MODELING ‘REPEATED’ RECORDS : COVARIANCE FUNCTIONS AND RANDOM REGRESSION MODELS TO ANALYSE ANIMAL BREEDING DATA

Karin Meyer

Institute of Cell, Animal and Population Biology, Edinburgh University, West Mains Road,
Edinburgh EH9 3JT, Scotland[1]

SUMMARY

The rôle of covariance functions in the analysis of repeated measurements data and their relationships with random regression models are reviewed. Parsimonious estimation of full or reduced rank covariance functions by Restricted Maximum Likelihood, fitting a random regression animal model, is described.

INTRODUCTION

There is a plethora of statistical literature on modeling and analysis of repeated records, often referred to as longitudinal or growth curve data. These can be analysed fitting a linear model using (restricted) maximum likelihood (Ware 1985). This framework accommodates a wide range of assumptions about the structural form of covariance matrices (e.g. Jennrich and Schluchter 1986). Until recently data from animal breeding applications have generally being analysed invoking one of the extremes of this model, namely a simple constant variance, univariate “repeatability” model or a fully parameterised, multivariate model with unstructured covariance. A specific feature of quantitative genetic analyses, not shared by other fields, is that we want to partition the variation due to individuals into its genetic and environmental components. This paper describes how an intermediate parameterisation can be achieved by fitting a *covariance function* (CF) for each source of variation, genetic and environmental, using a *random regression* (RR) model, and outlines Restricted Maximum Likelihood (REML) estimation of a parsimonious covariance structure.

COVARIANCE FUNCTION ESTIMATION

Consider measurements taken repeatedly for individuals along some continuous scale t, usually time, such as weights at various ages or test day records for milk yield at different stages of lactation.

What are covariance functions ? Literally, a CF gives the covariance between records taken at times t_i and t_j as a function of the times. A suitable class of functions is the family of orthogonal polynomials (Kirkpatrick *et al.* 1990). With potentially infinitely many records along the continuous scale t, CFs are, in essence, the ‘infinite-dimensional’ equivalent to covariance matrices. A covariance matrix for chosen times constructed from a CF has the same rank as the CF.

Random regression coefficients. Means over time or trajectories of repeated measures, e.g. growth or lactation curves, can be modeled by polynomial regressions. Regression coefficients are generally treated as fixed to account for overall trends or trends within some fixed classes. Equally, we can fit a set of random regression coefficients for each individual, to allow for individual variation in the shape of the trajectory (Laird and Ware 1982; Henderson 1982; Jamrozik *et al.* 1997). Fitting a RR model, we implicitly assume a certain covariance structure among the observations. This is determined by the covariances among the regression coefficients and can be characterised by a covariance function.

[1] on leave from : Animal Genetics and Breeding Unit, University of New England, Armidale, NSW 2351, Australia

Let y_i, the $i-$th record for an individual at time t_i, be determined by some fixed effects F, a set of RR coefficients β_m on functions $\phi_m(t_i)$ of t_i ($m = 0, \ldots, k-1$) and a measurement error ϵ_i

$$y_i = F + \sum_{m=0}^{k-1} \beta_m \phi_m(t_i) + \epsilon_i \tag{1}$$

This gives

$$Cov(y_i, y_j) = \sum_{m=0}^{k-1} \sum_{n=0}^{k-1} \phi_m(t_i)\phi_m(t_j) Cov(\beta_m, \beta_n) + Cov(\epsilon_i, \epsilon_j) = \mathcal{B}(t_i, t_j) + Cov(\epsilon_i, \epsilon_j) \tag{2}$$

where $\mathcal{B}(t_i, t_j)$ is the covariance function $\mathcal{B}$ due to $\boldsymbol{\beta}$ evaluated for times t_i and t_j. For regressions on orthogonal polynomials of t, $\mathcal{B}$ is a CF as described by Kirkpatrick *et al.* (1990), and k is the order of polynomial fit. For k equal to the number of observation, the covariance matrix among them given by $\mathcal{B}$ is unstructured, i.e. equal to that in the conventional multivariate model. Otherwise, we have a reduced order fit with less parameters and a smoothed covariance structure.

