

Estimation of genetic and phenotypic covariance functions for longitudinal or ‘repeated’ records by Restricted Maximum Likelihood

Karin Meyer¹ and William G. Hill

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

Abstract

Covariance functions are the equivalent of covariance matrices for traits with many, potentially infinitely many, records in which the covariances are defined as a function of age or time. They can be fitted for any source of variation, e.g. genetic, permanent environment or phenotypic. A suitable family of functions for covariance functions are orthogonal polynomials. These give the covariance between measurements at any two ages as a higher order polynomial of the ages at recording. Polynomials can be fitted to full or reduced order. The former is equivalent to a multivariate analysis estimating covariance components. A reduced order fit involves less parameters and smoothes out differences in estimates of covariances. It gives predicted covariance matrices of rank equal to the order of fit.

The coefficients of covariance functions can be estimated by Restricted Maximum Likelihood through a reparameterisation of existing algorithms to estimate covariance components. For a simple animal model with equal design matrices for all traits, computational requirements to estimate covariance functions are proportional to the order of fit for the genetic covariance function. Applications to simulated data and a set of beef cattle data are shown.

Keywords : Genetic parameters, covariance functions, repeated records, Restricted Maximum Likelihood

Introduction

Often biological characteristics such as body size or growth are measured on the same individual(s) at various times or ages. Such records are commonly referred to as longitudinal data. Potentially, there are infinitely many measurements per individual, and these typically are highly correlated. In some cases, they are treated as repeated records of the same trait, but more generally measurements at different ages are considered to represent different traits.

In many instances, the assumption of a univariate (‘repeatability’) model clearly does not hold whilst, on the other hand, a ‘full’ multivariate model with the number of traits equal to the number of ages (or equivalent) would result in a highly overparameterised analysis. This would be likely to impose unnecessary computational demands, give rise to problems associated with estimation at the bounds of the parameter space, and yield estimates with greatly increased sampling errors, especially in the context of variance component estimation. Hence the model which fits the least number of traits and describes the data adequately needs to be determined.

This paper reviews how a ‘reduced’ multivariate model can be identified using the *covariance function* model of Kirkpatrick and Heckman (1989), shows how it can be fitted in a Restricted Maximum Likelihood (REML) estimation framework, and demonstrates its application for simulated data and weight records of beef cattle.

Modelling ‘repeated’ records

‘Repeated’ measures are common to a wide range of applied statistics, for instance the analysis of growth curves, time series and spatial variation. Different approaches, terminology and models applied to such data are reviewed in detail by Lindsey (1993), who also gives an extensive bibliography.

Consider N individuals and t potential measurements. Without any assumptions about the structure of the (phenotypic) covariance of the t observations, i.e., treating each record as a separate trait, it has $t(t+1)/2$ parameters. To fit such an *unstructured multivariate* model in practical applications which require variance components to be estimated, t usually has to be small and N to be comparatively large for analyses to be feasible and estimates to be sufficiently accurate.

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

In other cases, the number of parameters fitted is reduced by incorporating a specific *covariance structure*. A typical example is the analysis of time series. For a stationary series, for instance, the covariance between two measurements is assumed to depend on the time lag between them via the so-called auto-covariance function. This can reduce the number of parameters considerably, especially for equally spaced observations. In some cases, a further reduction can be achieved by adopting a parametric model for the auto-covariance function, such as an exponential function (Diggle, 1990). In the animal breeding context, Wade *et al.* (1993), for instance, considered the application of an auto-regressive function to dairy cattle data.

In general, the objective is to fit a model with the least number of parameters which adequately describes the data. In the simplest case, all records are assumed to be measurements of the same trait ($t = 1$). In the animal breeding context, this is the so-called repeatability model. It implies genetic correlations of unity, i.e. that all genetic variances and covariances are of the same magnitude. Similarly, all phenotypic variances are considered identical, and phenotypic correlations and covariances among all measurements are assumed to be the same. Many of the models fitted when analysing repeated records invoke this model, often teamed with some correction(s) for trends (which in turn create a covariance structure among the corrected records). For example, Ptak and Schaeffer (1993) treat daily production of dairy cows throughout lactation as the same trait, adjusting for the shape of the lactation curve by fitting higher order regressions on stage of lactation and functions thereof.

Both imposing a structure on the covariance matrices and adjusting for differences in time or age at recording (or any other meta-meter, e.g. distance) require prior assumptions, e.g. about the number of different traits represented by the repeated measurements or the shape of time trends. In some instances, however, we do not know the forms of relationships between measurements involved or we are not prepared to decide on the number of traits *a priori*.

A common procedure to identify the number of independent combinations among the t records is the use of an eigenvalue or a canonical decomposition (e.g. Graybill, 1969). Hayes and Hill (1980) applied this to genetic and phenotypic covariance matrices. The latter procedure can be thought of as a linear transformation of the original measures to new variables which are genetically and phenotypically uncorrelated, have unit phenotypic variance and heritabilities equal to the resulting eigenvalues. For highly correlated measures, there are typically a number of eigenvalues close to zero, i.e., the genetic information supplied by the t measurements is almost entirely contained by the k linear combinations corresponding to the k largest eigenvalues. In other words, we can represent the t ‘repeated’ records by k ‘traits’. This approach has been used, for instance, by Wiggans *et al.* (1996) to identify 5 ‘canonical traits’ describing dairy production for milk, fat and protein yields at 10 individual days of lactation and thus to reduce the number of traits in a multivariate genetic evaluation model dramatically (from 30 to 5).

Application of the linear transformation(s) to create the reduced number of new variables is equivalent to assuming a genetic covariance matrix (of order t) of reduced rank (k). This can be obtained simply by setting the small eigenvalues identified equal to zero and pre- and postmultiplying the resulting diagonal matrix with the matrix of corresponding eigenvectors and its transpose. Note that this approach is invariant to the ordering of records (ages). For $k = 1$, all genetic correlations are assumed to be unity but differences in variability of the t records are preserved.

Covariance functions

An alternative, based on fitting a set of k orthogonal functions to a given covariance matrix (of order t) has been described recently by Kirkpatrick *et al.* (1990 and 1994) in the context of estimating covariance functions for (potentially) infinite-dimensional characters. In essence, a covariance function (CF) is merely the infinite-dimensional equivalent to a covariance matrix for a given number of records taken at different ages. It gives the covariance between any two records measured at given ages as a function of the ages and some coefficients. Theoretically, there are infinitely many coefficients, but in practice a limited number (up to $t(t + 1)/2$ for t ages) is estimated to provide an estimate of the CF. A suitable family of functions to describe CF are orthogonal polynomials. This applies to any type of covariance matrix, genetic, environmental or phenotypic. Like the corresponding covariance matrices, CFs are additive, i.e., assuming random effects are uncorrelated, the phenotypic CF can be estimated as the sum of its causal CFs (Kirkpatrick and Heckman, 1989).

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

$$\mathcal{T}(a_l, a_m) = \sum_{i=0}^{k-1} \sum_{j=0}^{k-1} \phi_i(a_l) \phi_j(a_m) K_{ij} = \sum_{i=0}^{k-1} \sum_{j=0}^{k-1} \tau_{ij} a_l^i a_m^j \quad (1)$$

where \mathcal{T} with factors τ_{ij} is the CF, k is the order of fit, \mathbf{K} with elements K_{ij} is the matrix of coefficients of the CF and a_m is the m -th age, standardised to the interval for which the polynomials are defined. Kirkpatrick *et al.* (1990 and 1994) use the so-called Legendre polynomials (see Abramowitz and Stegun, 1965; p.798) 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 ($k = t$), the observed covariance matrix can be rewritten as

$$\Sigma = \Phi \mathbf{K} \Phi' \quad (2)$$

i.e. \mathbf{K} can be estimated as

$$\mathbf{K} = \Phi^{-1} \Sigma (\Phi^{-1})' \quad (3)$$

(Kirkpatrick *et al.*, 1990). For a reduced order ($k < t$) fit, Φ has only k columns and, correspondingly, the number coefficients to be estimated is reduced to $k(k+1)/2$. As Φ is then rectangular and does not have an inverse, Kirkpatrick *et al.* (1990 and 1994) suggest a weighted least-squares procedure to estimate \mathbf{K} in this case. The authors give a step-by-step procedural guide and a detailed worked example.

Once a reduced fit matrix of coefficients has been estimated, it can be used to obtain a modified covariance matrix, Σ^* , among the t observations, using (1). Furthermore, $\hat{\mathcal{T}}$ can be used to interpolate, i.e. calculate the covariance for any two ages in the range for which it has been estimated, including those for which we do not have records.

