ESTIMATION OF GENETIC AND PHENOTYPIC COVARIANCE FUNCTIONS FOR LONGITUDINAL DATA

K. Meyer

Animal Genetics and Breeding Unit, University of New England, Armidale, NSW 2351

SUMMARY

Covariance functions are the 'infinite-dimensional' equivalents to covariance matrices for longitudinal data, i.e. many, 'repeated' records per individual taken over a period of time. Their properties are reviewed and illustrated with a numerical example. Restricted Maximum Likelihood estimation of genetic and phenotypic covariance functions fitting an animal model is described.

Keywords : Covariance functions, longitudinal data, genetic parameters, REML

INTRODUCTION

Biological characteristics such as body size or growth are often recorded at various times (ages), resulting in many, typically highly correlated measurements per individual. In some cases, these are repeated records of the same trait, but the assumption of a univariate ('repeatability') model is often clearly invalid, while a 'full' multivariate model with the number of traits equal to the number of ages would be highly overparameterised. This paper outlines how a 'reduced' multivariate model, fitting the least number of traits required to describe the data adequately, can be identified using the *covariance function* model of Kirkpatrick and Heckman (1989).

COVARIANCE FUNCTIONS

What is a covariance function? In essence, a covariance function (CF) is merely the 'infinite-dimensional' equivalent to a covariance matrix for records taken at a number of ages. It gives the covariance between any two records as a function of the ages at measurement. A suitable family of functions to describe CF are orthogonal polynomials. This applies to any type of covariance matrix, genetic, environmental or phenotypic.

Properties of Covariance functions Kirkpatrick and Heckman (1989) list three advantages of the CF model over the traditional, 'finite-dimensional' approach. Firstly, CFs produce a description for every point along the continuous (time) scale of measurement. This allows for easy interpolation to obtain covariances for ages not recorded. No prior assumptions about the shape of the curve (growth or equivalent) are required. Each record is used at its actual age making corrections for age superfluous, when data are recorded 'at all ages'. Secondly, CFs allow a more accurate prediction of response to selection. Each CF has a set of associated eigenvalues and eigenfunctions (infinite-dimensional analogues to eigenvectors) which provide valuable information about the directions in which mean curves are likely to change most rapidly under selection. Thirdly, the CF model makes more efficient use of the data. Fitting polynomials only to the minimum order required ensures that no unnecessary parameters are estimated, thus minimising sampling variation.

Full order fit. Let Σ denote the covariance matrix for observations at t ages, and Φ the matrix of orthogonal polynomial functions evaluated at the given ages with elements $\phi_{ij} = \phi_j(a_i)$, the $j-$th polynomial evaluated for age i. The covariance between records taken at ages l and m is then

$$\mathcal{S}(a_l, a_m) = \sum_{i=0}^{k-1}\sum_{j=0}^{k-1} \phi_i(a_l)\phi_j(a_m)K_{ij} = \sum_{i=0}^{k-1}\sum_{j=0}^{k-1} s_{ij}a_l^i a_m^j \qquad (1)$$

where $\mathcal{S}$ is the CF, k is the order of fit, $\mathbf{K}$ with elements K_{ij} is the matrix of coefficients of the CF, a_m is the $m-$th age, standardised to the interval for which the polynomials are defined, and $\mathbf{S}$ with elements s_{ij} is $\mathbf{K}$,

pre- and postmultiplied with the matrix of coefficients of the orthogonal polynomials. Kirkpatrick *et al.* (1990) use the so-called Legendre polynomials which span the interval from -1 to 1. Note that (1) includes a scalar term, i.e., that an order of fit of k includes functions of ages to the power 0 to $k-1$. Assuming a full-order polynomial fit ($k=t$), (1) gives $\mathbf{\Sigma} = \mathbf{\Phi K \Phi}'$, i.e $\mathbf{K}$ can be estimated as $\mathbf{K} = \mathbf{\Phi}^{-1}\mathbf{\Sigma}(\mathbf{\Phi}^{-1})'$.

