60th Annual Meeting of the European Federation of Animal Science
Barcelona, Spain August 24-27, 2009. Session 53, Paper 1

Validation of variance component estimation and BLUP software

Generating benchmark problems to evaluate variance component estimation software

M. Wensch-Dorendorf[1]*, J. Wensch[2], H. H. Swalve[1]

*presenting (monika.dorendorf@landw.uni-halle.de)

[1]Institute of Agricultural and Nutritional Sciences, University Halle-Wittenberg
[2]Institute of Scientific Computing, Technical University Dresden

Introduction

- Question 1: „ Is it possible to generate phenotypes such that a **genetic evaluation** hits the known values exactly?“ YES!

- Knowledge of variance components is a must for genetic evaluation

- Logical question: „Is it possible to generate phenotypes such that a **variance component estimation** hits the known values exactly?“

- But: **Variance component estimation is a more complex task**

  Precisely: **Nonlinear minimization problems** have to be solved by numerical algorithms

Simplest model: 1-way classification with balanced data

- The mathematical model for the j-th measurement of animal i:

  $$y_{ij} = \mu + u_i + e_{ij}$$

  - a animals, randomly selected, each animal has n measurements
  - μ is the overall mean
  - u_i are random animal effects with $u \sim N(0,\sigma^2_a)$
  - e_{ij} are residual effects with $e \sim N(0,\sigma^2_e)$
  - e und u are uncorrelated

  $(N=a \cdot n)$

Simplest model: 1-way classification with balanced data – the idea

- prescribe variances σ^2_a and σ^2_e for random effects u and e
- simulate random data u^0, e^0 based on the prescribed variances -> y^0
- **ANOVA** & **REML** :

  $E(SSA)=(a-1)(n\sigma^2_a+\sigma^2_e)$ and $E(SSE)=a(n-1)\sigma^2_e$
- **ML** : $E(SSA)=a(n\sigma^2_a+\sigma^2_e)$ and $E(SSE)=a(n-1)\sigma^2_e$

  (well known formulas)
- Find a minimum norm correction y for y^0 such that the prescribed variances are obtained as an estimator

$$\frac{1}{2}\left\|y-y^0\right\|_2^2 \to \min_y \quad \text{w.r.t} \quad SSA_y = E(SSA) \quad SSE_y = E(SSE)$$

Simplest model: 1-way classification with balanced data – ANOVA, REML, ML

❖ $SSA_y := y^T(C-B)y \quad SSE_y := y^T(\mathcal{I}-C)y$

$$\text{where } B=\frac{1}{N}1_N 1_N^T \quad C=\frac{1}{n}ZZ^T \quad (N=n\cdot a)$$

($B=B^T$, $B^2=B$ -> B is an ortho-projector as well as C, $\mathcal{I}$-B,$\mathcal{I}$-C,C-B)

❖ We couple the constraints by Lagrangian multipliers to the minimum norm condition, a necessary condition for optimality is then

$$\frac{\partial}{\partial y}\left(\frac{1}{2}\left\|y-y^0\right\|_2^2+\frac{\lambda_a}{2}\left(y^T(C-B)y-E(SSA)\right)+\frac{\lambda_e}{2}\left(y^T(\mathcal{I}-C)y-E(SSE)\right)\right)=0$$

❖ **Differentition yields** $\mathbf{y-y^0-\lambda_a(C-B)y-\lambda_e(\mathcal{I}-C)y=0}$
-> y^0 is a linear combination of y, By, Cy

Simplest model: 1-way classification with balanced data – ANOVA, REML, ML

→ $y = \beta_1 y^0 + \beta_2 B y^0 + \beta_3 C y^0$ with $\beta_1+\beta_2+\beta_3=1$ (use: B,C ortho-projectors)

- Minimization problems for ANOVA, ML, REML reduce to a biquadratic equation -> can be solved directly

- With $M=(y^0, By^0, Cy^0)$ and $\beta=(\beta_1, \beta_2, \beta_3)^T$ we have $y=M\beta$

- The constraints result in: $SSA_y = \beta^T M^T (C-B) M \beta = E(SSA)$
$$SSE_y = \beta^T M^T (I-B) M \beta = E(SSE)$$

→ three equations for the unknown coefficients β :

$$\mathbf{1^T \beta=1\ ,\ (\beta_1+\beta_3)^2 SSA_0=E(SSA)\ ,\ \beta^2{}_1 SSE_0=E(SSE)}$$

- Choosing the sign of the root such that $\beta=(1,0,0)^T$ when $SSA_0=E(SSA)$, $SSE_0=E(SSE)$ -> $\boldsymbol{\beta_1 = (E(SSE)/SSE_0)^{\frac{1}{2}}}$

$$\boldsymbol{\beta_3 = -\beta_1 + (E(SSA)/SSA_0)^{\frac{1}{2}}}$$
$$\boldsymbol{\beta_2 = 1 - \beta_1 - \beta_3}$$