Kirkpatrick *et al.* (1990) illustrate this for the additive genetic covariance matrix of body weights in mice at 2, 3 and 4 weeks of age reported by Riska *et al.* (1984),

$$\Sigma = \begin{bmatrix} 436.0 & 522.3 & 424.2 \\ 522.3 & 808.0 & 664.7 \\ 424.2 & 664.7 & 558.0 \end{bmatrix} \quad (4)$$

On the standardised scale, these ages are $-1, 0$ and 1 . This gives the matrix of Legendre polynomials evaluated for $k = 0, 1, 2$

$$\Phi = \begin{bmatrix} \phi_0(-1) & \phi_1(-1) & \phi_2(-1) \\ \phi_0(0) & \phi_1(0) & \phi_2(0) \\ \phi_0(1) & \phi_1(1) & \phi_2(1) \end{bmatrix} = \begin{bmatrix} 0.707 & -1.225 & 1.581 \\ 0.707 & 0 & -0.791 \\ 0.707 & 1.225 & 1.581 \end{bmatrix} \quad (5)$$

From (3), the estimated coefficient matrix is

$$\mathbf{K} = \begin{bmatrix} 1348.0 & 66.5 & -112.0 \\ 66.5 & 24.3 & -14.0 \\ -112.0 & -14.0 & 14.5 \end{bmatrix} \quad (6)$$

and the corresponding covariance function is $\mathcal{S}(a_i, a_j) = 808.0 + 71.2(a_i + a_j) + 36.4a_i a_j - 40.7(a_i^2 a_j + a_i a_j^2) - 215.0(a_i^2 + a_j^2) + 81.6a_i^2 a_j^2$ (Kirkpatrick *et al.*, 1990). Assume now, we want to obtain the covariance between weights at 3 and 3.5 weeks of age. On the standardised scale this is equal to $a_i = 0$ and $a_j = 0.5$, and the covariance is $808.0 + 71.2 \times 0.5 - 215.0 \times 0.5^2 = 789.9$. Similarly, \mathcal{S} gives the variance at 3.5 weeks as 775.7.

The scheme outlined above is Kirkpatrick *et al.*'s (1990) method of symmetric coefficients. Kirkpatrick *et al.* (1994) describe an asymmetric coefficients approach, the main difference in procedure being that only the elements of \mathbf{K} above and including the anti-diagonal are assumed to be non-zero and estimated, which leads to somewhat different properties of the estimated CF. The authors argue that it is often better behaved than the symmetric approach, resulting in less 'wiggly' functions by eliminating the product of two $(k-1)$ -th order polynomials.

For the symmetric \mathbf{K} , 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 (but this does not hold in the asymmetric approach). In contrast to the 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.

Kirkpatrick and Heckman (1989) list three advantages of the CF model over the ‘traditional’ multivariate, ‘finite-dimensional’ approach.

Firstly, it produces a description at every point along the continuous (time) scale of measurement. This allows for easy interpolation between the ages at which recording took place. Often we are interested in genetic parameter estimates at certain target ages while the data available spans a range of ages at recording. Using the CF approach, each record can be used at its actual age rather than having to correct for age differences, e.g. by fitting age as a covariable. As emphasized above, no restrictions on the form of the growth curve (or equivalent) are required for this, and we can obtain estimates for covariances for ages for which there are no observations.

Secondly, CFs allow a more accurate prediction of response to selection. Each CF has a set of associated eigenvalues and *eigenfunctions*, the latter being the infinite-dimensional analogues to the eigenvectors of a covariance matrix. These provide valuable information on the directions in which mean growth curves (or equivalent) are likely to change most rapidly under selection pressure because they exhibit most genetic variation. Moreover, the CF model allows the estimation of a continuous selection gradient function which describes the change in means due to a generation of selection; see Kirkpatrick and Heckman (1989) and Kirkpatrick *et al.* (1990) for further details.

Finally, the infinite-dimensional approach is likely to make more efficient use of the data. While the eigenvalues and -vectors of estimated covariances are expected to asymptote to the eigenvalues and -functions of the corresponding CF as more and more discrete ages (traits) are included in the analysis, Kirkpatrick and Heckman (1989) demonstrated in a simulation study quicker convergence for the CF model. Fitting polynomials only to the minimum order required to describe the data adequately, ensures that no unnecessary parameters are estimated, thus minimising sampling errors and reducing strong negative sampling correlations. Kirkpatrick *et al.* (1990) describe χ^2 test procedures to establish whether a reduced rank covariance matrix (derived from a reduced order fit CF) is consistent with the observed covariance matrix, and to test the hypothesis that one or more of the eigenvalues of the estimated CF are zero.

REML estimation

As discussed so far, both the canonical decomposition and the CF model required that estimates of the covariance matrices among records at the t observed ages were available. In practice, it would be preferable to estimate reduced rank covariance matrices directly from the data. Moreover, it would be desirable to do so sequentially, stopping when the data have been modelled adequately with the least number of parameters. This can be done readily within a maximum likelihood framework.

The canonical decomposition has been used in REML algorithms to estimate covariance matrices when design matrices were equal for all traits in order to reduce computational requirements, effectively carrying out a t -variate analysis in t corresponding univariate steps (Meyer, 1985 and 1991). Meyer (1991) used a reparameterisation — estimating the eigenvalues and eigenvectors of the canonical decomposition of the genetic and residual covariance matrices instead of the covariance components — to carry out a simple animal model analysis efficiently using a derivative-free REML algorithm. This resulted in a ‘full order’ fit, i.e. attempted to estimate all t eigenvalues simultaneously.

However, it can be adapted to a ‘reduced order fit’ simply by fixing a number of eigenvalues at zero and maximising the likelihood with respect to the remaining parameters only. This can be done successively. As the canonical decomposition does not use the ordering of traits, fitting k eigenvalues always estimates the k largest values. Moreover, it gives a likelihood ratio test criterion as a by-product : the minimum reduced order fit is reached when allowing for an additional non-zero eigenvalue does not cause a significant increase in likelihood.

Model of analysis

Let

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{u} + \mathbf{e} \tag{7}$$

denote the multivariate linear model of analysis, with \mathbf{y} , \mathbf{b} , \mathbf{u} and \mathbf{e} the vectors of observations, fixed effects, random effects and residual errors, respectively, and \mathbf{X} and \mathbf{Z} the incidence matrices pertaining to \mathbf{b} and \mathbf{u} . For an animal

model, \mathbf{u} always includes the vector of animals' additive genetic effects (\mathbf{a}), and may contain additional random effects, for instance, animals' maternal genetic effects, and permanent or common environmental effects such as litter effects in multiparous species.

Further, let $V(\mathbf{u}) = \mathbf{G}$, $V(\mathbf{e}) = \mathbf{R}$ and $Cov(\mathbf{u}, \mathbf{e}') = \mathbf{0}$, so that $V(\mathbf{y}) = \mathbf{V} = \mathbf{ZGZ}' + \mathbf{R}$. Assume a multivariate normal distribution, i.e. $\mathbf{y} \sim N(\mathbf{Xb}, \mathbf{V})$. Let there be t different ages measured per animal, with single records per age. For simplicity, consider a basic animal model with animals' additive genetic effects the only random effects fitted, i.e., $\mathbf{u} = \mathbf{a}$, and assume all individuals have records for all ages.

Let $\Sigma_A = \{\sigma_{A_{ij}}\}$ and $\Sigma_E = \{\sigma_{E_{ij}}\}$ denote the $t \times t$ matrices of additive genetic and residual covariances between measurements. This gives $\mathbf{G} = \mathbf{A} \times \Sigma_A$ where \mathbf{A} is the numerator relationship matrix and $' \times '$ denotes the direct matrix product. Similarly, assuming \mathbf{y} is ordered according to ages within animals, $\mathbf{R} = \mathbf{I}_N \times \Sigma_E$ with \mathbf{I}_N an identity matrix of size N , i.e. the residual covariance matrix is blockdiagonal for animals.

The REML (log) likelihood ($\log \mathcal{L}$) is then

$$\log \mathcal{L} = -\frac{1}{2} [\text{const} + N \ln |\Sigma_E| + N_A \ln |\Sigma_A| + t \ln |\mathbf{A}| + \ln |\mathbf{C}| + \mathbf{y}' \mathbf{P} \mathbf{y}] \quad (8)$$

where N_A is the total number of animals in the analysis, including any parents without records, \mathbf{C} is the coefficient matrix in the mixed model equations (MME) pertaining to (7) (or a full rank submatrix thereof), and \mathbf{P} is a matrix,

$$\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1} \mathbf{X} (\mathbf{X}' \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}' \mathbf{V}^{-1} \quad (9)$$

REML estimates of the (co)variance components to be estimated are obtained by maximising (8) with respect to the $t(t+1)$ distinct elements of the symmetric matrices Σ_A and Σ_E , using any suitable optimisation procedure. For this purpose, $\ln |\mathbf{A}|$ is a constant and can be omitted. Commonly, a derivative-free search has been used for such animal model analyses (Meyer, 1991), but recently algorithms using derivatives of the likelihood have become available (e.g. Meyer and Smith, 1996). In particular, the so-called "average information" procedure outlined by Johnson and Thompson (1995) for univariate analyses appears to perform well.

