COVARIANCE STRUCTURES FOR QUANTITATIVE GENETIC ANALYSES

Karin Meyer

Animal Genetics and Breeding Unit[1], University of New England, Armidale, NSW 2351

SUMMARY

Parsimonious estimation of covariance matrices for multiple traits or repeated records is reviewed. Emphasis is placed on flexible models which do not require prior assumptions about the structure of covariance matrices, in particular parameterisations involving genetic principal components.

INTRODUCTION

Covariance matrices in quantitative genetic analyses have, by and large, been considered 'unstructured', i.e. for q random variables, there are $q(q + 1)/2$ distinct covariance components. This implies that the number of parameters to be estimated increases quadratically with the number of variables. Multivariate analyses involving more than a few traits have been hampered by computational problems. Recent improvements in computer hardware, both speed and memory available, have made analysis of larger data sets and models feasible. In addition, methodology to estimate covariance components has seen substantial progress. For restricted maximum likelihood (REML) estimation in particular, there are now fast and reliable algorithms available, capable of dealing with analyses involving higher dimensional covariance matrices among numerous traits or random regression coefficients.

However, computational problems aside, an inherent problem remains: with many parameters to be estimated we rarely have sufficient data to support accurate estimation of all the elements of unstructured covariance matrices. Attempts to improve efficiency of multivariate estimation fall into two broad categories, 'shrinkage' and estimation assuming covariance matrices have a certain structure. This paper reviews approaches to structured estimation relevant to quantitative genetic analyses.

'REPEATED' RECORDS AND FARTHER

Substantial impetus for structured estimation has come from the analysis of 'repeated' records, i.e. traits measured repeatedly per subject and, almost invariably, recorded along some continuous scale, i.e. along a trajectory. Most commonly the 'control' (co)variable is time (longitudinal data) or distance (spatial data). While we generally assume that records at points close together are similar and highly correlated, only in special cases are the assumptions of a 'repeatability' model, i.e. equal variances and correlations throughout, justified. Often we have measurements at numerous different, irregular spaced points and observations not represented by a grid, i.e. many 'missing' points per individual. Hence, the other extreme, fitting a 'full' multivariate model, which treats observations at all points along the scale as different traits, with an unstructured (US) covariance matrix is seldom feasible or desirable. This has motivated a number of approaches to model covariance matrices assuming an underlying structure; see Jennrich and Schluchter (1986), Wolfinger (1996) and, in a genetic context, Jaffrézic and Pletcher (2000) for reviews. On the one hand, this reduces the number of parameters, often dramatically so, and thus facilitates more efficient estimation. On the other hand, the structure assumed may provide more easily interpretable results or facilitate interpolation at missing values.

[1] AGBU is a joint venture of NSW Department of Primary Industries and the University of New England

Table 1. Stationary RF

CS	$r_{ij} = \rho$
AR(1)	$r_{ij} = \exp\{-\theta\Delta_{ij}\}$
GAU	$r_{ij} = \exp\{-\theta\Delta_{ij}{}^2\}$
DEX	$r_{ij} = \exp\{-\theta\Delta_{ij}{}^\kappa\}$

Parametric correlation structures. Widely used in many areas of applied statistics are models assuming a parametric correlation structure. Let $\mathbf{\Sigma} = \mathbf{SRS}$, with $\mathbf{\Sigma}$ of size $q \times q$ an US matrix of covariances, $\mathbf{S} = \text{Diag}\,\{\sigma_i\}$ the diagonal matrix of standard deviations, and $\mathbf{R}$ the corresponding correlation matrix with elements r_{ij}. Assume the i–th variable has been recorded at value t_i of the control variable, t, and let $r_{ij} = 1$ for all $i = j$.

