sa pre štatistickú inferenciu štandarne využíva pivot - štatistika Waldovho typu.

- Waldov pivot možno využiť na štatistickú inferenciu o funkcii $\tilde{w} = \Lambda' \left(\tilde{b}', \tilde{u}'\right)' = K'\tilde{b} + L'\tilde{u}$:

  Teda testovanie nulovej hypotézy $H_0 : K'b = K'b_0$, kde b_0 je daná testovaná hodnota parametra pevných efektov, resp. konštrukciu konfidenčných/predikčných oblastí pre $K'b$, resp. w.

- *Exaktné (nulové, za platnosti pre $H_0 : w = w_0$) rozdelenie* je triviálne:

$$Q = (\tilde{w} - w_0)' \left(\Lambda' C \Lambda\right)^{-1} (\tilde{w} - w_0) \sim \chi^2_\ell,$$

  χ^2_ℓ tu označuje chi-kvadrát rozdelenie s $\ell = rank(\Lambda')$ stupňami voľnosti.

Štandardné metódy štatistickej inferencie pre pevné a náhodné efekty

- Predpokladajme, že variančné-kovariančné komponenty θ sú neznáme, ale máme k dispozícií odhad $\hat{\theta}$ spolu s odhadom $\hat{C}$ *[Problém č. 1 efektívne odhadovanie $\hat{\theta}$]*

- Na inferenciu o funkcii pevných a náhodných efektov $w = \Lambda' (b', u')'$ založenú na empirickom BLUPe (EBLUP), t.j. $\hat{w} = \Lambda' \left(\hat{b}', \hat{u}'\right)'$, sa typicky uvažuje štatistika

$$F = \frac{1}{\ell} (\hat{w} - w_0)' \left(\Lambda' \hat{C} \Lambda\right)^{-1} (\hat{w} - w_0),$$

  kde $\ell = rank(\Lambda')$.

- Špeciálnym prípadom je situácia, keď w je skalárna (jedno-dimenzionálna) veličina $w = \lambda' (b', u')' = k'b + l'u$, potom uvažujeme štatistiku

$$t = \frac{\hat{w} - w_0}{\sqrt{\lambda' \hat{C} \lambda}},$$

  kde $\hat{w}$ je EBLUP náhodnej premennej w.

Štandardné metódy štatistickej inferencie pre pevné a náhodné efekty

- Nulové rozdelenie týchto 't-štatistík' je netriviálne a najčastejšie sa aproximuje pomocou t-rozdelenia s ν (efektívnymi) stupňami voľnosti (DF), ktoré sa odhadujú pomocou Satterthwaiteovej aproximácie.

- Nulové rozdelenie 'F-štatistík' sa obyčajne aproximuje F-rozdelením s ν_1 a ν_2 stupňami voľnosti, pričom $\nu_1 = \ell$.

- Efektívne stupne voľnosti (ν_2) sú odhadované, štandardne zovšeobecnenou Satterthwaiteovou metódou, Fai-Cornelius (1996), alebo inými alternatívnymi aproximatívnymi metódami, Kenward-Roger (1997).

Satterthwaiteová metóda odhadovania efektívnych stupňov voľnosti

- Giesbrecht-Burns (1985) navrhli odhadovať rodelenie t-štatistiky pomocou Satterthwaiteovej aproximácie:

$$t = \frac{\hat{w} - w_0}{\sqrt{\lambda' \hat{C} \lambda}} \sim t_{\hat{\nu}},$$

$$\hat{\nu} = \frac{2\left(\lambda' \hat{C} \lambda\right)^2}{\widehat{Var}_{\hat{\theta}}\left(\lambda' \hat{C} \lambda\right)} \equiv \frac{2\left(\lambda' \hat{C} \lambda\right)^2}{\hat{g}' \hat{\Sigma} \hat{g}}$$

- $\hat{g}' \hat{\Sigma} \hat{g}$ je odhad variancie $Var_{\hat{\theta}}\left(\lambda' \hat{C} \lambda\right)$

  založený na Taylorovom rozvoji $\lambda' C \lambda$ s ohľadom na θ, pričom $\hat{\Sigma}$ označuje odhadnutú kovariančnú maticu odhadov variančných komponentov.
- $\hat{g}$ je odhadnutá verzia gradientu $g = \nabla_\theta(\lambda' C \lambda)$:

$$g = \begin{pmatrix} \frac{\partial(\lambda' C \lambda)}{\partial \theta_1} \\ \vdots \\ \frac{\partial(\lambda' C \lambda)}{\partial \theta_{n_V}} \end{pmatrix} \overset{\hat{\lambda}=\hat{C}\lambda}{\Longrightarrow} \hat{g} = \begin{pmatrix} \hat{\lambda}' \frac{\partial \hat{H}}{\partial \hat{\theta}_1} \hat{\lambda} \\ \vdots \\ \hat{\lambda}' \frac{\partial \hat{H}}{\partial \hat{\theta}_{n_V}} \hat{\lambda} \end{pmatrix} \overset{\hat{\theta}=(\hat{\sigma}_1^2,\ldots,\hat{\sigma}_{r+1}^2)}{\Longrightarrow} \hat{g} = \begin{pmatrix} \frac{-1}{(\hat{\sigma}_1^2)^2} \hat{\lambda}_1' \hat{\lambda}_1 \\ \vdots \\ \frac{-1}{(\hat{\sigma}_r^2)^2} \hat{\lambda}_r' \hat{\lambda}_r \\ \frac{-1}{(\hat{\sigma}_{r+1}^2)^2} \hat{\lambda}' H_{R_l} \hat{\lambda} \end{pmatrix}$$