Reparameterisation

The multivariate, "finite-dimensional" REML analysis can be adapted to the estimation of CFs or, more precisely their coefficient matrices, through a simple reparameterisation. Let \mathcal{A} and \mathcal{E} denote the covariance functions of additive genetic and residual errors with coefficients matrices \mathbf{K}_A and \mathbf{K}_E , respectively. From (2), $\Sigma_A = \Phi \mathbf{K}_A \Phi'$ and $\Sigma_E = \Phi \mathbf{K}_E \Phi'$, i.e. the likelihood (8) can be rewritten as a function of the coefficient matrices

$$\log \mathcal{L} = -\frac{1}{2} [\text{const} + N \ln |\mathbf{K}_E| + N_A \ln |\mathbf{K}_A| + \ln |\mathbf{C}| + \mathbf{y}' \mathbf{P} \mathbf{y} + (N + N_A) \ln |\Phi \Phi'| + t \ln |\mathbf{A}|] \quad (10)$$

and REML estimates of the distinct (method of symmetric coefficients) or non-zero (method of asymmetric coefficients) elements of \mathbf{K}_A and \mathbf{K}_E can be obtained by maximising (10) as above. The last two terms in (10), $\ln |\mathbf{A}|$ and $\ln |\Phi \Phi'|$, do not depend on the parameters to be estimated and can be omitted in determining the maximum of $\log \mathcal{L}$.

This accommodates both a full and reduced order fit. Moreover, polynomials of different order can be fitted for \mathcal{A} and \mathcal{E} , respectively. In that case, Φ has different numbers of columns (k_A and k_E), i.e. the constant, $(N + N_A) \ln |\Phi \Phi'|$, in (10), needs to be replaced by $N_A \ln |\Phi_A \Phi_A'| + N \ln |\Phi_E \Phi_E'|$.

Alternatively, a reduced order fit can be implemented by considering matrices \mathbf{K}_A , \mathbf{K}_E and Φ to be of size $t \times t$ but fixing the elements of rows and columns of \mathbf{K}_A and \mathbf{K}_E corresponding to higher order coefficients not fitted ($k_A + 1, \dots, t$ and $k_E + 1, \dots, t$, respectively) at zero and maximising the conditional likelihood with respect to the remaining $(k_A(k_A + 1) + k_E(k_E + 1))/2$ coefficients (considering the non-zero submatrices of \mathbf{K}_A and \mathbf{K}_E only in calculating $\ln |\mathbf{K}_A|$ and $\ln |\mathbf{K}_E|$). In this framework, the reparameterisation can also be thought of as a transformation of the data to $\mathbf{y}^* = (\Phi^{-1} \times \mathbf{I}_N) \mathbf{y}$. Reducing the order of polynomial fit is then equivalent to assuming that the variables on the new scale corresponding to the omitted eigenvalues have variance zero.

Measurement errors

Usually, records are assumed to be affected by both permanent and temporary environmental effects. The latter are often assumed to be uncorrelated or even identically distributed and then called measurement errors or, in the analysis of time series, ‘white noise’. Alternatively, we might consider temporary environmental effects to represent a certain random process with a corresponding structured, non-diagonal covariance matrix. For instance, a stationary time series would result in ‘auto-correlated’ errors, records separated by a time lag of i assumed to have a correlation amongst temporary environmental errors of ρ^i (where ρ is the auto-correlation).

Let ε denote the vector of temporary environmental effects pertaining to \mathbf{y} with covariance matrix $\mathbf{I}_N \times \Sigma_\varepsilon$, and $\Sigma_R = \{\sigma_{R_{ij}}\}$ the matrix of permanent environmental covariances. Unless animals are measured repeatedly for the same traits (ages) and a corresponding permanent environmental effect, \mathbf{r} , is included in the model of analysis, $\mathbf{e} = \mathbf{r} + \varepsilon$ and $\Sigma_E = \Sigma_R + \Sigma_\varepsilon$, i.e. under the ‘finite’ model we cannot disentangle permanent and temporary environmental variation when estimating (co)variance components.

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 phenotypic (or residual) covariance matrix due to measurement errors. In essence, this involves extrapolating to the diagonals after the coefficients of the CF have been estimated using only the off-diagonals of the estimated phenotypic (or residual) covariance matrix. For t ages, it implies that the maximum order of fit for the CF is $t - 1$ rather than t . The authors illustrate the procedure for their method of asymmetric coefficients, considering the example of test day records for milk yield of dairy cows.

This can be implemented analogously in the REML framework by fitting a CF for the permanent environmental effects due to the individual, \mathcal{R} , together with explicit measurement errors rather than a CF for residuals, \mathcal{E} . Assume that temporary environmental effects are independent but allow for differences in variability, i.e., let $\Sigma_\varepsilon = \text{Diag}\{\sigma_{\varepsilon_i}^2\}$ for $i = 1, \dots, t$. Replacing Σ_E by $\Sigma_R + \Sigma_\varepsilon$ and rewriting Σ_R as $\Phi \mathbf{K}_R \Phi'$ then gives

$$\log \mathcal{L} = -\frac{1}{2} \left[\text{const} + N \ln |\Phi \mathbf{K}_R \Phi' + \text{Diag}\{\sigma_{\varepsilon_i}^2\}| + N_A \ln |\mathbf{K}_A| + \ln |\mathbf{C}| + \mathbf{y}' \mathbf{P} \mathbf{y} + N_A \ln |\Phi \Phi'| + t \ln |\mathbf{A}| \right] \quad (11)$$

where \mathbf{K}_R is the matrix of coefficients pertaining to \mathcal{R} .

As for Kirkpatrick *et al.*'s (1994) least-squares procedure, the maximum order of fit for \mathcal{R} in (11) is $t - 1$ rather than t . In fact, fitting \mathcal{E} to the order t and fitting \mathcal{R} to the order $t - 1$ together with t independent measurement errors yields equivalent models, with the same number of parameters used to described permanent and temporary environmental variation and the same likelihood. More generally, (11) has $(k_A(k_A + 1) + k_R(k_R + 1))/2 + t$ parameters, at least $t + 2$ and at most $t(t + 1)$, with $k_A \leq t$ and $k_R < t$ denoting the order of fit for \mathcal{A} and \mathcal{R} , respectively. Fitting both \mathcal{A} and \mathcal{R} to a reduced order then implies that on the ‘ \mathbf{y}^* ’ scale’ (see above) the corresponding number of ‘traits’ has phenotypic variance equal to measurement errors only. This allows for a reduced order fit for environmental effects while ‘preserving’ the phenotypic variance for all traits, thus making likelihoods for different orders of fit directly comparable.

Maximum likelihood estimation of covariance functions

General case Calculation of the REML log likelihood for the multivariate “finite-dimensional” case has been considered in detail by Meyer (1991). This involves setting up and factoring a matrix \mathbf{M} , the coefficient matrix of the MME pertaining to (7) augmented by the vector of right hand sides and its transpose and a quadratic in the data, to evaluate $\ln |\mathbf{C}|$ and $\mathbf{y}' \mathbf{P} \mathbf{y}$. Since a given coefficient matrix \mathbf{K} yields, for given Φ , a unique estimate of the corresponding covariance matrix for the t measurements involved, the same procedure can be used to evaluate (11). For each likelihood evaluation, Σ_A and Σ_E are simply calculated as $\Phi \mathbf{K}_A \Phi'$ and $\Phi \mathbf{K}_E \Phi' + \Sigma_\varepsilon$ before $\log \mathcal{L}$ is calculated on the ‘variance component’ scale.

Canonical scale For the special case of a simple animal model with equal design matrices for all traits, Meyer (1991) showed that $\log \mathcal{L}$ can be determined trait by trait, exploiting a transformation to canonical scale. Let $\mathbf{X} = \mathbf{I}_t \times \mathbf{X}_0$ and $\mathbf{Z} = \mathbf{I}_t \times \mathbf{Z}_0$. For Σ_A positive semi-definite and Σ_E positive definite, there exists a matrix \mathbf{Q} such that

$$\mathbf{Q} \Sigma_A \mathbf{Q}' = \Lambda$$

and

$$\mathbf{Q}\Sigma_E\mathbf{Q}' = \mathbf{I}_t$$

(e.g. Graybill, 1969), where Λ is a diagonal matrix with elements $\lambda_i \geq 0$ which are the eigenvalues of $\Sigma_E^{-1/2}\Sigma_A\Sigma_E^{-1/2}$. Transforming the data to ‘canonical’ variables

$$\mathbf{y}^* = (\mathbf{Q} \times \mathbf{I}_N)\mathbf{y}$$

then yields t new traits which are uncorrelated and have unit error variances. This makes the corresponding coefficient matrix of the MME block diagonal for traits, and the factorisation of the t -variate matrix \mathbf{M} can be carried out by factoring t univariate matrices

$$\mathbf{M}_i^* = \begin{bmatrix} \mathbf{X}'_0\mathbf{X}_0 & \mathbf{X}'_0\mathbf{Z}_0 & \mathbf{X}'_0\mathbf{y}_i^* \\ \mathbf{Z}'_0\mathbf{X}_0 & \mathbf{Z}'_0\mathbf{Z}_0 + \lambda_i^{-1}\mathbf{A}^{-1} & \mathbf{X}'_0\mathbf{y}_i^* \\ \mathbf{y}_i^{*\prime}\mathbf{X}_0 & \mathbf{y}_i^{*\prime}\mathbf{Z}_0 & \mathbf{y}_i^{*\prime}\mathbf{y}_i^* \end{bmatrix} \quad (12)$$

where \mathbf{y}_i^* is the subvector of \mathbf{y}^* for the i -th trait. This yields terms $\ln|\mathbf{C}_i^*|$ and $\mathbf{y}_i^{*\prime}\mathbf{P}_i^*\mathbf{y}_i^*$, and (8) can be rewritten as

$$\log \mathcal{L} = -\frac{1}{2} \left[N_A \sum_{i=1}^t \ln \lambda_i + \sum_{i=1}^t \ln |\mathbf{C}_i^*| + \sum_{i=1}^t \mathbf{y}_i^{*\prime} \mathbf{P}_i^* \mathbf{y}_i^* + t \ln |\mathbf{A}| \right. \\ \left. + (N - r(\mathbf{X}_0)) \ln |\Sigma_E| \right] \quad (13)$$

see Meyer (1991) for further details.

