Random regression models for analyses of longitudinal data in animal breeding

Karin Meyer
University of New England, Animal Genetics and Breeding Unit
Armidale, NSW 2351, Australia
kmeyer@didgeridoo.une.edu.au

1. Introduction

Animal breeding is concerned with the genetic improvement of farmed livestock. A central task is the estimation of genetic parameters or, equivalently, variance components, required to design selection programmes and in identification of genetically superior animals. Statistical techniques used rely heavily on linear, mixed models. Whilst some traits of interest are measured only once per animal, others are recorded repeatedly and may change, gradually and continually, as time progresses. Typical examples are test day records for dairy cows, with milk production at the beginning and end of lactation having quite different means and variances but high genetic correlations, and growth of meat-producing animals. Recently, so-called random regression (RR) models have become popular for the analysis of such data, as they allow complete 'growth curves' to be fitted within the linear, mixed model framework, correctly modelling changes in mean and dispersion with time, and are suitable for large scale applications. This paper reviews the use of RR models in analyses of data from livestock improvement schemes, concentrating on variance component estimation.

2. Random Regression Models

Typically, RR models for genetic analyses include at least two sets of RR coefficients for each animal i, representing the direct, additive genetic (α_{im}) and permanent environmental (γ_{im}) effects of the animal. Let y_{ij} denote the $j-$th record for animal i taken at age (or time) a_{ij}. Covariables are functions of age, $\phi_m(a_{ij})$. Orthogonal polynomials are the most common choice, as they require few assumptions about the shape of the trajectory to be modelled, but other functions, such as spline or trigonometric functions have been used. This gives the linear model

$$(1) \quad y_{ij} = F_{ij} + \sum_{m=0}^{k_A-1} \alpha_{im}\phi_m(a_{ij}) + \sum_{m=0}^{k_R-1} \gamma_{im}\phi_m(a_{ij}) + \varepsilon_{ij}$$

where F_{ij} denotes the fixed effects and ε_{ij} the temporary environmental effect or 'measurement error' affecting y_{ij}. The number of regression coefficients fitted is given by k_A and k_R. In matrix notation,

$$(2) \quad \mathbf{y} = \mathbf{Xb} + \mathbf{\Phi}_A\boldsymbol{\alpha} + \mathbf{\Phi}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

with $\mathbf{y}$, $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\varepsilon}$ the vectors of observations, RR coefficients and residuals, respectively, assumed to be ordered according to animals, and $\mathbf{b}$ denoting the vector of fixed effects fitted. $\mathbf{X}$, $\mathbf{\Phi}_A$ and $\mathbf{\Phi}$ are the corresponding design matrices. Assume that $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\varepsilon}$ are uncorrelated and that

$$\begin{array}{lll} E[\boldsymbol{\alpha}] = \mathbf{0} & E[\boldsymbol{\gamma}] = \mathbf{0} & E[\boldsymbol{\varepsilon}] = \mathbf{0} \\ Var(\boldsymbol{\alpha}) = \mathbf{A}\otimes\mathbf{K}_A = \mathbf{G} & Var(\boldsymbol{\gamma}) = \mathbf{I}_N\otimes\mathbf{K}_R = \mathbf{R} & Var(\boldsymbol{\varepsilon}) = Diag\left\{\sigma^2_k\right\} = \mathbf{\Sigma}_\varepsilon \end{array}$$

with N denoting the number of animals with records, and $\mathbf{I}_N$ an identity matrix of size N. This gives

$$(3) \quad Var(\mathbf{y}) = \mathbf{\Phi}_A(\mathbf{A}\otimes\mathbf{K}_A)\mathbf{\Phi}_A{}' + \mathbf{\Phi}(\mathbf{I}_N\otimes\mathbf{K}_R)\mathbf{\Phi}' + Diag\left\{\sigma^2_k\right\} = \mathbf{\Phi}_A\mathbf{G}\mathbf{\Phi}_A{}' + \mathbf{\Phi}\mathbf{R}\mathbf{\Phi}' + \mathbf{\Sigma}_\varepsilon = \mathbf{V}$$

Partitioning of animal effects into their genetic and non-genetic components requires information on relationships between animals, which is provided by the numerator relationship matrix $\mathbf{A}$. Animals

in the pedigree only are included in $\boldsymbol{\alpha}$, which has length $N_A \times k_A$ for $N_A > N$, with the corresponding rows of $\boldsymbol{\Phi}_A$ having elements zero. $\mathbf{K}_A = \{K_{Aij}\}$ and $\mathbf{K}_R = \{K_{Rij}\}$ are the matrices of covariances between RR coefficients, and σ_k^2 denote the measurement error variances. If assumed heterogeneous, changes in σ_k^2 with age at recording are commonly modelled through a step function or a low order polynomial variance function of σ_k^2 or $\log(\sigma_k^2)$.