Model of analysis. To impose a structure on both genetic and environmental covariances, we need to fit corresponding sets of RR coefficients. Consider a simple animal model

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}^* \boldsymbol{\alpha} + \mathbf{Z}_D^* \boldsymbol{\gamma} + \boldsymbol{\epsilon} \tag{3}$$

with $\mathbf{y}$ the vector of N observations measured on N_D animals, $\mathbf{b}$ the vector of fixed effects, $\boldsymbol{\alpha}$ the vector of $k_A \times N_A$ additive-genetic random regression coefficients ($N_A \geq N_D$ denoting the total number of animals in the analysis, including parents without records), $\boldsymbol{\gamma}$ the vector of $k_R \times N_D$ permanent environmental random regression coefficients, and $\boldsymbol{\epsilon}$ the vector of N measurement errors. $\mathbf{X}$, $\mathbf{Z}^*$ and $\mathbf{Z}_D^*$ are the corresponding 'design' matrices, and k_A and k_R denote the order of fit for $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ and corresponding genetic and permanent environmental CF $\mathcal{A}$ and $\mathcal{R}$, respectively. The superscript '*' marks matrices with non-zero elements equal to functions $\phi_m(t_i)$. Each observation gives rise to k_A or k_R non-zero elements in $\mathbf{Z}^*$ or $\mathbf{Z}_D^*$ rather than a single element of 1 in the usual, finite-dimensional model.

REML estimation. Assume that the fixed part of (3) accounts for systematic trends so that $\boldsymbol{\alpha} \sim N(\mathbf{0}, \mathbf{K}_A \otimes \mathbf{A})$ and $\boldsymbol{\gamma} \sim N(\mathbf{0}, \mathbf{K}_R \otimes \mathbf{I}_{N_D})$, with $\mathbf{A}$ the numerator relationship between animals, $\mathbf{I}_{N_D}$ an identity matrix of size N_D and $\otimes$ denoting the direct matrix product. The matrices of covariances between regression coefficients, $\mathbf{K}_A$ and $\mathbf{K}_R$, are equal to the coefficient matrices of the corresponding CFs, $\mathcal{A}$ and $\mathcal{R}$. Let $Cov(\boldsymbol{\alpha}, \boldsymbol{\gamma}') = \mathbf{0}$ and, for generality, let $V(\boldsymbol{\epsilon}) = \mathbf{R}$. This gives log likelihood ($\mathcal{L}$)

$$\log \mathcal{L} = -\frac{1}{2} \left(N_A \log |\mathbf{K}_A| + k_A \log |\mathbf{A}| + N_D \log |\mathbf{K}_R| + \log |\mathbf{R}| + \log |\mathbf{C}^*| + \mathbf{y}'\mathbf{P}^*\mathbf{y} \right) \tag{4}$$

The log determinant of the coefficient matrix $\mathbf{C}^*$ and the residual sums of squares $\mathbf{y}'\mathbf{P}^*\mathbf{y}$ can be obtained by factoring the mixed model matrix pertaining to (3)

$$\mathbf{M}^* = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}^* & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}_D^* & \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}^{*'}\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^{*'}\mathbf{R}^{-1}\mathbf{Z}^* + \mathbf{K}_A^{-1} \otimes \mathbf{A}^{-1} & \mathbf{Z}^{*'}\mathbf{R}^{-1}\mathbf{Z}_D^* & \mathbf{Z}^{*'}\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}_D^{*'}\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}_D^{*'}\mathbf{R}^{-1}\mathbf{Z}^* & \mathbf{Z}_D^{*'}\mathbf{R}^{-1}\mathbf{Z}_D^* + \mathbf{K}_R^{-1} \otimes \mathbf{I}_{N_D} & \mathbf{Z}_D^{*'}\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{y}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{y}'\mathbf{R}^{-1}\mathbf{Z}^* & \mathbf{y}'\mathbf{R}^{-1}\mathbf{Z}_D^* & \mathbf{y}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix} \tag{5}$$

$\mathbf{M}^*$ has $N_F + k_A N_A + k_R N_D + 1$ rows and columns (with N_F the total number of fixed effects fitted), i.e., its size and thus computational requirements are proportional to the order of fit of CFs.