Stationary. Common, simple correlation functions (RF) assume stationarity (e.g. Diggle *et al.* 1994), i.e. that the correlation between points t_i and t_j depends only on the lag, $\Delta_{ij} = |t_i - t_j|$, rather than t_i or t_j. Let $\rho > 0$ denote the lag 1 correlation and $\theta = -\ln(\rho)$. Well known structures determined by a single parameter are compound symmetry (CS), first-order autoregressive (AR(1)) and Gaussian (GAU) RFs, shown in Table 1. These are special cases of the 'damped' exponential (DEX) (Muñoz *et al.* 1992) for $\kappa = 0, 1, 2$. An alternative form of AR(1) is the autocorrelation function, $r_{ij} = \rho^{\Delta_{ij}}$. AR(1), GAU and DEX define correlations which decay with increasing lag. Other, less common single parameter RFs have been considered by Pletcher and Geyer (1999).

Non-stationary. In other cases we cannot assume equidistant records to be equicorrelated. A simple extensions of the above RF to account for non-stationarity is a deformation of the control variable t. Related to time series are the so-called ante-dependence models, AD(s). These assume that the i–th 'repeated' record per individual depends (at most) on the s preceding observations. In its unstructured form, the corresponding RF has $sq - s(s + 1)/2$ parameters, which are the elements the first s sub-diagonals of $\mathbf{R}$. The remaining elements of $\mathbf{R}$ are a function of these parameters. For $s = 1$, these are $r_{ij} = \prod_{l=i+1}^{j-1} r_{l,l+1}$ (for $i = 1, q$ and $j = i + 2, q$), e.g. $r_{13} = r_{12}r_{23}$ and $r_{14} = r_{12}r_{23}r_{34}$. This gives $\mathbf{\Sigma}^{-1}$ which is banded, with only the first s sub-diagonals non-zero. Structured ante-dependence (SAD(s)) models (Núñez-Antón and Zimmerman 2000) impose a functional relationship on the parameters of an AD(s) model. The RF defined by SAD(s) has $2s$ parameters, ρ_n and γ_n for $n = 1, s$. Correlations on the n–th sub-diagonal of $\mathbf{R}$ are then given as $r_{ij} = \exp\{\ln(\rho_n)\Delta^n_{ij}\}$ for $i = n + 1, q$ and $j = i - n$, with $\Delta^n_{ij} = f(t_i, \gamma_n) - f(t_j, \gamma_n)$. The function $f(\cdot)$ represents a Box-Cox transformation, i.e. $f(t, \gamma) = \ln(\gamma)$ for $\gamma = 0$ and $f(t, \gamma) = (t^\gamma - 1)/\gamma$ otherwise. For $s = 1$ and $\gamma = 1$, SAD(1) reduces to AR(1).

An alternative, encompassing many stationary and non-stationary RFs as special cases, is the 'generalised autoregressive parameter' (GARP) model (Pourahmadi 1999). This models the off-diagonal elements of $\mathbf{T}$ for $\mathbf{T\Sigma T}' = \mathbf{D}$, with $\mathbf{D}$ is the diagonal matrix of 'innovation' variances. The unit, lower triangular matrix $\mathbf{T}$ has off-diagonal elements, u_{ij} which are (negative values of) the regression coefficient predicting the i–th record from the $i - 1$ preceding observations. These are unconstrained can thus be modelled as a function of some covariates $\mathbf{z}_{ij}$ and parameters $\boldsymbol{\gamma}' = (\gamma_1, \cdots, \gamma_s)$, $u_{ij} = g(\mathbf{z}_{ij}, \boldsymbol{\gamma})$.

More dimensions. Generalisations to more than 1 dimension are available in the literature, in particular in the field of geostatistics (e.g. Wackernagel 2003). For example, Zimmermann and Harville (1991) discuss spherical, exponential and Gaussian covariance functions for random fields, and Gaspari and Cohn (1999) consider RFs with up to 3 dimensions.