3. Estimation of Breeding Values

For known variances, solutions to the mixed model equations (4) yield best linear unbiased predictions (BLUP) of $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$. Coefficients α_{im} for animal i define it's genetic merit for all ages

$$(4) \quad \begin{bmatrix} \mathbf{X}'\boldsymbol{\Sigma}_\varepsilon^{-1}\mathbf{X} & \mathbf{X}'\boldsymbol{\Sigma}_\varepsilon^{-1}\boldsymbol{\Phi}_A & \mathbf{X}'\boldsymbol{\Sigma}_\varepsilon^{-1}\boldsymbol{\Phi} \\ \boldsymbol{\Phi}_A'\boldsymbol{\Sigma}_\varepsilon^{-1}\mathbf{X} & \boldsymbol{\Phi}_A'\boldsymbol{\Sigma}_\varepsilon^{-1}\boldsymbol{\Phi}_A + \mathbf{A}^{-1}\otimes\mathbf{K}_A^{-1} & \boldsymbol{\Phi}_A'\boldsymbol{\Sigma}_\varepsilon^{-1}\boldsymbol{\Phi} \\ \boldsymbol{\Phi}'\boldsymbol{\Sigma}_\varepsilon^{-1}\mathbf{X} & \boldsymbol{\Phi}'\boldsymbol{\Sigma}_\varepsilon^{-1}\boldsymbol{\Phi}_A & \boldsymbol{\Phi}'\boldsymbol{\Sigma}_\varepsilon^{-1}\boldsymbol{\Phi} + \mathbf{I}\otimes\mathbf{K}_R^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\boldsymbol{\alpha}} \\ \hat{\boldsymbol{\gamma}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\boldsymbol{\Sigma}_\varepsilon^{-1}\mathbf{y} \\ \boldsymbol{\Phi}_A'\boldsymbol{\Sigma}_\varepsilon^{-1}\mathbf{y} \\ \boldsymbol{\Phi}'\boldsymbol{\Sigma}_\varepsilon^{-1}\mathbf{y} \end{bmatrix}$$

within the range considered, and estimated breeding values for target ages a_k are obtained simply by evaluating the regression curve, $\sum_{m=0}^{k_A-1}\alpha_{im}\phi_m(a_k)$. In addition, functions of the curve may be of interest, e.g. integrals of estimated lactation curves to estimate total lactation genetic merit for dairy cows, or turning points of growth curves to distinguish between early and late maturing animals.

4. Estimation of Covariance Functions

Covariance functions (CF) defining genetic ($\mathscr{A}$) and permanent environmental ($\mathscr{R}$) covariances between ages a_i and a_j are given by $\mathbf{K}_A$ and $\mathbf{K}_R$.

$$(5) \quad \mathscr{A}(a_i,a_j) = \sum_{m=0}^{k_A-1}\sum_{n=0}^{k_A-1}\phi_m(a_i)\phi_n(a_j)K_{Amn} \qquad \text{and} \qquad \mathscr{R}(a_i,a_j) = \sum_{m=0}^{k_R-1}\sum_{n=0}^{k_R-1}\phi_m(a_i)\phi_n(a_j)K_{Rmn}$$

With $\mathscr{A}$ and $\mathscr{R}$ generally fitted to reduced order, i.e. k_A and k_R smaller, often much smaller, than the number of ages in the data, the resulting estimates of covariance matrices among observations are smoothed and have reduced rank. Estimates of CFs can be obtained by restricted maximum likelihood (REML), using a derivative-free or an 'average information' (Gilmour et al., 1995) algorithm, or Bayesian analysis. For $\mathbf{C}$ the coefficient matrix in (4) and $\mathbf{y}'\mathbf{P}\mathbf{y}$ is the weighted sum of squares of residuals ($\mathbf{P}\mathbf{y} = \boldsymbol{\Sigma}_\varepsilon^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}} - \boldsymbol{\Phi}_A\hat{\boldsymbol{\alpha}} - \boldsymbol{\Phi}\hat{\boldsymbol{\gamma}})$), the REML log likelihood for (2) is

$$(6) \quad -2\log\mathscr{L} = const + \log|\mathbf{G}| + \log|\mathbf{R}| + \log|\boldsymbol{\Sigma}_\varepsilon| + \log|\mathbf{C}| + \mathbf{y}'\mathbf{P}\mathbf{y}$$