The number of non-zero elements of Λ is equal to the number of non-zero eigenvalues, or the rank of Σ_A . Hence, for the method of symmetric coefficients, fitting \mathcal{A} to reduced order k_A implies that $t - k_A$ eigenvalues λ_i are zero. For $\lambda_i = 0$, $\ln|\mathbf{C}_i^*|$ reduces to $\ln|\mathbf{X}'_0\mathbf{X}_0|$, i.e., is a constant which needs to be evaluated only once per analysis. Similarly, for $\lambda_i = 0$,

$$\mathbf{y}_i^{*\prime} \mathbf{P}_i^* \mathbf{y}_i^* = \mathbf{y}_i^{*\prime} \mathbf{P}_0 \mathbf{y}_i^* = \mathbf{y}_i^{*\prime} (\mathbf{I}_N - \mathbf{X}_0(\mathbf{X}'_0\mathbf{X}_0)^{-1}\mathbf{X}'_0) \mathbf{y}_i^* \quad (14)$$

which depends on the canonical transformation only. For each likelihood evaluation, it can be evaluated as a linear combination of the corresponding residual sums of squares and crossproducts on the original scale

$$\mathbf{y}_i^{*\prime} \mathbf{P}_0 \mathbf{y}_i^* = \sum_{k=1}^t \sum_{m=1}^t q_{ik} q_{im} \mathbf{y}_k' \mathbf{P}_0 \mathbf{y}_m \quad (15)$$

where q_{km} denotes the km -th element of \mathbf{Q} . Again, terms $\mathbf{y}_k' \mathbf{P}_0 \mathbf{y}_m$ need to be determined only once per analysis. Effectively, this makes the computational requirements for each likelihood evaluation for t traits in the equal design matrix case proportional to that for k_A corresponding univariate analyses.

Maximising $\log \mathcal{L}$ The likelihood can then be maximised with respect to the elements of \mathbf{K}_A , \mathbf{K}_E and Σ_e using a simple derivative-free search strategy such as Nelder and Mead’s (1965) simplex procedure or Powell’s (1965) method of conjugate directions, or a Quasi-Newton algorithm which approximates both first and second derivatives of the likelihood (Meyer, 1989). Carrying out a constrained maximisation which forces the estimated coefficient matrices (\mathbf{K}) to be positive (semi-) definite then ensures, for Φ of full rank, that the estimated covariance functions are positive (semi-) definite (Graybill, 1969), eliminating the major drawback of Kirkpatrick *et al.*’s (1990) weighted least-squares procedure.

While derivative-free algorithms are simple to implement and easy to use, they have been shown to be slow to converge, especially for analyses involving multiple traits ($t \geq 4$) and thus a high-dimensional search. In that case algorithms using derivatives of the likelihood are preferable. Such procedures have been described recently for the estimation of (co)variance components fitting a multivariate animal model (Jensen *et al.*, 1996; Madsen *et al.*, 1994; Meyer, 1994; Meyer and Smith, 1996). They can be adapted to the estimation of covariance functions in the same way as described

above for a derivative-free algorithm. In essence, only a simple reparameterisation is required — the likelihood and its derivatives can, as before, be calculated on the variance component scale. Derivatives with respect to the parameters of the covariance function model can then be obtained as linear combinations of those with respect to the variance components, and then be used in a Newton type estimation procedure. Details are not within the scope of this paper; see Meyer (1996) for a description of an ‘average information’ REML algorithm to estimate covariance functions.

Extension to other models

Missing records and additional random effects For simplicity, only the simplest scenario of an animal model without any additional random effects and equal design matrices for all traits has been considered so far. As shown, there are simplifications for this case which make computational requirements proportional to a function of the order of fit for the genetic CF. The methodology presented, however, is by no means restricted to this case and extends readily to a general multivariate linear model. Albeit, computational effort required for this is generally a function of the number of ages measured, t , rather than the order of polynomial fit of the CFs.

For models with additional random effects, such as (uncorrelated) maternal genetic or permanent environmental effects, a covariance function for each of the effects can be fitted analogously to that for animals’ additive genetic effects. Instead of the covariances between ages for this effect, the coefficients of the corresponding CF are then estimated. As emphasized above, $\log \mathcal{L}$ can be calculated on the variance component scale. For missing records, diagonal blocks of \mathbf{R} are submatrices of Σ_E , and $N \ln |\Sigma_E|$ in (8) is replaced by summing corresponding terms, $N_w \ln |\Sigma_E^w|$, for the various combinations of traits recorded; see Meyer (1991) for details.

Calculation of $\log \mathcal{L}$ then requires setting up and factoring the t -variate mixed model matrix \mathbf{M} . For large t this may impose considerable computational demands. For this, the inverses of all covariance matrices due to random effects, Σ_T for $T = A, E, \dots$, are needed. Conceptually, a reduced fit for any CF \mathcal{T} implies that the corresponding covariance matrix Σ_T has one or several eigenvalues equal to zero. In practice, however, these are set to an operational zero to avoid a generalised inverse with a number of rows and columns ‘zeroed out’, resulting in a numerically full rank inverse. This implies that regardless of the order of fit for any CFs, the work required to factor \mathbf{M} when there are missing records or additional random effects is proportional to t .

Correlated measurement errors Extensions to other models can be perceived. For instance, the assumption of a diagonal Σ_e with t distinct elements may not be appropriate, in particular for traits measured at short time intervals, e.g. daily feed intake of animals. Instead we might assume that the measurement errors represent a stationary time series and that Σ_e is a Toeplitz matrix, i.e. can be represented by only two parameters, namely the error variance σ_e and the auto-correlation ρ (Graybill, 1969).

Multivariate covariance functions In some cases, it might be desired to fit more than one covariance function for some cause of variation or a covariance matrix because measurements taken clearly represent different characters or physiological processes. Examples are data for weight and feed intake at different ages or milk, fat and protein yield for daily production of dairy cows, or direct and maternal genetic effects for the same trait which are correlated. In other instances, we might have different meta-meters (e.g. age, distance, production level) for different traits. We then need to fit a CF to describe the variation across ‘repeated’ records for each trait, and to estimate ‘cross-covariance functions’ which model the covariation between traits over time, in other words a multivariate CF. This can be accommodated in the above framework by a slightly different reparameterisation.

Let records for n traits be ordered according to age (or time) within trait, and Σ_{ij} denote the submatrix of Σ (standing in turn for Σ_A, Σ_R , etc.) for the i -th and j -th trait. Σ can then be rewritten as (see (2))

$$\begin{bmatrix} \Sigma_{11} & \cdots & \Sigma_{1n} \\ \vdots & \ddots & \vdots \\ \Sigma_{n1} & \cdots & \Sigma_{nn} \end{bmatrix} = \begin{bmatrix} \Phi_1 & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \Phi_n \end{bmatrix} \begin{bmatrix} \mathbf{K}_{11} & \cdots & \mathbf{K}_{1n} \\ \vdots & \ddots & \vdots \\ \mathbf{K}_{n1} & \cdots & \mathbf{K}_{nn} \end{bmatrix} \begin{bmatrix} \Phi'_1 & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \Phi'_n \end{bmatrix} \quad (16)$$

where a diagonal block Φ_i is the matrix Φ for the i -th trait as described above for ‘univariate’ CFs, For traits with the same meta-meter (e.g. measured at the same time), Φ_i are identical. This gives $\Sigma_{ij} = \Phi_i \mathbf{K}_{ij} \Phi'_j$, i.e., $n(n+1)/2$ CFs and their coefficient matrices have to be estimated for each covariance matrix Σ .