$$\mathbf{\Sigma} = \begin{bmatrix} 436.0 & 522.3 & 424.2 \\ 522.3 & 808.0 & 664.7 \\ 424.2 & 664.7 & 558.0 \end{bmatrix}$$

$$\mathbf{\Phi} = \begin{bmatrix} \phi_0(-1) & \phi_1(-1) & \phi_2(-1) \\ \phi_0(0) & \phi_1(0) & \phi_2(0) \\ \phi_0(1) & \phi_1(1) & \phi_2(1) \end{bmatrix} = \begin{bmatrix} 0.707 & -1.225 & 1.581 \\ 0.707 & 0 & -0.791 \\ 0.707 & 1.225 & 1.581 \end{bmatrix}$$

$$\mathbf{K} = \begin{bmatrix} 1348.0 & 66.5 & -112.0 \\ 66.5 & 24.3 & -14.0 \\ -112.0 & -14.0 & 14.5 \end{bmatrix}$$

Example. Kirkpatrick *et al.* (1990) consider body weights of mice at ages 2, 3 and 4 weeks, which are $-1, 0$ and 1 on the standardised scale. The covariance original matrix $\mathbf{\Sigma}$, $\mathbf{\Phi}$ evaluated for $k = 0, 1, 2$ and $\mathbf{K}$ are shown on the left. This gives CF $\mathcal{S}(a_i, a_j) = 808.0 + 71.2(a_i + a_j) + 36.4a_ia_j - 40.7(a_i^2 a_j + a_i a_j^2) - 215.0(a_i^2 + a_j^2) + 81.6a_i^2 a_j^2$. Assume we want to determine the covariance between weights at 3 and 3.5 weeks of age. This gives $a_i = 0$ and $a_j = 0.5$ (standardised scale) and covariance $808.0 + 71.2 \times 0.5 - 215.0 \times 0.5^2 = 789.9$. Similarly, $\mathcal{S}$ gives the variance at 3.5 weeks as 775.7.

Reduced order fit. For a reduced order ($k < t$) fit, $\mathbf{\Phi}$ has only k columns and, correspondingly, the number coefficients to be estimated is reduced to $k(k+1)/2$. As $\mathbf{\Phi}$ is then rectangular and does not have an inverse, Kirkpatrick *et al.* (1990) use a weighted least-squares procedure to estimate $\mathbf{K}$ in this case. Once a reduced fit matrix of coefficients has been estimated, it can be used to obtain a modified covariance matrix, $\mathbf{\Sigma}^*$, among the t observations, using (1). Reducing the order of fit by one has a similar effect to setting an eigenvalue to zero, i.e., it reduces the rank of the matrix by one. In contrast to a canonical decomposition, however, the CF approach explicitly accounts for the ordering of records and spacing of ages. For $k = 1$, all (co)variances are equal, which implies that all correlations are unity.

$$\mathbf{\Sigma}^* = \begin{bmatrix} 360.5 & 324.1 & 287.7 \\ 324.1 & 312.2 & 300.3 \\ 287.7 & 300.3 & 312.9 \end{bmatrix}$$

Example. For $k = 2$, Kirkpatrick *et al.* (1990) obtained an estimated CF $\mathcal{S}(a_i, a_j) = 312.2 + 11.9(a_i + a_j) + 24.5a_ia_j$. This gives $\mathbf{\Sigma}^*$ of rank 2 as shown on the left.

Measurement errors. Under the 'finite' model and with single records per age, we usually cannot disentangle permanent and temporary environmental effects and their (co)variances. This can be done indirectly, however, using the CF model. Kirkpatrick *et al.* (1994) describe how to correct for the bias in the diagonal elements of the estimated residual (or phenotypic) covariance matrix due to measurement errors, i.e. temporary environmental effects, by extrapolating to the diagonals, after the coefficients of the CF have been estimated using only the off-diagonals of the estimated covariance matrix. This implies that the maximum order of fit for the CF is $t-1$ rather than t.