Both $\mathbf{y}'\mathbf{P}\mathbf{y}$ and $\log|\mathbf{C}|$ can be evaluated by factoring (Graser et al., 1987)

$$(7) \quad \mathbf{M} = \begin{bmatrix} \mathbf{X}'\boldsymbol{\Sigma}_\varepsilon^{-1}\mathbf{X} & \mathbf{X}'\boldsymbol{\Sigma}_\varepsilon^{-1}\boldsymbol{\Phi}_A & \mathbf{X}'\boldsymbol{\Sigma}_\varepsilon^{-1}\boldsymbol{\Phi} & \mathbf{X}'\boldsymbol{\Sigma}_\varepsilon^{-1}\mathbf{y} \\ \boldsymbol{\Phi}_A'\boldsymbol{\Sigma}_\varepsilon^{-1}\mathbf{X} & \boldsymbol{\Phi}_A'\boldsymbol{\Sigma}_\varepsilon^{-1}\boldsymbol{\Phi}_A + \mathbf{A}^{-1}\otimes\mathbf{K}_A^{-1} & \boldsymbol{\Phi}_A'\boldsymbol{\Sigma}_\varepsilon^{-1}\boldsymbol{\Phi} & \boldsymbol{\Phi}_A'\boldsymbol{\Sigma}_\varepsilon^{-1}\mathbf{y} \\ \boldsymbol{\Phi}'\boldsymbol{\Sigma}_\varepsilon^{-1}\mathbf{X} & \boldsymbol{\Phi}'\boldsymbol{\Sigma}_\varepsilon^{-1}\boldsymbol{\Phi}_A & \boldsymbol{\Phi}'\boldsymbol{\Sigma}_\varepsilon^{-1}\boldsymbol{\Phi} + \mathbf{I}\otimes\mathbf{K}_R^{-1} & \boldsymbol{\Phi}'\boldsymbol{\Sigma}_\varepsilon^{-1}\mathbf{y} \\ \mathbf{y}'\boldsymbol{\Sigma}_\varepsilon^{-1}\mathbf{X} & \mathbf{y}'\boldsymbol{\Sigma}_\varepsilon^{-1}\boldsymbol{\Phi}_A & \mathbf{y}'\boldsymbol{\Sigma}_\varepsilon^{-1}\boldsymbol{\Phi} & \mathbf{y}'\boldsymbol{\Sigma}_\varepsilon^{-1}\mathbf{y} \end{bmatrix}$$

while the other terms in (6) can be determined indirectly. First derivatives are obtained similarly, using an 'automatic differentiation' of the Cholesky factor of $\mathbf{M}$ (Smith, 1995), which only requires derivatives of $\mathbf{M}$ to be evaluated. For $\mathbf{M} = \mathbf{L}\mathbf{L}'$, of size $M \times M$, l_{ii} the $i-$th diagonal element of $\mathbf{L}$, θ_k the $k-$th variance component to be estimated and $\partial\mathbf{L}/\partial\theta_k = \{\partial l_{ij}/\partial\theta_k\}$,

$$(8) \quad \log|\mathbf{C}| = 2\sum_{i=1}^{M-1}\log(l_{ii}) \qquad \partial\log|\mathbf{C}|/\partial\theta_k = 2\sum_{i=1}^{M-1} l_{ii}^{-1}\partial l_{ii}/\partial\theta_k$$

$$(9) \quad \mathbf{y}'\mathbf{P}\mathbf{y} = l_{MM}^2 \qquad \partial\mathbf{y}'\mathbf{P}\mathbf{y}/\partial\theta_k = l_{MM}\,\partial l_{MM}/\partial\theta_k$$

$$(10) \quad \log|\boldsymbol{\Sigma}_\varepsilon| = \sum_k \log(\sigma_k^2) \qquad \log|\mathbf{G}| = N_A\log|\mathbf{K}_A| + k_A\log|\mathbf{A}| \qquad \log|\mathbf{R}| = N\log|\mathbf{K}_R|$$