'Character process' models. Generally, parametric correlation structures are used to model within-subject or residual covariance matrices. Pletcher and Geyer (1999) suggested to model both the genetic and residual covariance matrices for longitudinal data in this way, dubbing the resulting models character process (CP) models. CP models can involve any of the RFs described above. Extensions to repeated records for multiple traits have been described by Jaffrézic *et al.* (2004).

Variance functions. Similarly, changes in variances with t can be modelled parsimoniously through a (link) function. Most commonly, this is done parameterising to logarithmic values, thus removing the need for constraints, i.e. $\ln(\sigma_i^2) = v(\mathbf{z}_i, \boldsymbol{\upsilon})$ with $\mathbf{z}_i$ a vector of covariates and $\boldsymbol{\upsilon}$ the vector of parameters. Simple variance functions, $v(\cdot)$, are step functions or low degree polynomials of t. Alternatively, trigonometric or spline functions may be appropriate to model periodic or more arbitrary patterns of change. In most cases, $\mathbf{z}_i$ will comprise functions of t only, but more complicated dependencies are readily accommodated. In mixed model analyses, for instance, this may involve conditioning on fixed effects, i.e. effectively fitting a 'double' mixed model (Ruppert *et al.* 2003).

Random regression models. A less parsimonious, but more flexible alternative is the covariance structure defined in random regression (RR) models; see Meyer and Kirkpatrick (2005b) for a detailed review. RR models imply that traits are 'function-valued' and that these functions can be represented as regression equations, $g(t) = \sum_{j=1}^{k} \alpha_j \phi_j(t) = \boldsymbol{\alpha}' \boldsymbol{\phi}(t)$. The underlying idea is that any trajectory can modelled as the weighted sum of a set of basis functions, $\boldsymbol{\phi}(t) = \{\phi_j(t)\}$. Suitable bases are, for instance, orthogonal polynomials, trigonometric or spline functions. Conceptually, there are infinitely many functions in the set, but, in practice, a small number of k functions often suffices for a good approximation. This allows non-linear trajectories to be fitted within the standard, linear mixed model.

In particular, we can model the trajectory for any random effect by fitting a corresponding set of RR coefficients, $\boldsymbol{\alpha}_i = \{\alpha_{ij}\}$, for each level i. Let $\mathrm{V}(\boldsymbol{\alpha}) = \mathbf{K}$ denote the $k \times k$ covariance matrix among the RR coefficients. The covariance between two measures, at points t_i and t_j, is then given by the covariance function $\mathcal{G} = \sum_{l=1}^{k} \sum_{n=1}^{k} K_{ln} \phi_l(t_i) \phi_n(t_j)$, with K_{ln} the ln–th element of $\mathbf{K}$. In turn, this gives $\boldsymbol{\Sigma} = \boldsymbol{\Phi}' \mathbf{K} \boldsymbol{\Phi}$ with $\boldsymbol{\Phi} = \{\boldsymbol{\phi}(t_1) \cdots \boldsymbol{\phi}(t_q)\}$ the $k \times q$ matrix of basis functions evaluated for the q points represented in $\boldsymbol{\Sigma}$. This is equivalent to the covariance function in the 'infinite-dimensional' model proposed by Kirkpatrick *et al.* (1990), i.e. RR models provide a convenient way to estimate such functions.

Generally, $\mathbf{K}$ is considered to be US, i.e. the $q(q+1)/2$ elements of $\boldsymbol{\Sigma}$ are modelled by $p = k(k+1)/2$ parameters. If $\mathbf{K}$ is estimated at reduced rank, $m < k$, this is reduced to $p = m(2k - m + 1)/2$. In special cases, a more rigid structure can be imposed on $\mathbf{K}$. For instance, when fitting a RR on natural, cubic smoothing splines only the intercept and linear terms are assumed correlated and all quadratic terms are considered to be *i.i.d.* distributed, yielding a low number of $p = 4$ parameters (White *et al.* 1999).