Table 1 : REML estimates of the coefficients of the genetic covariance function (\mathbf{K}_A), the resulting covariance function (\mathcal{A}) and the resulting genetic covariance (Σ_A) matrix Kirkpatrick *et al.*'s (1990) example, assuming a data set consisting of records for 1000 half-sib families of size 5 and simulating a measurement error of 500.

| Fit ^a | | Element | | | | | |
|------------------|---|--------------|--------------|---------------|-------------|--------------|-------------|
| | | 11 | 12 | 13 | 22 | 23 | 33 |
| \mathbf{K}_A | 1 | 1047.1 | | | | | |
| | 2 | 1120.9 | 75.8 | | 68.1 | | |
| | 3 | 1348.4 | 66.4 | -111.8 | 24.2 | -14.0 | 14.6 |
| \mathcal{A}^b | 1 | 523.6 | | | | | |
| | | <i>324.0</i> | | | | | |
| | 2 | 560.5 | 65.6 | | 102.0 | | |
| | | <i>312.2</i> | <i>-11.9</i> | | <i>24.5</i> | | |
| | 3 | 808.3 | 71.1 | -214.9 | 36.3 | -40.7 | 82.0 |
| | | <i>808.0</i> | <i>71.2</i> | <i>-215.0</i> | <i>36.4</i> | <i>-40.7</i> | <i>81.6</i> |
| Σ_A | 1 | 523.6 | 523.6 | 523.6 | 523.6 | 523.6 | 523.6 |
| | 2 | 531.2 | 494.8 | 458.4 | 560.5 | 626.1 | 793.8 |
| | 3 | 436.1 | 522.4 | 424.2 | 808.3 | 664.6 | 557.8 |

^aOrder of polynomial fit for \mathcal{A}

^bFirst line : REML estimates, second line (in italic) : Kirkpatrick *et al.*'s (1990) weighted least squares estimates

Numerical Examples

Example 1 : Kirkpatrick's example

This section contrasts estimates of covariance functions obtained directly from the data by REML with estimates obtained from an estimated covariance function using a least-squares approach, for the example given by Kirkpatrick *et al.* (1990).

As shown above, the contribution of the data to $\log \mathcal{L}$ is essentially a function of the sums of squares (SS) and crossproducts (CP) in the data vector. For a given family structure, a 'deterministic' simulation can be carried out by calculating the SS/CP contributed by each family directly from the population values of covariances between traits. The non-data part of $\log \mathcal{L}$ then depends on the model of analysis and the assumed values for the parameters to be estimated. Hence, varying these and maximising the resulting $\log \mathcal{L}$, maximum likelihood estimates for a given data structure and population values under different models can be obtained. This is equivalent to sampling with many replicates; see Meyer (1992) for an application and procedural details.

This technique was applied to the example of three body weights in mice given by Kirkpatrick *et al.* (1990). It was assumed that the genetic covariance matrix given was the matrix of population values, and that there were no permanent environmental effects but that all traits were affected by measurement errors with variance 500. Data were considered to have a simple, balanced paternal-half sib structure, consisting of 1000 sire families of size 5.

Estimates of the coefficient matrix of the genetic CF, the CF itself and the resulting genetic covariance matrix among the 3 ages, fitting \mathcal{A} to the order 1, 2 and 3 are summarised in Table 1. While estimates for the full order fit agreed closely with those of Kirkpatrick *et al.* (1990), REML estimates for a reduced fit are quite different from the weighted least-squares estimates obtained by them. Presumably this is due in part to the fact that measurement errors as well as additive genetic effects were considered at the same time. As shown in Table 2, estimates of the variances due to measurement errors are biased for a reduced fit. In agreement with Kirkpatrick *et al.*'s χ^2 test criterion, likelihood values indicate that only the full fit CF describes the data adequately.

Example 2 : Repeatability model

This example shows estimates of covariance functions and corresponding log likelihoods for increasing orders of fit on a simulated data set. In particular, the behaviour when overparameterising, i.e. fitting CFs to an order higher than necessary, is examined.

Table 2 : REML log likelihoods ($\log \mathcal{L}$) together with estimates of measurement errors ($\sigma_{\varepsilon_i}^2$) and eigenvalues of the additive genetic covariance function (λ_K) and resulting genetic covariance matrix (λ_A) for Kirkpatrick *et al.*'s (1990) example, assuming a data set consisting of records for 1000 half-sib families of size 5 and simulating a measurement error of 500.

| Fit ^a : | $k_A = 1$ | $k_A = 2$ | $k_A = 3$ |
|----------------------------|------------|------------|------------|
| $\log \mathcal{L}$ | -58,159.95 | -58,075.32 | -58,032.54 |
| $\sigma_{\varepsilon_1}^2$ | 529.1 | 428.3 | 500.0 |
| $\sigma_{\varepsilon_2}^2$ | 651.5 | 672.4 | 499.9 |
| $\sigma_{\varepsilon_3}^2$ | 520.5 | 298.2 | 500.1 |
| λ_{A1} | 1570.8 | 1698.7 | 1713.8 |
| λ_{A2} | 0 | 186.8 | 82.2 |
| λ_{A3} | 0 | 0 | 6.3 |
| λ_{K1} | 1047.2 | 1126.3 | 1361.1 |
| λ_{K2} | | 62.6 | 24.5 |
| λ_{K3} | | | 1.6 |

^aOrder of fit for genetic covariance function

Data for 4 multivariate normally distributed ‘traits’, measured at equally spaced ages, were simulated (single replicate) for a repeatability model, i.e. assuming the population values for Σ_A and Σ_R were described by CFs of order 1, namely $\mathcal{A} = 250$ and $\mathcal{R} = 750$ and allowing for measurement errors $\sigma_{\varepsilon_i}^2 = 100$. Records were generated for a balanced hierarchical full-sib design with 1000 sires each mated to 3 dams and 2 progeny per family, i.e. 10,000 animals in total (records available for both generations). These data were analysed as described above using a derivative-free algorithm to maximise $\log \mathcal{L}$, fitting CFs \mathcal{A} and \mathcal{R} to the order $k_A \leq k_R = 1, 2, 3$ together with measurement errors and fitting CFs \mathcal{A} and \mathcal{E} for $k_A = 1, \dots, 4$ and $k_E = 4$.

Estimates of elements of the coefficient matrices and measurement errors together with the maximum for $\log \mathcal{L}$ are given in Table 3. Overall, estimates agree with the population values and are very consistent for the different orders of fit. Note that for $k_R = 4$, \mathcal{R} (\mathcal{E}) includes the measurement errors, and thus has a different expected value. While estimates of linear and higher order coefficients were close to zero (other than for \mathbf{K}_R for a full order fit for residual, $k_r = 4$), likelihoods increased slightly with increasing order of polynomial fit.

Had analyses been performed sequentially, we would have stopped at fit 22, as estimating an extra 4 parameters (compared to fit 11) only increased $\log \mathcal{L}$ by 0.77. In this and other simulations not shown, convergence of the derivative-free algorithm used was slow and frequently unreliable with increasing numbers of parameters, in particular when attempting to estimate higher order coefficients which had population values of zero. Note that $\log \mathcal{L}$ for orders of fit 31, 32 and 33 was expected to be the same as for 41, 42 and 43, respectively. Consistently somewhat lower values for the former highlight the existence of convergence problems for these analyses.

Example 3 : Analysis of beef cattle data

This example gives an application of the covariance function model to a data set with missing records arising from a selection experiment in beef cattle. In particular, the assumptions about the shape of the covariance functions for different orders of fit are illustrated.

Mean January weights at two to six years of age for a total of 913 ‘Wokalup’ cows are given in Table 4. This is a synthetic breed, part of a selection experiment in beef cattle, named after the research station of that name in Western Australia. Records are a subset of the data considered by Meyer (1995) who, excluding weight at two years, treated them as repeated measurements of mature weight, fitting either a repeatability model or a Gompertz growth curve for each animal.

Covariance functions of increasing order were estimated for these data, fitting genetic and environmental CFs to the same order throughout ($k_A = k_R = k$). In addition, ‘finite-dimensional’ multivariate analyses were carried out using an average information algorithm, which yielded approximate lower bound sampling errors of covariance component estimates as a by-product. The model of analysis was a simple animal model, fitting year-paddock subclasses as the only fixed effects. Including parents without records, there were a total of 1125 animals in the analysis.