Simplest model: 1-way classification with balanced data – simulation of the data

- Prescribe variances σ^2_a and σ^2_e for random effects u and e

- simulate random data u^0, e^0 based on the prescribed variances -> y^0

- Evaluate SSE_0, SSA_0 for y^0, and E(SSE), E(SSA) by using the prescribed variance components -> evaluate β_1 , β_2 , β_3 -> evaluate the corrected y by using $y= \beta_1 y^0+\beta_2 By^0+\beta_3 Cy^0$

- Estimate the variance components for y (SAS, proc mixed f.i.) -> the solutions are the prescribed variances σ^2_a and σ^2_e

1-way classification with unbalanced data

❖ The mathematical model for the j-th measurement of animal i:

$$y_{ij} = \mu + u_i + e_{ij}$$

- a animals, randomly selected, animal i has n_i measurements
- μ is the overall mean
- u_i are random animal effects with $u \sim N(0,\sigma^2_a)$
- e_{ij} are residual effects with $e \sim N(0,\sigma^2_e)$
- e und u are uncorrelated

  $(\Sigma_i n_i = N)$

1-way classification with unbalanced data - ML

- Use L'=-k log L instead of likelihood L (k=constant factors in L independent of the parameters) $\left(\text{with } \xi_i = n_i\sigma_a^2 + \sigma_e^2\right)$

$$L' = \frac{1}{2}\sum_{i=1}^{a}\log(\xi_i) + \frac{1}{2}(M-a)\log\sigma_e^2 + \frac{1}{2\sigma_e^2}\sum_{i,j}(y_{ij}-\mu)^2 - \frac{1}{2\sigma_e^2}\sum_{i=1}^{a}n_i\frac{n_i\sigma_a^2}{\xi_i}(\bar{y}_i-\mu)^2$$

- setting the corresponding partial derivatives to zero ($\partial L'/\partial \mu$, $\partial L'/\partial \sigma^2_a$, $\partial L'/\partial \sigma^2_e$)-> necessary condition for a minimizer

1-way classification with unbalanced data - ML

- Use L'=-k log L instead of likelihood L (k=constant factors in L independent of the parameters) $\left(\text{with } \xi_i = n_i\sigma_a^2 + \sigma_e^2\right)$

$$L' = \frac{1}{2}\sum_{i=1}^{a}\log(\xi_i) + \frac{1}{2}(M-a)\log\sigma_e^2 + \frac{1}{2\sigma_e^2}\sum_{i,j}(y_{ij}-\mu)^2 - \frac{1}{2\sigma_e^2}\sum_{i=1}^{a} n_i \frac{n_i\sigma_a^2}{\xi_i}(\bar{y}_i-\mu)^2$$

- setting the corresponding partial derivatives to zero ($\partial L'/\partial\mu$, $\partial L'/\partial\sigma^2_a$, $\partial L'/\partial\sigma^2_e$)-> necessary condition for a minimizer

  - $$\frac{\partial L'}{\partial\mu} = 0 \quad \Rightarrow \sum_{i=1}^{a}\frac{n_i}{\xi_i}(\bar{y}_i-\mu) = 0$$
  - $$\frac{\partial L'}{\partial\sigma_a^2} = 0 \;\Rightarrow \sum_{i=1}^{a}\frac{n_i^2}{\xi_i^2}(\bar{y}_i-\mu)^2 = S_1 := \sum_{i=1}^{a}\frac{n_i}{\xi_i}$$
  - $$\frac{\partial L'}{\partial\sigma_e^2} = 0 \;\Rightarrow \sum_{i,j}(y_{ij}-\bar{y}_i)^2 + \sum_{i=1}^{a}\frac{n_i\sigma_e^4}{\xi_i^2}(\bar{y}_i-\mu)^2 = S_2 := \sum_{i=1}^{a}\frac{\sigma_e^4}{\xi_i} + (M-a)\sigma_e^2$$

1-way classification with unbalanced data - ML

- Use L'=-k log L instead of likelihood L (k=constant factors in L independent of the parameters) $\left(\text{with } \xi_i = n_i\sigma_a^2 + \sigma_e^2\right)$

$$L' = \frac{1}{2}\sum_{i=1}^{a}\log(\xi_i) + \frac{1}{2}(M-a)\log\sigma_e^2 + \frac{1}{2\sigma_e^2}\sum_{i,j}(y_{ij}-\mu)^2 - \frac{1}{2\sigma_e^2}\sum_{i=1}^{a} n_i \frac{n_i\sigma_a^2}{\xi_i}(\bar{y}_i-\mu)^2$$

- setting the corresponding partial derivatives to zero ($\partial L'/\partial \mu$, $\partial L'/\partial \sigma^2_a$, $\partial L'/\partial \sigma^2_e$)-> necessary condition for a minimizer
- Following the same procedure as for balanced data: prescribe μ, σ^2_a, σ^2_e and simulate u^0, e^0 , evaluate y^0
- Determine a minimum norm correction for the resulting $y^0 = \mu 1_M + Zu^0 + e^0$ such that the ML-estimate of the corrected value y yields the precribed variances and μ-value