RR analyses in quantitative genetics usually involve covariance functions for at least 2 sources of variation, individuals' additive genetic and permanent environmental effects. In addition, temporary environmental effects are considered. These are generally considered independently distributed and often modelled through a variance function, as described above. Extensions to multiple traits or more than one control variable are conceptually straightforward but can be complex in practical applications; see Meyer and Kirkpatrick (2005b) for some discussion.

Other. Other notable approaches to model covariances among repeated records include a factor-analytic (FA) structure for $\boldsymbol{\Sigma}$, and methods of 'covariance selection' (Dempster 1972) aimed at identifying zero elements of the 'concentration' matrix, $\boldsymbol{\Sigma}^{-1}$, or its Cholesky factor (Smith and Kohn 2002).

PRINCIPAL COMPONENTS AND BEYOND

In a more general scenario, we want to estimate covariances among numerous, correlated traits where there is no 'natural' ordering and – assuming few, if any repeated records – no obvious structure. In special cases, there may be prior knowledge or restrictions for individual elements of $\boldsymbol{\Sigma}$, e.g. that a covariance between two traits is zero or that a correlation has an absolute value of unity. In addition, some of the 'covariance' selection procedures (see above) may be applicable. Parsimonious estimation for this case in general, however, requires a different approach.

Principal components (PC) have long been used to summarise multivariate information in a number of areas, going back to Hotelling (1933) and earlier. They are based on an eigen-decomposition of the covariance matrix. Moreover, eigen-values and -vectors are ubiquitous in the statistical literature on matrices. Yet, apart from use of the 'canonical transformation' to reduce computational requirements of multi-trait analyses, there has been little interest in parameterisations involving these quantities.

So far, genetic PCs have generally been estimated in two steps, carrying out an eigen-decomposition of an initial, US estimate of the genetic covariance matrix. Recently, Kirkpatrick and Meyer (2004) advocated direct estimation of PCs, showing that this involved little more than a straightforward reparameterisation of the standard linear, mixed model.

Principal components. Let $\mathbf{\Sigma} = \mathbf{E\Lambda E}'$ represent the eigen-decomposition of $\mathbf{\Sigma}$, with $\mathbf{\Lambda} = \text{Diag}\{\lambda_i\}$ the diagonal matrix of eigen-values, and $\mathbf{E} = \{\mathbf{e}_i\}$ the matrix whose columns are the corresponding eigen-vectors. $\mathbf{E}$ is orthogonal, i.e. $\mathbf{EE}' = \mathbf{I}$. Assume the λ_i and $\mathbf{e}_i$ are in descending order of λ_i, i.e. $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_q$. Further, let $\mathbf{\Sigma} = \text{V}(\mathbf{v})$ with $\mathbf{v}$, of length q, a vector of random variables. The i–th PC is then given by $\mathcal{P}_i = \mathbf{e}_i'\mathbf{v}$, has variance λ_i and is uncorrelated with all other PCs. Moreover, $\mathcal{P}_i$ is the linear function of $\mathbf{v}$ which explains most variation, given $\mathcal{P}_1$ to $\mathcal{P}_{i-1}$.

Reduced rank estimation. Hence, any PCs with corresponding eigen-values close to zero contribute little and can be omitted with negligible loss of information. This is the principle underlying the use of PC analysis as a dimension reduction technique. Considering the leading m PCs only reduces the number of effects in a mixed model analysis and thus computational requirements and sampling errors. Let $\mathbf{E}_m$ denote $\mathbf{E}$ truncated to the first m columns and $\mathbf{\Lambda}_m$ the corresponding sub-matrix of $\mathbf{\Lambda}$. This yields $\mathbf{\Sigma}_m = \mathbf{E}_m\mathbf{\Lambda}_m\mathbf{E}_m'$, i.e a parameterisation of $\mathbf{\Sigma}$ which defines a reduced rank matrix . It involves $p = m(2q - m + 1)/2$ parameters, m values λ_i and $m(2q - m - 1)/2$ elements of $\mathbf{E}_m$. The remaining $m(m + 1)/2$ elements of $\mathbf{E}_m$ are determined by the orthogonality constraints on its columns.