The average of observed and expected information (AI) is proportional to second derivatives of $\mathbf{y}'\mathbf{P}\mathbf{y}$, the 'data part' of $\log\mathscr{L}$, and thus considerably easier to compute than either of the former, $\partial^2\mathbf{y}'\mathbf{P}\mathbf{y}/\partial\theta_k\partial\theta_m = \mathbf{b}_k'\mathbf{P}\mathbf{b}_m$ with $\mathbf{b}_k = \partial\mathbf{V}/\partial\theta_k\mathbf{P}\mathbf{y}$. Let $\mathbf{B}$ be the matrix of column vectors $\mathbf{b}_k$, and expand $\mathbf{M}$ to $\mathbf{M}_B$ by replacing $\mathbf{y}$ and $\mathbf{y}'$ in $\mathbf{M}$ by $\mathbf{B}$ and $\mathbf{B}'$, respectively. Absorbing rows 1 to $M-1$ of $\mathbf{M}_B$ then replaces elements of $\mathbf{B}'\boldsymbol{\Sigma}_\varepsilon^{-1}\mathbf{B}$ with $\mathbf{b}_k'\mathbf{P}\mathbf{b}_m$. With $\mathbf{M}$ already factored in evaluating $\log\mathscr{L}$, additional computations needed are undemanding. The AI can be used in a modified Newton-Raphson procedure to maximise $\log\mathscr{L}$, parameterising to elements of the Cholesky decompositions of $\mathbf{K}_A$ and $\mathbf{K}_R$, taking logarithms of their diagonal elements and σ_k to remove constraints on the parameters.

Bayesian estimates can be obtained using Gibbs Sampling. Location parameters $\mathbf{b}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ can be sampled sequentially, considering blocks of RR coefficients, from their fully conditional distributions. Variances are sampled from inverted Wishart (IW) or χ^2 distributions. For priors $\boldsymbol{\Sigma}_\alpha^{-1}$ and $\boldsymbol{\Sigma}_\gamma^{-1}$, matrices of sums of squares $\mathbf{S}_\alpha = \{\boldsymbol{\alpha}_k'\mathbf{A}^{-1}\boldsymbol{\alpha}_m\}$ and $\mathbf{S}_\gamma = \{\boldsymbol{\gamma}_k'\boldsymbol{\gamma}_m\}$ and degrees of freedom ν_α and ν_γ

$$(11)\quad \mathbf{K}_A \sim IW\left((\mathbf{S}_\alpha+\boldsymbol{\Sigma}_\alpha)^{-1}, N_A+\nu_\alpha\right) \qquad \text{and} \qquad \mathbf{K}_R \sim IW\left((\mathbf{S}_\gamma+\boldsymbol{\Sigma}_\gamma)^{-1}, N+\nu_\gamma\right)$$

where $\boldsymbol{\alpha}_m$ and $\boldsymbol{\gamma}_m$ are the subvectors of $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ gathering the $m-$th RR coefficients for all animals. Further details are given, for instance, by Jamrozik and Schaeffer (1997).

5. Eigenfunctions and beyond

Eigenvalues and eigenfunctions of genetic CFs provide valuable insights on how the trajectory modelled is likely to change due to selection (Kirkpatrick and Heckman, 1989). These can be estimated through an eigenvalue decomposition of the corresponding covariance matrix of RR coefficients, $\mathbf{K}_A = \mathbf{Q}\,Diag\{\lambda_i\}\mathbf{Q}'$. The eigenvector pertaining to the largest eigenvalue λ_1 gives the linear function of random regression coefficients which explains most genetic variation, and the corresponding eigenfunction shows the expected change at each age when selecting on this combination. For traits changing gradually with age, correlations between measurements at different ages are generally high, and a few eigenfunctions suffice to account for almost all genetic variance.

Reparameterisation of (4) to estimate $\boldsymbol{\alpha}^* = \mathbf{Q}'\boldsymbol{\alpha}$ directly yields estimates of genetic values for the eigenfunctions. Omitting estimation of coefficients α^* corresponding to eigenvalues close to zero then results in little loss of information, but can yield substantial computational savings. Analogous arguments apply for $\boldsymbol{\gamma}$. Similarly, the number of variance components to be estimated can be reduced by imposing rank restrictions on estimates of $\mathbf{K}_A$ or $\mathbf{K}_R$.