Table 3 : Estimates of coefficient matrices for covariance functions due to additive genetic (K_{Aij}) and permanent environmental (K_{Rij}) (or residual) effects together with measurement errors ($\sigma_{\epsilon_i}^2$) and corresponding log likelihood.

| | Pop. ^a | Fit 11 ^b | Fit 21 | Fit 22 | Fit 31 | Fit 32 | Fit 33 | Fit 41 | Fit 42 | Fit 43 | Fit 44 |
|-------------------------|-------------------|---------------------|---------|---------|---------|---------|--------|------------|--------|--------|--------|
| K_{A11} | 500.0 | 449.2 | 449.2 | 447.5 | 448.7 | 452.5 | 451.9 | 448.7 | 448.1 | 453.4 | 452.2 |
| K_{A12} | | | | -3.5 | | -3.4 | -3.5 | | -3.5 | -3.6 | -1.3 |
| K_{A13} | | | | | | | -1.8 | | | -1.8 | -1.9 |
| K_{A14} | | | | | | | | | | | -1.9 |
| K_{A22} | | | | 0.2 | | 0.2 | 0.2 | | 0.3 | 0.3 | 0.5 |
| K_{A23} | | | | | | | 0.0 | | | 0.0 | 0.3 |
| K_{A24} | | | | | | | | | | | -0.2 |
| K_{A33} | | | | | | | 0.3 | | | 0.3 | 0.4 |
| K_{A34} | | | | | | | | | | | -0.2 |
| K_{A44} | | | | | | | | | | | 0.1 |
| K_{R11} | 1500.0 | 1568.9 | 1568.8 | 1570.2 | 1574.4 | 1571.7 | 1571.5 | 1634.7 | 1635.9 | 1630.8 | 1631.2 |
| K_{R12} | | | 0.2 | 3.1 | 0.2 | 3.1 | 3.2 | -1.5 | 1.5 | 1.6 | -0.3 |
| K_{R13} | | | | | -3.4 | -3.4 | -1.8 | -19.5 | -19.5 | -17.9 | -17.8 |
| K_{R14} | | | | | | | | 1.2 | 1.2 | 1.2 | 2.7 |
| K_{R22} | | | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 68.1 | 67.8 | 67.8 | 67.5 |
| K_{R23} | | | | | -0.0 | -0.0 | -0.0 | 0.2 | 0.2 | 0.1 | -0.1 |
| K_{R24} | | | | | | | | -33.0 | -33.0 | -33.0 | -32.8 |
| K_{R33} | | | | | 0.2 | 0.3 | 0.0 | 22.0 | 22.0 | 21.7 | 21.6 |
| K_{R34} | | | | | | | | -0.3 | -0.3 | -0.3 | -0.1 |
| K_{R44} | | | | | | | | 27.9 | 27.9 | 27.9 | 27.8 |
| $\sigma_{\epsilon 1}^2$ | 100.0 | 98.4 | 98.4 | 98.0 | 98.2 | 97.5 | 98.0 | Not fitted | | | |
| $\sigma_{\epsilon 2}^2$ | 100.0 | 97.9 | 98.0 | 98.0 | 97.5 | 97.4 | 97.2 | | | | |
| $\sigma_{\epsilon 3}^2$ | 100.0 | 95.5 | 95.55 | 95.5 | 94.9 | 94.8 | 95.0 | | | | |
| $\sigma_{\epsilon 4}^2$ | 100.0 | 96.5 | 96.5 | 96.1 | 96.6 | 96.0 | 95.4 | | | | |
| No. par.s ^c | | 6 | 8 | 10 | 11 | 13 | 16 | 11 | 13 | 16 | 20 |
| $\log \mathcal{L}^d$ | | -102.24 | -102.23 | -101.47 | -100.89 | -100.07 | -99.51 | -100.78 | -99.83 | -99.26 | -98.71 |

^aPopulation values

^bOrder of fit of covariance functions : $k_R k_A$

^cNo. of parameters fitted

^d \log likelihood + 138,400

Estimates of the coefficients of the genetic CF \mathcal{A} and the resulting genetic covariance matrix ('regenerated' for the ages in the data) are summarised in Table 5. The re-generated covariance matrix for the full-order fit ($k = 5$) was expected to agree with the estimate of the genetic covariance matrix from the conventional multivariate analysis. Some small discrepancies exist in this case but they are well within the range of sampling errors, and, as discussed below, can be attributed to convergence problems.

For five equally spaced ages, the standardised age for the middle value is zero, i.e. the estimate of the genetic variance at 4 years of age is equal to the scalar term in \mathcal{A} . Suppose we want to determine the genetic covariance between weight at ages 4.5 years and 5.5 years. On the standardised scale (range -1 to 1) these are equal to 0.25 and 0.75 , respectively. For $k = 3$, this gives the covariance as (see (1))

$$\begin{bmatrix} 1 & 0.25 & 0.25^2 \end{bmatrix} \begin{bmatrix} 1815.8 & 424.0 & -431.5 \\ 424.0 & 182.0 & -121.6 \\ -431.5 & -121.6 & 141.0 \end{bmatrix} \begin{bmatrix} 1 \\ 0.75 \\ 0.75^2 \end{bmatrix} = 1986.4$$

Table 6 gives the corresponding eigenvalues of the CFs fitted, estimates of measurement error and phenotypic variances, and $\log \mathcal{L}$. Clearly, a simple repeatability model does not describe the data adequately, while augmenting the order of fit from 2 to 3 (estimating 6 additional parameters) increased $\log \mathcal{L}$ only by 5.22 compared to a value of $\chi_{6,5\%}^2 = 18.31$. Although $\log \mathcal{L}$ did not increase significantly above $k = 2$, estimated covariance matrices, variances and eigenvalues of the CFs were very similar only from $k = 3$ onwards.

Table 4 : January weights of Wokalup cows at 2 to 6 years of age.

| Age (years) | 2 | 3 | 4 | 5 | 6 |
|----------------------|-------|-------|-------|-------|-------|
| No. records | 808 | 662 | 513 | 440 | 372 |
| Mean (kg) | 447.5 | 522.2 | 584.0 | 611.9 | 625.7 |
| SD ^a (kg) | 57.3 | 70.7 | 71.8 | 72.5 | 74.0 |

^aStandard Deviation

Table 5 : Estimates of the coefficients of the genetic covariance function \mathcal{A} and the resulting genetic covariances ($\sigma_{A_{ij}}$) for increasing order of polynomial fit (k), together with corresponding estimates of covariance components and their approximate sampling errors (s.e.) from ‘finite-dimensional’, multivariate analyses (Cov).

| | Coefficients of \mathcal{A} | | | | | $\sigma_{A_{ij}}$ | Genetic covariances | | | | | Cov \pm s.e. |
|-----|-------------------------------|---------|---------|---------|---------|-------------------|---------------------|---------|---------|---------|---------|----------------|
| | $k = 1$ | $k = 2$ | $k = 3$ | $k = 4$ | $k = 5$ | | $k = 1$ | $k = 2$ | $k = 3$ | $k = 4$ | $k = 5$ | |
| 0,0 | 884.9 | 1222.7 | 1815.8 | 1789.1 | 1828.5 | 11 | 885 | 736 | 672 | 691 | 695 | 672 \pm 197 |
| 0,1 | | 341.4 | 424.0 | 291.3 | 289.6 | 12 | 885 | 809 | 858 | 885 | 872 | 855 \pm 246 |
| 0,2 | | | -431.5 | -393.7 | -582.4 | 13 | 885 | 881 | 960 | 930 | 964 | 1048 \pm 296 |
| 0,3 | | | | 173.9 | 171.7 | 14 | 885 | 954 | 978 | 920 | 907 | 790 \pm 290 |
| 0,4 | | | | | 179.1 | 15 | 885 | 1026 | 911 | 951 | 958 | 944 \pm 314 |
| 1,1 | 196.4 | 182.0 | 349.5 | 419.7 | | 22 | 885 | 930 | 1261 | 1368 | 1347 | 1277 \pm 461 |
| 1,2 | | -121.6 | -144.2 | -134.8 | | 23 | 885 | 1052 | 1496 | 1523 | 1528 | 1659 \pm 453 |
| 1,3 | | | -205.7 | -275.4 | | 24 | 885 | 1174 | 1563 | 1535 | 1496 | 1383 \pm 455 |
| 1,4 | | | | -81.1 | | 25 | 885 | 1295 | 1463 | 1587 | 1580 | 1553 \pm 504 |
| 2,2 | | 141.0 | 136.9 | 232.8 | | 33 | 885 | 1223 | 1816 | 1789 | 1829 | 2221 \pm 643 |
| 2,3 | | | -3.0 | 21.1 | | 34 | 885 | 1393 | 1920 | 1858 | 1860 | 1914 \pm 556 |
| 2,4 | | | | -113.9 | | 35 | 885 | 1564 | 1808 | 1861 | 1887 | 2090 \pm 607 |
| 3,3 | | | 250.1 | 317.2 | | 44 | 885 | 1613 | 2048 | 1965 | 1946 | 1722 \pm 691 |
| 3,4 | | | | -21.9 | | 45 | 885 | 1833 | 1948 | 1932 | 1924 | 1828 \pm 654 |
| 4,4 | | | | 117.5 | | 55 | 885 | 2102 | 1881 | 1963 | 1966 | 1988 \pm 793 |

This can be seen more clearly in Figure 1 which shows the ‘surface’ of estimated genetic covariances for different orders of fit. For $k = 1$ (not shown), all covariances are equal. Graphically, that is a plane parallel to the base. For $k = 2$, covariances are linear functions of the ages, resulting in a tilted plane. Including quadratic ages for $k = 3$ then gives a parabolic surface. Considering cubic ($k = 4$) or quartic ($k = 5$) terms in addition then adds a few creases to the surface but does not change its shape dramatically. As expected for an estimate of the fifth eigenvalue of \mathcal{A} of zero, surfaces for $k = 4$ and $k = 5$ are virtually indistinguishable. Figure 2 gives the corresponding plots for the phenotypic covariances among the 5 ages. Surfaces are dominated by peaks for the variances, reflecting substantial variances due to measurement errors. Again, there are few differences between plots for $k \geq 3$.