More useful forms for estimation are $\mathbf{\Sigma}_m = \mathbf{\Gamma}_m\mathbf{\Gamma}_m'$ with $\mathbf{\Gamma}_m = \mathbf{E}_m\mathbf{\Lambda}_m^{1/2}$, or $\mathbf{\Sigma}_m = \mathbf{L}_m\mathbf{L}_m'$ where $\mathbf{L}_m$ denotes the Cholesky factor of $\mathbf{\Sigma}$ (obtained pivoting on the largest diagonal), truncated to the first m columns. This utilises that, for $\mathbf{\Sigma} = \mathbf{LL}'$, $\mathbf{L} = \mathbf{E\Lambda}^{1/2}\mathbf{U}'$ with $\mathbf{UU}' = \mathbf{I}$, i.e. that the columns of $\mathbf{L}$ can be interpreted as rotated PCs (Smith *et al.* 2001). $\mathbf{L}$ has $m(m-1)/2$ elements of zero (above the diagonal), i.e. the rotation is a convenient way of imposing the necessary constraint on the number of parameters.

Factor-analytic models. Closely related, but with a somewhat different emphasis, is the assumption of a FA structure for $\mathbf{\Sigma}$, i.e. $\mathbf{\Sigma}_m^+ = \mathbf{\Gamma}_m\mathbf{\Gamma}_m' + \mathbf{\Psi}$, where $\mathbf{\Psi} = \text{Diag}\{\psi_i\}$ represents the matrix of 'specific' variances ψ_i, for $i = 1, q$. This increases the number of parameters to $p = q(m + 1) - m(m - 1)/2$, and thus imposes a restriction on m, as p cannot exceed the value in the US case, i.e. $p \leq q(q + 1)/2$. While PC analysis is concerned with identifying variables which successively explain maximum amounts of variation, factor analysis attempts to attribute covariances between variables to common factors. This implies a latent variable model $\mathbf{v} = \mathbf{\Gamma}_m\mathbf{f} + \boldsymbol{\delta}$, with m factors $\mathbf{f} \sim N(\mathbf{0}, \mathbf{I}_m)$ and residuals $\boldsymbol{\delta} \sim N(\mathbf{0}, \mathbf{\Psi})$. Elements of $\mathbf{\Gamma}_m$ are generally referred to as 'factor loadings' .

Mixed model implementations. REML estimation of a covariance matrix $\mathbf{\Sigma}$ assuming a FA structure might simply involve a 'standard' multivariate analysis for q traits, fitting a corresponding random effect, $\mathbf{v}$ with q effects per level, and maximising the corresponding likelihood imposing the structural constraint on $\mathbf{\Sigma}$. Alternatively, we might fit the components of the latent model separately, i.e. fit m

factors $\mathbf{f}$ and q effects δ_i per level of the random effect, to yield an 'extended FA' model (Thompson *et al.* 2003). While this increases the total number of effects, this is advantageous as the δ_i are independent, resulting in sparse mixed model equations. If specific effects and their variances are considered to be zero, this collapses to the reduced rank model considered by Meyer and Kirkpatrick (2005a) to directly estimate the leading m PCs only.

Common principal components. When there are several, independent groups of individuals with the same q variables recorded, we may want to model the corresponding covariance matrices, $\mathbf{\Sigma}_j$ for $j = 1, g$, simultaneously. Assuming a common correlation structure, $\mathbf{R}_1 = \cdots = \mathbf{R}_g$, reduces the number of parameters from $p = gq(q+1)/2$ for the full, US case to $p = gq + q(q-1)/2$. If we can assume variances to be proportional, $\mathbf{\Sigma}_j = \tau_j \mathbf{\Sigma}_1$ for $j = 2, g$, $p = q(q+1)/2 + g - 1$.