Faiova-Corneliusová metóda odhadovania efektívnych stupňov voľnosti

- Fai a Cornelius (1996) navrhli zovšeobecnenie Satterthwaiteovej aproximácie pre prípad mnohorozmernej funcie $w = \Lambda' \left(b', u'\right)'$, teda pre odhad efektívnych stupňov voľnosti (DDF) nulového rozdelenia F-štatistiky

$$F = \frac{1}{\ell} \left(\hat{w} - w_0\right)' \left(\Lambda' \hat{C} \Lambda\right)^{-1} \left(\hat{w} - w_0\right) \sim F_{\ell,\nu},$$

  kde $\ell = rank(\Lambda')$.
- Metóda je založená na rozložení viacrozmerného problému na viac jednorozmerných problémov. Faiova-Corneliusová aproximácia je

$$\hat{\nu} = \frac{2\hat{E}}{\hat{E} - \ell}, \quad \text{kde} \quad \hat{E} = \sum_{i=1}^{\ell} \frac{\hat{\nu}_i}{\hat{\nu}_i - 2} 1_{\{\hat{\nu}_i > 2\}} .$$

- $\hat{\nu}_i$, $i = 1, \ldots, \ell$, sú efektívne stupňe voľnosti odhadnuté Satterthwaiteovou aproximáciou t-štatistík pre funkcie $\hat{w}_i = \hat{\lambda}_i' \left(\hat{b}', \hat{u}'\right)'$, kde $\hat{\lambda}_i$, $i = 1, \ldots, \ell$, sú stĺpce matice $\hat{\Lambda}_{FC} = \Lambda \hat{U}$
- $\hat{U}$ označuje ortonormálnu maticu zo spektrálneho rozkladu matice $\Lambda' \hat{C} \Lambda$, teda, je to taká matica, že platí $\hat{U}' \Lambda' \hat{C} \Lambda \hat{U} = \hat{S}$, kde $\hat{S}$ je diagonálna matica.

Kenwardova-Rogerová aproximácia

- Harville (2008) dlhodobo poukazoval na nekorektnosť použitia MSE matice prediktora BLUP $\tilde{w}$ v situácii, keď je inferencia založená na EBLUPe - empirickom prediktore $\hat{w}$ - a namiesto jeho skutočnej MSE matice, $M_{\hat{w}}$, sa nekorektne použije MSE matica BLUP prediktora $\tilde{w}$, $M_{\tilde{w}} = \Lambda' C \Lambda$, (resp. odhadnuté verzie týchto matíc).
- Prvý zdroj vychýlenia takéhoto nesprávneho odhadu MSE matice možno vysvetliť rozkladom predikčnej chyby:

$$(\hat{w} - w) = (\tilde{w} - w) + (\hat{w} - \tilde{w}),$$

- Teda pre MSE maticu empirického prediktora $\hat{w}$ dostávame

$$M_{\hat{w}} = M_{\tilde{w}} + M_{\delta\hat{w}},$$

kde $M_{\delta\hat{w}} = E\left((\hat{w} - \tilde{w})(\hat{w} - \tilde{w})'\right) = Var(\hat{w} - \tilde{w})$.

- Odtiaľ teda dostávame dôkaz o podhodnotení skutočnej matice MSE prediktora EBLUP: $M_{\hat{w}} \geq M_{\tilde{w}}$.

Kenwardova-Rogerová aproximácia

- S využitím aproximácie $\dot{M}_{\delta\hat{w}}$ založenej na Taylorovom rozvoji na rozvoji predikčnej chyby $\hat{w} - \tilde{w}$, dostávame aproximáciu MSE matice EBLUP prediktora $\hat{w}$ v tvare

$$\dot{M}_{\hat{w}} = M_{\tilde{w}} + \dot{M}_{\delta\hat{w}}, \qquad \text{pričom} \qquad \dot{M}_{\delta\hat{w}} = -\frac{1}{2}\sum_{i=1}^{n_V}\sum_{j=1}^{n_V}\Sigma_{ij}M_{\tilde{w}}^{(i,j)},$$

kde $M_{\tilde{w}}^{(i,j)} = \Lambda' C \frac{\partial^2 H}{\partial\theta_i \partial\theta_j} C\Lambda = \tilde{\Lambda}' \frac{\partial^2 H}{\partial\theta_i \partial\theta_j}\tilde{\Lambda}$, pričom $\tilde{\Lambda} = C\Lambda$.