REML ESTIMATION

Model of analysis. Consider a simple animal model with single measurements at t ages available for all N individuals.

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{r} + \varepsilon \tag{2}$$

with $\mathbf{y}$, $\mathbf{b}$, $\mathbf{a}$, $\mathbf{r}$ and ε the vectors of observations, fixed effects, animals additive genetic effects, permanent environmental effects, and measurement errors, respectively, and $\mathbf{X}$ and $\mathbf{Z}$ the corresponding incidence matrices. Assume a multivariate normal distribution with $V(\mathbf{a}) = \Sigma_A \times \mathbf{A}$, $V(\mathbf{r}) = \Sigma_R \times \mathbf{I}_N$, $V(\varepsilon) = \Sigma_\varepsilon \times \mathbf{I}_N$ and

zero covariances between **a**, **r** and ε. Here $\Sigma_A = \{\sigma_{A_{ij}}\}$, $\Sigma_R = \{\sigma_{R_{ij}}\}$ and Σ_ε= Diag$\{\ \sigma^2_{\varepsilon_i}\}$ denote the $t \times t$ matrices of additive genetic, permanent and temporary environmental covariances between measurements, **A** is the numerator relationship matrix, $\mathbf{I}_N$ is an identity matrix of size N, and $'\times'$ is the direct matrix product. The REML (log) likelihood ($\mathcal{L}$) is then (Meyer, 1991)

$$\mathcal{L} = -\frac{1}{2}\left[const + N\ln|\Sigma_R + \Sigma_\varepsilon| + N_A \ln|\Sigma_A| + t\ln|\mathbf{A}| + \ln|\mathbf{C}| + \mathbf{y}'\mathbf{P}\mathbf{y}\right] \tag{3}$$

where N_A is the total number of animals in the analysis, including any parents without records, **C** is the coefficient matrix in the mixed model equations (MME) pertaining to (2) (or a full rank submatrix thereof), and $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}$ with **V**=Var(**y**).

Reparameterisation. As shown by Meyer and Hill (1996), the multivariate, "finite-dimensional" REML analysis can be adapted to the estimation of CFs through simple reparameterisation. Let $\mathcal{A}$ and $\mathcal{R}$ denote the covariance functions of additive genetic and permanent environmental effects with coefficients matrices $\mathbf{K}_A$ and $\mathbf{K}_R$, respectively. As for Kirkpatrick *et al.*'s (1994) least-squares procedure, the maximum order of fit for $\mathcal{R}$ is $t-1$ rather than t; fitting $\mathcal{R}$ to the order $t-1$ together with t independent measurement errors is equivalent to a full order fit for environmental effects. Rewriting $\Sigma_A = \Phi_A\mathbf{K}_A\Phi'_A$ and $\Sigma_R = \Phi_R\mathbf{K}_R\Phi'_R$, (3) becomes a function of the coefficient matrices of the CF

$$\begin{aligned}\mathcal{L} \quad = \quad & -\frac{1}{2}\left[const + N\ln\left|\Phi_\mathbf{R}\mathbf{K}_R\Phi'_\mathbf{R} + \mathrm{Diag}\{\sigma^2_{\varepsilon_\mathbf{i}}\}\right| + N_A\ln|\mathbf{K}_A| + \ln|\mathbf{C}| + \mathbf{y}'\mathbf{P}\mathbf{y}\right. \\ & \left. + N_A\ln|\Phi_\mathbf{A}\Phi'_\mathbf{A}| + t\ln|\mathbf{A}|\,\right]\end{aligned} \tag{4}$$

This accommodates both a full and reduced order fit. Moreover, polynomials of different order, $k_A \le t$ and $k_R < t$, can be fitted for $\mathcal{A}$ and $\mathcal{R}$, respectively. REML estimates of the distinct elements of $\mathbf{K}_A$ and $\mathbf{K}_R$ and the measurement errors ($(k_A(k_A+1) + k_R(k_R+1))/2 + t$) parameters, at least $t+2$ and at most $t(t+1)$) can be obtained using a suitable optimisation procedure to locate the maximum of $\log\mathcal{L}$. This can be done forcing estimates of variances $\sigma^2_{\varepsilon_i}$ to be positive and of matrices **K** to be (semi) positive definite, thus guaranteeing estimated CFs to be (semi) positive definite. A likelihood ratio test can be used to determine the minimum order of fit.