Alternatively, we can base classification of the degree of similarity of the $\mathbf{\Sigma}_j$ on their PCs. The common PC (CPC) model assumes all matrices have the same eigen-vectors, $\mathbf{E}_1 = \cdots = \mathbf{E}_g$, but allows for different eigen-values (Flury 1984). This implies that all $\mathbf{\Sigma}_j$ can be simultaneously diagonalised. The number of parameters is the same as for the common correlation model, $p = gq + q(q-1)/2$. Less restrictive is the a partial CPC structure, where we assume that the first $m \leq q - 2$ eigen-vectors are the same in all groups, i.e. $\mathbf{E}_{m\,1} = \cdots = \mathbf{E}_{m\,g}$, while the remaining eigen-vectors are group specific. This gives $p = (gq(q+1) - m(g-1)(2q-m-1))/2$. The so-called 'Flury hierarchy' has had considerable uptake in evolutionary biology to characterise differences in genetic covariance matrices between species or its change over time; see Phillips and Arnold (1999), or Steppan *et al.* (2002) for a review.

While generally described for a full rank scenario, there are obvious extensions of the CPC models to reduced rank estimation, with corresponding reductions in the number of parameters. For instance, we might have $m < q$ PCs of interest and $k < m$ CPCs. The parameters then consist of gm eigen-values, $k(2q - k - 1)/2$ elements of the common and $g(m - k)(2q - m - k - 1)/2$ elements of the non-common eigen-vectors. Other authors consider common subspaces of PCs. In particular, Schott (1999) examines the case where the groups have 'almost' common subspace of dimension $m + n$ with n a small number.

Similarly, CPCs can be useful in modelling patterned covariance matrices which can be partitioned into blocks of similar matrices. For instance, Klingenberg *et al.* (1996) investigate q morphological traits measured at each of g growth stages. Assuming CPCs, the sub-matrices for stages i and j are $\mathbf{\Sigma}_{ij} = \mathbf{E}\mathbf{\Lambda}_{ij}\mathbf{E}'$ with $\mathbf{E}$ the matrix of common eigen-vectors. This reduces the number of parameters from $p = gq(gq+1)/2$ in the US case to $p = q(g(g+1) + q - 1)/2$. Extensions to reduced rank, partial CPC models for other types of multivariate repeated records or function-valued traits are readily conceived, but have not been considered so far.

Related models. Again utilising the relationship, $\mathbf{L} = \mathbf{E}\mathbf{\Lambda}^{1/2}\mathbf{U}$, between the Cholesky factor and eigen-vectors of a matrix, Pourahmadi *et al.* (2007) propose a common GARP model. This allows a similar hierarchy to the (partial) CPC models, and, due to the unconstrained nature of its parameters offers the scope for finer hierarchies as well as computational advantages. An even more gradual change in the number of parameters is afforded by 'spectral' model of Boik (2002), which comprises a number of CPC and common subspace models as special cases. It achieves flexibility by modelling the eigen-values and -vectors of several matrices simultaneously, allowing for relationships between eigen-values across groups in addition to communality of PCs or their spaces.

DISCUSSION

Estimation of high(er) dimensional genetic covariance matrices will often require assumptions of an underlying, structural relationship between individual covariance components. There are a substantial number of alternatives. While structured estimation is likely to make more efficient use of the data available, there is clearly a trade-off between parsimony, potential bias and complexity of analyses, and judicious choices need to be made.