- MSE matica $M_{\hat{w}}$ ako aj jej aproximácia $\dot{M}_{\hat{w}}$ však závisí od matice Σ a od neznámych variančných-kovariančných komponentov.
- Ak nahradíme tieto neznáme parametre ich odhadmi (REML) dostaneme prirodzený odhad MSE matice EBLUP prediktora v tvare

$$\widehat{\dot{M}}_{\hat{w}} = \widehat{M}_{\tilde{w}} + \widehat{\dot{M}}_{\delta\hat{w}}$$

- Dôležitým kritériom korektnosti použitia odhadu MSE matice EBLUP prediktora je kritérium nevychýlenosti. Od rozumného odhadu teda očakávame, že

$$E\left(\widehat{\dot{M}}_{\hat{w}}\right) \approx M_{\hat{w}}$$

- To však v prípade odhadu $\widehat{\dot{M}}_{\hat{w}}$ nie je pravda! Hoci platí

$$E\left(\widehat{\dot{M}}_{\delta\hat{w}}\right) \approx M_{\delta\hat{w}}$$

Kenwardova-Rogerová aproximácia

- Ako ukázali Kenward a Roger (1997), pre LMM s lineárnou parametrizáciou kovariančnej matice platí

$$E\left(\widehat{M}_{\tilde{w}}\right) \approx M_{\tilde{w}} - M_{\delta\hat{w}}$$

- Optimálnym odhadom MSE matice EBLUP prediktora je teda

$$\widehat{\dot{M}}_{\hat{w},A} = \widehat{M}_{\tilde{w}} + 2\widehat{\dot{M}}_{\delta\hat{w}}$$

Kenwardova-Rogerová aproximácia

- Kenward a Roger (1997) navrhli aproximáciu rozdelenia testovacej F-štatistiky pre testovanie nulovej hypotézy pre lineárne funckcie pevných efektov $K'b$ založenej na približne nevychýlenom odhade kovariančnej matice empirického BLUE $K'\hat{b}$, teda odhade matice $K'\hat{C}_{11}K$.

- Prirodzeným zovšeobecnením tohto prístupu je aproximácia rozdelenia F-štatistiky pre vektor lineárnych funckií pevných aj náhodnách efektov:

$$F = \frac{1}{\ell}(\hat{w} - w_0)' \left(\widehat{\dot{M}}_{\hat{w},A}\right)^{-1} (\hat{w} - w_0)$$

  kde $\widehat{\dot{M}}_{\hat{w},A}$ je približne nevychýlený odhad MSE matice.

- Uvažovaný typ aproximácie rozdelenia je v tvare

$$\kappa F \overset{approx.}{\sim} F_{\ell,\nu},$$

  pričom neznáme parametre κ a ν sú odhadované z pozorovaných dát.

Kenwardova-Rogerová aproximácia

- Analogicky ako v prípade aproximácie pre funkcie fixných efektov, na základe momentovej metódy, v jednoduchom LMM sú odhady parametrov κ a ν dané v tvare

$$\begin{aligned}\hat{\kappa} &= \frac{\hat{\nu}}{\hat{E}\,(\hat{\nu}-2)},\\ \hat{\nu} &= 4+\frac{2+\ell}{\ell\hat{\varrho}-1},\end{aligned}$$

$$\begin{aligned}\hat{\varrho} &= \frac{\hat{V}}{2\hat{E}^2}, \quad \hat{E}=1+\frac{\hat{A}_2}{\ell},\\ \hat{V} &= \frac{2}{\ell}\left(1+\hat{B}\right), \quad \hat{B}=\frac{1}{2\ell}\left(\hat{A}_1+6\hat{A}_2\right),\end{aligned}$$

$$\begin{aligned}\hat{A}_1 &= \sum_{i=1}^{n_V}\sum_{j=1}^{n_V}\hat{\Sigma}_{ij}\,tr\left(\widehat{M}_{\tilde{w}}^{-1}\widehat{M}_{\tilde{w}}^{(i)}\right)tr\left(\widehat{M}_{\tilde{w}}^{-1}\widehat{M}_{\tilde{w}}^{(j)}\right),\\ \hat{A}_2 &= \sum_{i=1}^{n_V}\sum_{j=1}^{n_V}\hat{\Sigma}_{ij}\,tr\left(\widehat{M}_{\tilde{w}}^{-1}\widehat{M}_{\tilde{w}}^{(i)}\widehat{M}_{\tilde{w}}^{-1}\widehat{M}_{\tilde{w}}^{(j)}\right).\end{aligned}$$

kde $\widehat{M}_{\tilde{w}}^{(i)}=\frac{\partial\widehat{M}_{\tilde{w}}}{\partial\hat{\theta}_i}=\Lambda'\hat{C}\frac{\partial\hat{H}}{\partial\hat{\theta}_i}\hat{C}\Lambda=\hat{\Lambda}'\frac{\partial\hat{H}}{\partial\hat{\theta}_i}\hat{\Lambda}$, pričom $\hat{\Lambda}=\hat{C}\Lambda$.