As above, there were some convergence problems : analyses fitting CFs to full order and conventional multivariate analyses are expected to give the same covariance estimates and $\log \mathcal{L}$. Presumably this is due to the high dimension of derivative-free search (30 parameters) in conjunction with the fact that for $k = 4$ or $k = 5$ we are attempting to estimate parameters which are not necessary to describe the data. Algorithms using derivatives of the likelihood may perform better in this case.

Discussion

As shown, covariance functions enable us to model our data with the least number of parameters necessary. This avoids problems associated with overparameterised models and makes efficient use of the data. Moreover, they allow covariances between ages for which no records are available to be estimated. These should, however, be in the range of ages covered by the data.

While conceptually able to cope with data coming in ‘at all ages’, practical problems remain for the covariance function model in this case. As emphasized above, computational requirements generally increase with the number of observed ages rather than the order fit of covariance functions estimated. It may then be necessary to make the grid of ages coarser than desired, for instance years rather than months of age, possibly in conjunction with some adjustment for

Table 6 : Estimates of eigenvalues of the genetic (λ_A) and residual or error ($\lambda_{R(E)}$) covariance function for weights of Wokalup cows, together with estimates of measurement errors (σ_ϵ^2) and corresponding phenotypic variances (σ_P^2) for various orders of polynomial fit, $k_A = k_R = k$ and multivariate analyses (Cov).

| | | $k = 1$ | $k = 2$ | $k = 3$ | $k = 4$ | $k = 5$ | Cov \pm s.e. ^a |
|------------------------|---|---------|---------|---------|---------|---------|-----------------------------|
| λ_A | 1 | 1770 | 2511 | 3170 | 3152 | 3128 | |
| | 2 | | 66 | 59 | 86 | 92 | |
| | 3 | | | 6 | 8 | 8 | |
| | 4 | | | | 2 | 1 | |
| | 5 | | | | | 0 | |
| $\lambda_{R(E)}$ | 1 | 1002 | 2149 | 1795 | 1739 | 2635 | |
| | 2 | | 43 | 114 | 105 | 1291 | |
| | 3 | | | 1 | 37 | 991 | |
| | 4 | | | | 1 | 120 | |
| | 5 | | | | | 53 | |
| σ_ϵ^2 | 1 | 547 | 547 | 496 | 472 | | |
| | 2 | 2118 | 2006 | 1968 | 1725 | | |
| | 3 | 1984 | 1694 | 1582 | 1613 | | |
| | 4 | 2201 | 1565 | 1573 | 1452 | | |
| | 5 | 2165 | 866 | 736 | 694 | | |
| σ_P^2 | 1 | 1933 | 1631 | 1628 | 1590 | 1599 | 1597 \pm 127 |
| | 2 | 3504 | 3536 | 3774 | 3739 | 3643 | 3664 \pm 311 |
| | 3 | 3370 | 3915 | 4158 | 4158 | 4239 | 4324 \pm 424 |
| | 4 | 3587 | 4721 | 4879 | 4626 | 4576 | 4536 \pm 470 |
| | 5 | 3552 | 5202 | 4886 | 4902 | 5065 | 5043 \pm 549 |
| $\log \mathcal{L}^b$ | | -100.24 | -16.03 | -10.81 | -7.08 | -1.29 | 0 |
| No. parm. ^b | | 7 | 11 | 17 | 25 | 30 | 30 \pm |

^aApproximate sampling error

^b \log likelihood, expressed as deviation from value from multivariate analysis

^bNo. of parameters

age differences within age class in the model of analysis, e.g by fitting a covariable. While relinquishing some of the advantages of the covariance function, namely that no prior assumptions about the nature of time trends are required, this appears still preferable over more traditional growth curve type of analyses which necessitate assumptions about the shape of the curve over the complete range of ages.

Recently, longitudinal records arising in the animal breeding context such as growth data in pigs (Andersen and Pedersen, 1995) and test day records in dairy cattle (Schaeffer and Dekkers, 1994) have been modeled fitting a linear model comprising random regression coefficients on time. It has been noted (M. Goddard, 1995 pers. comm.) that the covariance function model is equivalent to a random regression model. Consider a simple regression model

$$z_l = \sum_i \gamma_i x_{il} + e_l \quad (17)$$

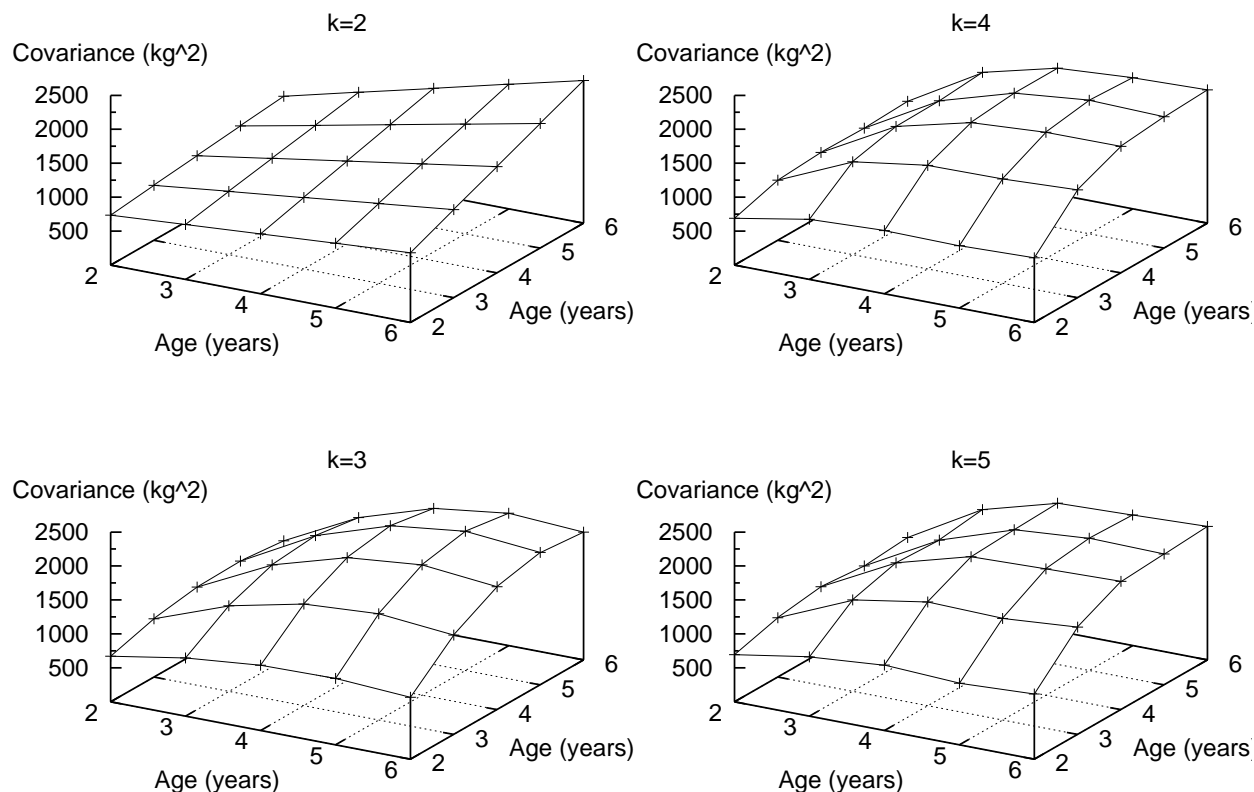
where z_l denotes an observation for the l -th individual, γ_i is the i -th regression coefficient and x_{il} is the corresponding covariable for z_l , and e_l denotes the pertaining residual. Assume that z_l has been recorded at age a_l (as above standardised to the interval from -1 to 1). Further, let the i -th covariable in (17) be equal to the i -th Legendre polynomial of a_l , i.e. $x_{il} = \phi_i(a_l)$, for $i = 0, \dots, k-1$. This gives

$$z_l = \sum_{i=0}^{k-1} \gamma_i \phi_i(a_l) + e_l \quad (18)$$

and for random regression coefficients

$$Cov(z_l, z_m) = \sum_{i=0}^{k-1} \sum_{j=0}^{k-1} \phi_i(a_l) \phi_j(a_m) Cov(\gamma_i, \gamma_j) + cov(e_l, e_m) \quad (19)$$

Figure 1 : Estimates of additive genetic covariance components for January weights of Wokalup cows at 2 to 6 years of age, calculated from estimated covariance functions fitted to the order k .



Ignoring the error covariance, the right hand side of (19) clearly describes a covariance function (c.f. (1)) with $Cov(\gamma_i, \gamma_j)$ equal to K_{ij} , the ij -th element of the coefficient matrix of the CF. Further research is required to examine the utility of this equivalence for the estimation of CFs. In particular, it should enhance the scope for dealing with data observed at many ages, as only k regression coefficients and their $k(k+1)/2$ covariances need to be estimated for each source of variation in a univariate analysis.