Model selection. With likelihood based estimation, a likelihood ratio test (LRT) is an obvious way to compare the fit of models assuming different covariance structures. A less often used alternative is a score test, which also allows for one-sided hypothesis testing (Verbeke and Molenberghs 2003). LRTs are limited to nested models, and care must be taken to account for boundary conditions when testing hypotheses involving parameter values at their limits (Self and Liang 1987), e.g. whether an eigen-value or variance component is zero. Moreover, the LRT is known to favour the most detailed model. Hence, the so-called information criteria, which adjust for the number of parameters fitted, are widely used alternatives, in particular, the Akaike (AIC) and Bayesian (BIC) information criterion; see Burnham and Anderson (2004) for a comprehensive exposé. REML forms are $-2\ln\mathcal{L}+2p$ and $-2\ln\mathcal{L}+\ln(d)\,p$, respectively, with $\mathcal{L}$ the likelihood and d the degrees of freedom (Wolfinger 1993), i.e. BIC generally involves a considerably more stringent penalty for higher numbers of parameters than AIC.

While these statistics are commonly reported for and used with mixed models, it has to be noted that they were originally derived in the context of regression analyses. Concern has been voiced that some of the underlying assumptions are violated when the model of analysis includes random effects (Ripley 2004). Vaida and Blanchard (2005) propose a conditional AIC for mixed effect models, based on the conditional likelihood given the random effects. Further work is needed to evaluate how robust and consistent information criterion based model selection procedures are, in particular with reference to discriminating between covariance structures.

How many PCs ? A crucial question for factor or reduced rank analysis is how many PCs should be fitted or, equivalently, which eigenvalues are different from zero. In addition to the likelihood based criteria, a number of tests addressing this question are in use. These range from Lawley's (1956) adaptation of Bartlett's test to heuristic procedures like the scree test (Cattell 1966). Disconcertingly, limited simulation studies available (Jackson 1993; Peres-Neto *et al.* 2005) show inconsistent results, indicating that these methods should be applied with caution. For a half-sib design, Hine and Blows (2006) report a tendency for reduced rank REML analyses together with an AIC based choice, to underestimate the number of PCs of the genetic covariance matrix at low heritabilities.

Perils of parsimony. Imposing any constraints on covariance matrices introduces bias. A typical example is the bias in REML estimates of US covariance matrices, generated by forcing them to be non-negative definite. Parsimonious estimation assuming covariance matrices are structured reduces sampling variances in comparison to the US scenario. However, the mean square error of estimation is reduced only if any corresponding bias created is small or negligible. Clearly the bias acceptable depends on the particular analysis and data available.

For analyses involving more than one covariance matrix, structured estimation may create biased partitioning of variation. Loosely speaking, any excess variance not accommodated by the structure imposed is likely to be 'picked up' in a covariance matrix subject to less restrictions. Gilmour and Thompson (2006) recommend that all 'strata' should be fitted at the same degree of complexity. However, this may not suffice. Jaffrézic *et al.* (2002) encountered substantially inflated estimates of genetic variances when fitting a CP model for longitudinal data. Providing an 'outlet' for permanent environmental variation not modelled by the CP by allowing residual effects to be correlated and to have heterogeneous variances alleviated the problem. This emphasises that care must be taken when fitting highly parsimonious and restrictive covariance structures.

Similarly, as shown by Meyer and Kirkpatrick (2007), fitting too few PCs in reduced rank analyses can substantially bias estimates of the leading PCs. For a simple animal model, ignoring genetic PCs with non-zero eigenvalues is likely to yield underestimates of the genetic and overestimates of the residual variances. For models with additional random effects, the resulting pattern is less readily predictable and depends on the relative numbers of PCs fitted and assumptions on the structure of the residual covariance matrix. For reduced rank analyses of longitudinal data, it appears prudent to fit more PCs for permanent environmental than genetic effects (Meyer 2005).

Structure and shrinkage. Early work on efficient multivariate estimation, most notably by Stein (see e.g. Dey and Srinivasan (1985) for references), has considered 'shrinkage' estimators. In particular, regression of the eigenvalues of a matrix towards their mean has been suggested. In a genetic context, Hayes and Hill (1981) advocated this approach to improve sampling properties of selection indexes. Recently, there has been renewed interest in such estimators. In particular, Daniels and Kass (2001) consider estimation which combines some 'squeezing' of eigenvalues with shrinkage of the estimate towards some specified, parametric structure, and Daniels and Pourahmadi (2002) extend this work to a Bayesian setting. Such 'data-driven' shrinkage can alleviate the ill effects of miss-specifying an underlying structure. Future work should examine its applicability and properties for genetic models with more than one covariance matrix.