6. Example

REML estimates of variance components were obtained for $21,053$ weights of $N = 3,417$ beef calves, recorded at monthly intervals from birth to 280 days of age, fitting RRs on Legendre polynomials of age with $k_A = 5$ and $k_R = 6$. These were records taken prior to weaning, i.e. maternal genetic (M) and permanent environmental (C) effects had to be taken into account in addition to animals' direct effects. Hence two additional sets of RR with $k_M = k_C = 3$ were included in the model of analysis. Changes in σ_k^2 were modelled as a quadratic function of age at weighing, yielding a total of 52 parameters to be estimated. Calves were offspring of $1,023$ dams and 174 sires, and including parents without records yielded $N_A = 3,794$. Figure 1 shows estimates of the direct variance components for the ages in the data derived from estimated CFs and measurement error variances (bottom), as well as estimates of genetic correlations with contour lines from $0.95,\ldots,0.65$ in steps of 0.05.

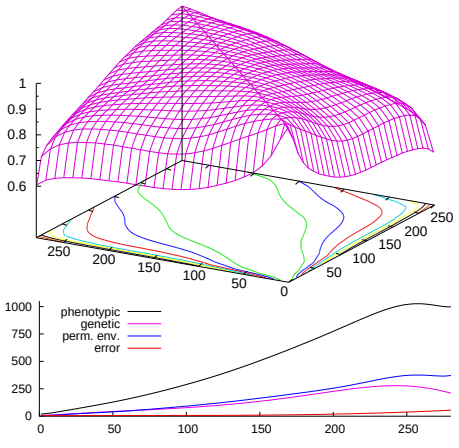

Figure 1. Genetic correlations and variance components

7. Discussion

Impetus for the uptake of RR models in animal breeding has come from the work of Kirkpatrick and co-workers (e.g. Kirkpatrick and Heckman, 1989) who introduced the concept of covariance functions to quantitative genetics. Analyses of longitudinal or similar data in other areas of applied statistics frequently assume a parametric correlation structure, with 'random coefficient' models often found to be of little advantage, requiring high orders of fit, and yielding less readily interpretable covariance functions than a parametric correlation function. In contrast, animal breeders have embraced RR models for the analysis of longitudinal data, in particular test-day records of dairy cows and growth data for pigs and beef cattle.

Quantitative genetic analyses are invariably concerned with the variation between animals, while other areas of statistics are often content with modelling within-subject covariances only. Fitting a set of additive genetic RR coefficients provides estimates of genetic merit for the whole range of ages considered, and allows ranking of animals to change with time. RR models are thus an obvious choice if we are concerned with (genetic) differences between individuals. Assuming a RR model and resulting covariance structure on a genetic level, it seems natural to apply the same model to other random effects such as permanent environmental effects. RR models account for changes in variances with time and do not require specific assumptions about the shape of the resulting CF, other than implied by the choice of covariables.

Estimates of genetic covariance matrices arising from RR model analyses can be thought of as smoothed versions of corresponding estimates from an unstructured, multivariate analysis treating records at different ages as different traits. Estimates of the eigenvalues and eigenfunctions of CFs can be obtained directly from estimates of covariances among RR coefficients. For genetic covariance functions, these statistics provide valuable insight into the effects of selection for the trait considered.

Last, but not least, RR models provide a computationally feasible way to estimate CF for large data sets with records coming in at 'all ages', as are typical for data from livestock recording schemes. Estimating coefficients of CFs as the matrix of covariances between RR coefficients requires manipulation of mixed model equations of size proportional to the number of regression coefficients to be manipulated, rather than proportional to the number of ages or even the number of records. Nevertheless, computational requirements of RR can be large, in particular for variance component estimation.

REFERENCES

Gilmour, A. R., Thompson, R. and Cullis, B. R. (1995). Average Information REML, an efficient algorithm for variance parameter estimation in linear mixed models. Biometrics 51, 1440–1450

Graser, H. U., Smith, S. P. and Tier, B. (1987). A derivative-free approach for estimating variance components in animal models by restricted maximum likelihood. Journal of Animal Science 64, 1362–1370

Jamrozik, J. and Schaeffer, L. R. (1997). Estimates of genetic parameters for a test day model with random regressions for yield traits of first lactation Holsteins. Journal of Dairy Science 80, 762–770

Kirkpatrick, M. and Heckman, N. (1989). A quantitative genetic model for growth, shape, reaction norms, and other infinite-dimensional characters. Journal of Mathematical Biology 27, 429–450

Smith, S. P. (1995). Differentiation of the Cholesky algorithm. Journal of Computational and Graphical Statistics 4, 134–147

RÉSUMÉ

On descrive des modèles de régression aléatoire pour analyse des données longitudinales en génétique animale.