Conclusions

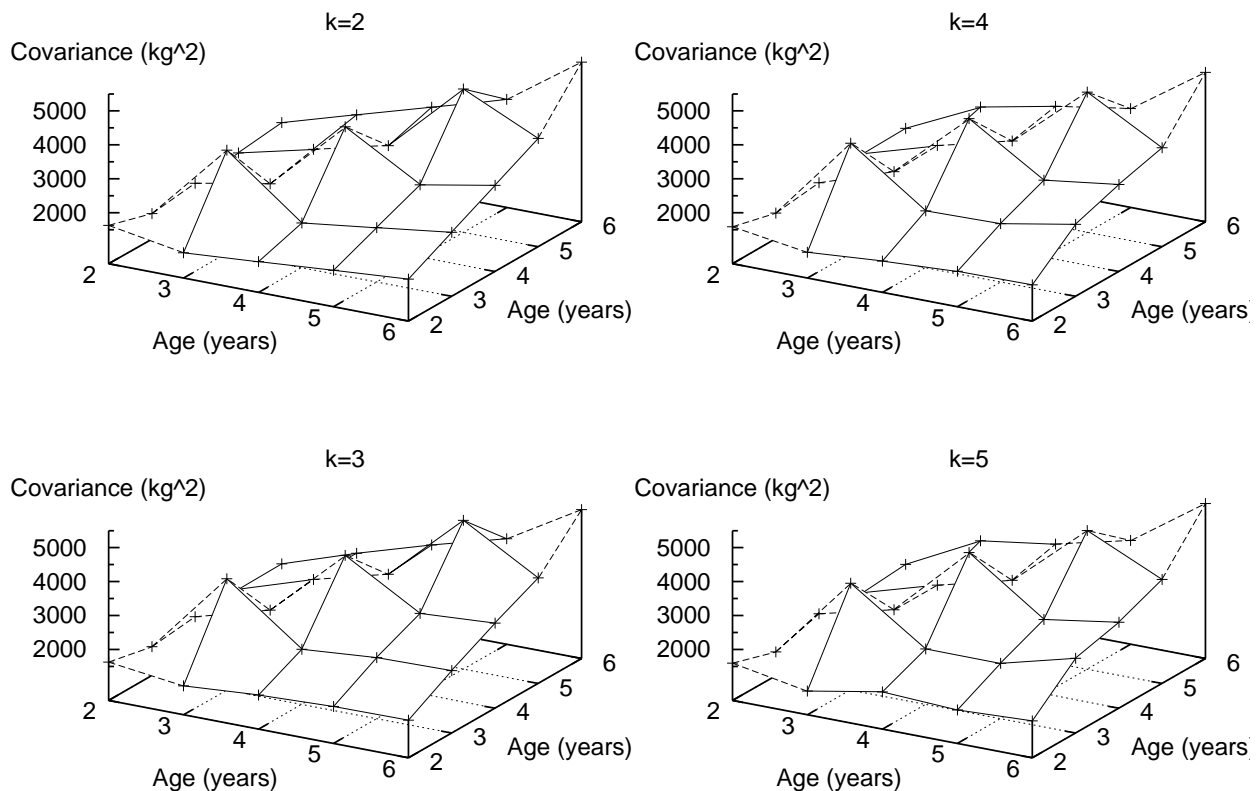
Covariance functions can be estimated readily using maximum likelihood. In essence, this involves only a simple reparameterisation of existing procedures to estimate covariance components. In contrast to the weighted least-squares approach used by Kirkpatrick *et al.* (1990 and 1994), it guarantees estimated CFs to be positive semi-definite. A likelihood ratio test can be used to determine the minimum order of fit.

The covariance function model provides a useful alternative to the analyses of repeated records used to date. In particular, it does not require any *a priori* assumptions about the number of different ‘traits’ represented by a series of measurements or the shape of any trends. Moreover, the eigenvalues and eigenvectors of a CF have an interpretation of their own, providing information on the directions in which mean growth trajectories are likely to change under selection. Potentially they could be used to characterise or summarise differences between breeds or species for sequentially measured ‘traits’ as growth.

Acknowledgements

This work was funded by a grant from the BBSRC, U.K., and grant UNE35 of the Australian Meat Research Corporation (MRC).

Figure 2 : Estimates of phenotypic covariances for January weights of Wokalup cows at 2 to 6 years of age, calculated from estimated covariance functions fitted to the order k and estimates of measurement errors.



References

- ABRAMOWITZ, M. AND STEGUN, I.A. 1965. *Handbook of Mathematical Functions*. Dover, New York.
- ANDERSEN, S. AND PEDERSEN, B. 1995. Statistical analysis of growth curve data. 2nd European Workshop on Advanced Biometrical Methods in Animal Breeding, Salzburg, Austria, June 12–20, 1995.
- DIGGLE, P.J. 1990. *Time Series – A Biostatistical Introduction*. Oxford Statistical Science Series, Clarendon Press, Oxford.
- GRAYBILL, F.A. 1969. *Introduction to Matrices with Applications in Statistics*. Wadsworth Publishing Company, Inc., Belmont, California.
- HAYES, J.F. AND HILL, W.G. 1980. A reparameterization of a genetic selection index to locate its sampling properties. *Biometrics* **36**:237–248.
- JENSEN, J., MANTYSAARI, E., MADSEN, P. AND THOMPSON, R. 1996. Restricted Maximum Likelihood estimation of (co)variance components in multivariate linear models using average of observed and expected information. (in preparation).
- JOHNSON, D.L. AND THOMPSON, R. 1995. Restricted Maximum Likelihood estimation of variance components for univariate animal models using sparse matrix techniques and average information. *J. Dairy Sci.* **78**:449–456.
- KIRKPATRICK, M. AND HECKMAN, N. 1989. A quantitative genetic model for growth, shape and other infinite-dimensional characters. *J. Math. Biol.* **27**:429–450.
- KIRKPATRICK, M., LOFSVOLD, D. AND BULMER, M. 1990. Analysis of the inheritance, selection and evolution of growth trajectories. *Genetics* **124**:979–993.

- KIRKPATRICK, M., HILL, W.G. AND THOMPSON, R. 1994. Estimating the covariance structure of traits during growth and ageing, illustrated with lactations in dairy cattle. *Genet. Res.* **64**:57–69.
- LINDSEY, J.K. 1993. *Models for Repeated Measurements*. Oxford Statistical Science Series, Clarendon Press, Oxford.
- MADSEN, P., JENSEN, J. AND THOMPSON, R. 1994. Estimation of (co)variance components by REML in multivariate mixed linear models using average of observed and expected information. *5th World Congr. Genet. Appl. Livest. Prod.* **Vol 22**:19–22.
- MEYER, K. 1989. Restricted Maximum Likelihood to estimate variance components for animal models with several random effects using a derivative-free algorithm. *Genet. Select. Evol.* **21**:317–340.
- MEYER, K. 1991. Estimating variances and covariances for multivariate Animal Models by Restricted Maximum Likelihood. *Genet. Select. Evol.* **23**:67–83.
- MEYER, K. 1992. Bias and sampling covariances of estimates of variance components due to maternal effects. *Genet. Select. Evol.* **24**:487–509.
- MEYER, K. 1994. Derivative-Intense Restricted Maximum Likelihood Estimation of Covariance Components for Animal Models. *5th World Congr. Genet. Appl. Livest. Prod.* **Vol 18**:365–369.
- MEYER, K. 1995. Estimates of genetic parameters for mature weight of Australian beef cows and its relationship to early growth and skeletal measures. *Livest. Prod. Sci.* **44**:125–137.
- MEYER, K. 1996. An “average information” Restricted Maximum Likelihood algorithm for estimating reduced rank genetic covariance matrices or covariance functions for animal models with equal design matrices. *Genet. Select. Evol.* :(submitted).
- MEYER, K. AND SMITH, S.P. 1996. Restricted Maximum Likelihood estimation for animal models using derivatives of the likelihood. *Genet. Select. Evol.* **28**:23–49.
- NELDER, J.A. AND MEAD, R. 1965. A simplex method for function minimization. *Computer J.* **7**:147–151.
- POWELL, M.J.D. 1965. An efficient method for finding the minimum of a function of several variables without calculating derivatives. *Computer J.* **7**:155–162.
- PTAK, E. AND SCHAEFFER, L.R. 1993. Use of test day yields for genetic evaluation of dairy sires and cows. *Livest. Prod. Sci.* **34**:23–34.
- RISKA, B.W., ATCHLEY, W.R. AND RUTLEDGE, J.J. 1984. A genetic analysis of targeted growth in mice. *Genetics* **107**:79–101.
- SCHAEFFER, L.R. AND DEKKERS, J.C.M. 1994. Random regressions in animal models for test-day production in dairy cattle. *5th World Congr. Genet. Appl. Livest. Prod.* **Vol. 18**:443–446.
- WADE, K.M., QUAAS, R.L. AND VAN VLECK, L.D. 1993. Estimation of parameters involved in a first-order autoregressive process for contemporary groups. *J. Dairy Sci.* **76**:3033–3040.
- WIGGANS, G.R., GODDARD, M.E. AND MEYER, K. 1996. Test day model with 30 traits and a genetic (co)variance matrix of reduced rank. *J. Dairy Sci.* :(in preparation).