Example. Data for 6 ages (equally spaced) were simulated for CFs of order 3. Figure 1 shows the covariance matrices among the 6 ages, 'reconstructed' from REML estimates of the genetic CF fitted to orders 2, 3 and 6, respectively. For $k = 2$, components lie on a tilted plane. This changes into a quadratic surface for $k = 3$. With 3 representing the true number of 'traits', surfaces for $k = 4, 5$ (not shown) and 6 are little different, except for more 'wiggles' depicting increased sampling variation.

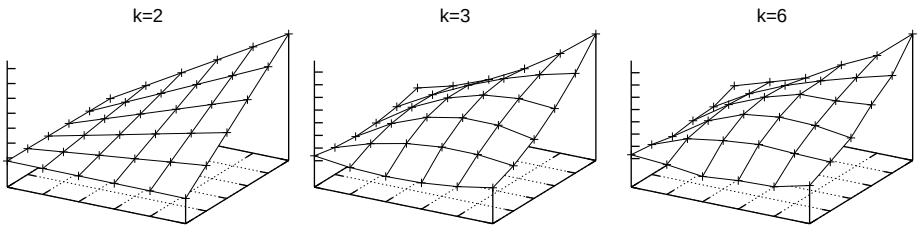

Figure 1. Covariance matrices for different orders of fit.

Extension to other models. While described above for the simplest case of a basic animal model without missing records, the procedure is readily applicable to more general cases, e.g. models including additional random effects and data with few ages recorded for each individual. Other assumptions about the structure of Σ_ϵ and multivariate CFs can be accommodated (Meyer and Hill, 1996).

DISCUSSION

The covariance function model provides a useful alternative to the analyses of repeated records used to date. CFs enable us to model our data with the least number of parameters necessary, avoiding problems associated with overparameterised models. CFs do not require any *a priori* assumptions about the number of different 'traits' represented by a series of measurements or the shape of trends. Eigenvalues and -functions of CFs have an interpretation of their own. On the genetic level, they give the directions in which mean growth trajectories are likely to change under selection. Potentially these could be used to characterise differences between breeds for sequentially measured 'traits' such as growth. CFs can readily be estimated by REML. This involves only a simple reparameterisation of existing procedures to estimate covariance components.

Random regressions. An equivalent model to the CF model is a random regression model with covariables equal to orthogonal polynomials of age at recording. This implies fitting k_A genetic and k_R permanent environmental random regression coefficients for each animal. These replace the respective random effects (t each) in (2). The covariances between the regression coefficients are equal to the coefficients of the CFs (elements K_{ij}), and can be estimated in a corresponding REML analysis. Moreover, the k_A genetic regression coefficients fully describe the generit merit of an animal for the 'trait(s)' recorded over a period of time, i.e. this approach can reduce the number of breeding values (EBVs) to be estimated and thus simplify selection decision, especially for 'traits' like growth where a range of EBVs for points along the growth curve (like birth, weaning, yearling, final and mature weight in beef cattle) is replaced by estimates of genetic regression coefficients describing the complete growth curve.

REFERENCES

Kirkpatrick, M. and Heckman, N. (1989) *J. Math. Biol.* **27** : 429.
Kirkpatrick, M., Lofsvold, D. and Bulmer, M. (1990) *Genetics* **124** : 979.
Kirkpatrick, M., Hill, W.G. and Thompson, R. (1994) *Genet. Res.* **64** : 57.
Meyer, K. (1991) *Genet. Select. Evol.* **23** : 67.
Meyer, K. and Hill, W.G. (1996) *Livest. Prod. Sci.* : (in press).