1-way classification with unbalanced data - ML

- Thus **minimize** $\left\| y - y^0 \right\|$ **under the necessary condition for a L' minimizer** = constraints

- We couple the constraints by Lagrangian multipliers λ_μ, λ_e, λ_a to the minimum norm condition, a necessary condition for optimality is then

$$\frac{\partial}{\partial y_{ij}}\left[\frac{1}{2}\sum_{i,j}(y_{ij}-y_{ij}^0)^2+\lambda_\mu\sum_i\frac{n_i}{\xi_i}(\overline{y}_i-\mu)+\frac{1}{2}\lambda_a\sum_i\frac{n_i^2}{\xi_i^2}(\overline{y}_i-\mu)^2 + \frac{1}{2}\lambda_e\left[\sum_{i,j}(y_{ij}-\overline{y}_i)^2+\sum_i\frac{n_i\sigma_e^4}{\xi_i^2}(\overline{y}_i-\mu)^2\right]\right]=0$$

- N equations for N+3 variables: y_{ij}, λ_μ, λ_e, λ_a

- Together with three constraints: N+3 equations for N+3 variables

1-way classification with unbalanced data - ML

- We substitue $z_i := \bar{y}_i - \mu$ and $e_{ij} := y_{ij} - \bar{y}_i$ in the constraints and minimizing condition and obtain

$$(1)\ \sum_i \frac{n_i}{\xi_i} z_i = 0 \qquad (2)\ \sum_i \frac{n_i^2}{\xi_i^2} z_i^2 = S_1 \qquad (3)\ \sum_{i,j} e_{ij}^2 + \sigma_e^4 \sum_i \frac{n_i}{\xi_i^2} z_i^2 = S_2$$

$$(4)\ e_{ij} + z_i + \lambda_\mu \frac{1}{\xi_i} + \lambda_a \frac{n_i}{\xi_i^2} z_i + \lambda_e \left[e_{ij} + \sigma_e^4 \frac{1}{\xi_i^2} z_i \right] = y_{ij}^0 - \mu \qquad \forall i, j$$

$$(5)\ \sum_j e_{ij} = 0 \qquad \forall i$$

- (4), (5) are equivalent to (6), (7) (summing over j in (4), using (5))

$$(6)\ z_i + \lambda_\mu \frac{1}{\xi_i} + \lambda_a \frac{n_i}{\xi_i^2} z_i + \lambda_e \sigma_e^4 \frac{1}{\xi_i^2} z_i = \bar{y}_{ij}^0 - \mu \qquad \forall i$$

$$(7)\ e_{ij} = \frac{1}{1 - \lambda_e} (y_{ij}^0 - \bar{y}_i^0) \qquad \forall i, j$$

1-way classification with unbalanced data - ML

- We set: $x_i = \frac{n_i}{\xi_i} z_i \quad , \alpha_i = \frac{\sigma_e^4}{n_i}, \ \beta_i = \frac{\xi_i}{n_i}, S_4 = \sum_{i,j} (y_{ij}^0 - \bar{y}_i^0)^2 \ , \ R_i = \xi_i (\bar{y}_i^0 - \mu)$

$$\sum_i x_i = 0$$

$$\sum_i x_i^2 = S_1$$

$$(1-\lambda_e)^2 \left[\sum_i \alpha_i x_i^2 - S_2 \right] = S_4$$

where $\quad x_i := \frac{R_i - \lambda_\mu}{\beta_i + \lambda_a + \lambda_e \alpha_i}$

→ a system of three equations for the 3 unknowns λ_μ, λ_a, λ_e

- no further simplification available -> we have to rely on a numerical solution of these equations to generate benchmark sets

Summary

- The projection method correct simulated phenotypic data such that the estimated variances are equal to the precribed variances
- Balanced case: analytical solution is available
- Unbalanced case: use high accuracy numerical solution
- Work in progress: 2-way classification & models with pedigree

Example: 1-way balanced data

- Prescribe variances: σ^2_a =4 and σ^2_e =16 , μ=20.0
- ANOVA/REML: β_1 = 1.142099, β_2 =0.242734, β_3 =-0.38483
- ML: β_1 = 1.142099, β_2 =0.153351, β_3 =-0.29545

animal	y-original	y-corr_anova	y-corr_ml	measurement
1	21.31103230	19.94709525	20.27204752	1
1	28.08764151	27.68665567	28.01160794	2
2	17.90074808	18.64633517	18.36876571	3
2	18.01613969	18.77812384	18.50055439	4
3	19.63635102	18.84951758	18.98515726	5
3	25.52633969	25.57646933	25.71210901	6
4	23.28091046	21.94554247	22.32887414	7
4	27.42403753	26.67740487	27.06073653	8
5	12.23411615	13.41782504	12.85147090	9
5	17.22104056	19.11338776	18.54703362	10

- Use SAS proc mixed with y-corr_ml/_anova (ml/reml option) -> estimates are the prescribed variances σ^2_a =4 and σ^2_e =16

Example: 1-way balanced data