Measurement error variances. Usually **R** can be described by few parameters, measurement errors often being assumed uncorrelated. In the simplest case, $\mathbf{R} = \sigma^2_\epsilon \mathbf{I}_N$, $\log|\mathbf{R}| = N \log \sigma^2_\epsilon$, and σ^2_ϵ can be factored from (5), so that $\mathbf{M}^*$ can be set up as for a univariate analysis, and σ^2_ϵ can be estimated directly as $\sigma^2_\epsilon = \mathbf{y}'\mathbf{P}^*\mathbf{y}/(N - r(\mathbf{X}))$. Other assumptions might involve heterogeneous variances, i.e. **R** diagonal, or errors following an auto-regressive or moving average process.

Maximising the likelihood. Estimates of the distinct elements of $\mathbf{K}_A$ and $\mathbf{K}_R$ and the parameters determining **R** can be obtained maximising (3), using existing REML algorithms for the estimation of covariance components. This may involve a simple derivative-free search (Meyer 1991) or, more efficiently, a method utilising derivatives of log $\mathcal{L}$ such as the 'average information' algorithm (Johnson and Thompson 1995). The minimum order(s) of fit required to model the data adequately can be determined using a likelihood ratio test (LRT). This encourages an upwards strategy, increasing the order(s) of fit stepwise until no significant increase in likelihood is achieved.

Reduced rank covariance functions. Fitting a CF to order k requires $k(k+1)/2$ elements of the corresponding covariance matrix among RR coefficients, **K**, to be estimated. By nature, polynomial regression coefficients are highly correlated, i.e. **K** is likely to have m dominating eigenvalues with the remainder, $k - m$ close to zero. It implies that most variation is in the directions given by the eigenvectors corresponding to the large eigenvalues. Thus little information is lost by ignoring the others, i.e. fixing them at zero (or a small positive value as, strictly speaking, (5) is not defined for indefinite **K**). This not only reduces the number of parameters to be estimated but also alleviates convergence problems frequently encountered at the bounds of the parameter space. It appears especially advantageous where a high order of fit k is required to model the shape of the trajectory adequately, but a subset of m directions suffices.

Reparameterisation. Consider the Cholesky decomposition of **K**, pivoting on the largest diagonal

$$\mathbf{K} = \mathbf{L}\mathbf{D}\mathbf{L}' = \sum_{i=1}^{k} d_i \, \mathbf{l}_i \, \mathbf{l}_i' \tag{6}$$

where **L** is a lower diagonal matrix with diagonal elements of unity, $\mathbf{l}_i$ the $i-$th column vector of **L**, and **D** is a diagonal matrix with elements d_i. A reparameterisation to the non-zero off-diagonal elements of **L** and the elements of **D** has been advocated to remove constraints on the parameter space or to improve convergence rates in an iterative (RE)ML estimation scheme (Lindstrom and Bates 1988, Groeneveld 1994, Meyer and Smith 1996). Moreover the elements of **D** are the eigenvalues of **K**. Thus, assuming descending order, $d_i > d_{i+1}$, we can estimate

$$\mathbf{K}^+ = \sum_{i=1}^{m} d_i \, \mathbf{l}_i \, \mathbf{l}_i' \tag{7}$$

which has rank m and is described by $km - m(m-1)/2$ parameters on the Cholesky scale. A LRT can be used to determine the minimum rank of $\mathbf{K}^+$.

Breeding value estimation. Estimates of $\boldsymbol{\alpha}$ in the RR model take the place of estimated breeding values in the finite, multivariate model. In some instances, the shape parameters of the trajectory or functions thereof have an interpretation in their own right and are used to rank animals. In other cases functions of the trajectory are of interest, for example the first derivative of a growth curve gives an estimate of growth rate, and integrating over the lactation curve estimates lactation yield.

As CFs give a continuous description of the covariance structure over the time interval considered, RR coefficients provide a continuous estimate of the additive genetic merit for each animal, and conventional breeding value estimates at selected times are readily derived. The RR/CF approach provides the flexibility to treat measurements along a trajectory as different traits and, at the same time, facilitates efficient estimation through a reduction in dimensionality to the order of fit (or rank) of the CF. This can be exploited in a transformation to canonical scale as described by Van der Werf *et al.* (1997).

Selection response. The equivalent to the eigenvalue decomposition of a covariance matrix for a CF is given by its eigenvalues (λ_i) and eigenfunctions ($\boldsymbol{\zeta}_i$) (Kirkpatrick *et al.* 1990). These are estimated as the eigenvalues of the corresponding coefficient matrix **K** and as a function of the eigenvectors $\boldsymbol{\nu}_i$ of **K** with elements ν_{ij}

$$\boldsymbol{\zeta}_i = \sum_{j=0}^{k-1} \nu_{ij}\,\phi_j(t) \tag{8}$$

Note that $\phi_j(t)$ in (8) is not evaluated for any particular time, i.e. $\boldsymbol{\zeta}_i$ is a continuous (polynomial) function of t. Any change in the mean trajectory can be expressed as a weighted sum of the eigenfunctions with the rate of change determined by the eigenvalues. A decomposition of the genetic CF $\mathcal{A}$ thus provides valuable insight in the potential response to selection along the complete trajectory - eigenfunctions of $\mathcal{A}$ representing possible deformations and the corresponding eigenvalues quantifying the amount of genetic variation available in each direction (Kirkpatrick *et al.* 1990).

CONCLUSIONS

Repeated measurements for traits gradually changing with time can be modeled using RR coefficients. Regressing on orthogonal polynomials of time does not require any prior assumptions about the shape of the trajectory to be modeled, and the resulting covariance structure can be described by a CF as proposed by Kirkpatrick *et al.* (1990). Conversely, genetic and environmental CF can be estimated by REML fitting RR coefficients for each source of variation. A parsimonious model can be achieved by fitting reduced rank CF. Minimum rank and order of fit can be determined using a likelihood ratio test.

REFERENCES

Groeneveld, E. 1994. *Genet. Sel. Evol.* **26** : 537–545.
Henderson, C.R. Jr. 1982. *Biometrics* **38** : 623–640.
Jamrozik, J., Schaeffer, L.R. and Dekkers, J.C.M. 1997. *J. Dairy Sci.* **80** : 1217–1226.
Jennrich, R.I. and Schluchter, M.D. 1986. *Biometrics* **42** : 805–820.
Johnson, D.L. and Thompson, R. 1995. *J. Dairy Sci.* **78** : 449–456.
Kirkpatrick, M., Lofsvold, D. and Bulmer, M. 1990. *Genetics* **124** : 979-993.
Laird, N.M. and Ware, J.H. 1982. *Biometrics* **38** : 963–974.
Lindstrom, M.J. and Bates, D.M. 1988. *J. Amer. Stat. Ass.* **83** : 1014–1022.
Meyer, K. 1991. *Genet. Sel. Evol.* **23** : 67–83.
Meyer, K. and Smith, S.P. 1996. *Genet. Sel. Evol.* **28** : 23–49.
Van der Werf, J., Goddard, M. and Meyer, K. 1997. *J. Dairy Sci.* : (submitted)
Ware, J. H. 1985. *Amer. Stat.* **39** : 95–101.