CONCLUSIONS

Structured estimation of covariance matrices provides a means of modelling patterns of dispersion in more than a few dimensions parsimoniously. In particular, reduced rank estimation considering the leading principal components only requires few initial assumptions, can reduce computational requirements, and is highly appealing. It is likely to see increasing use in future.

REFERENCES

Boik, R. J. (2002) *Biometrika* **89**:159.

Burnham, K. P. and Anderson, D. R. (2004) *Sociol. Meth. Res.* **33**:261.

Cattell, R. B. (1966) *Multiv. Behav. Res.* **1**:245.

Daniels, M. and Kass, R. E. (2001) *Biometrics* **57**:1173.

Daniels, M. J. and Pourahmadi, M. (2002) *Biometrika* **89**:553.

Dempster, A. P. (1972) *Biometrics* **28**:157.

Dey, D. and Srinivasan, C. (1985) *Ann. Stat.* **13**:1581.

Diggle, P. J., Liang, K. Y. and Zeger, S. L. (1994) *Analysis of Longitudinal Data*. Oxford Science Publications, Clarendon Press, Oxford.

Flury, B. N. (1984) *J. Amer. Stat. Ass.* **79**:892.

Gaspari, G. and Cohn, S. (1999) *Quart. J. Roy. Meteor. Soc.* **125**:723.

Gilmour, A. R. and Thompson, R. (2006) *CD-ROM Eighth World Congr. Genet. Appl. Livest. Prod.* Communication No. 25–02.

Hayes, J. F. and Hill, W. G. (1981) *Biometrics* **37**:483.

Hine, E. and Blows, M. W. (2006) *Genetics* **173**:1135.

Hotelling, H. (1933) *J. Educ. Psych.* **24**:417.
Jackson, D. A. (1993) *Ecology* **74**:2204.
Jaffrézic, F. and Pletcher, S. D. (2000) *Genetics* **156**:913.
Jaffrézic, F., Thompson, R. and Pletcher, S. D. (2004) *Genetics* **168**:477.
Jaffrézic, F., White, I. M. S., Thompson, R. and Visscher, P. M. (2002) *J. Dairy Sci.* **84**:968.
Jennrich, R. I. and Schluchter, M. D. (1986) *Biometrics* **42**:805.
Kirkpatrick, M., Lofsvold, D. and Bulmer, M. (1990) *Genetics* **124**:979.
Kirkpatrick, M. and Meyer, K. (2004) *Genetics* **168**:2295.
Klingenberg, C. P., Neuenschwander, B. E. and Flury, B. D. (1996) *Syst. Biol.* **45**:135.
Lawley, D. N. (1956) *Biometrika* **43**:128.
Meyer, K. (2005) *Austr. J. Exp. Agric.* **45**:847.
Meyer, K. and Kirkpatrick, M. (2005a) *Genet. Select. Evol.* **37**:1.
Meyer, K. and Kirkpatrick, M. (2005b) *Phil. Trans. R. Soc. B* **360**:1443.
Meyer, K. and Kirkpatrick, M. (2007) *Proc. Ass. Advan. Anim. Breed. Genet.* **17**:154.
Muñoz, A., Carey, V., Schouten, J. P., Segal, M. and Rosner, B. (1992) *Biometrics* **48**:733.
Núñez-Antón, V. and Zimmerman, D. L. (2000) *Biometrics* **56**:699.
Peres-Neto, P. R., Jackson, D. A. and Somers, K. M. (2005) *Comp. Stat. Dat. Anal.* **49**:974.
Phillips, P. C. and Arnold, S. J. (1999) *Evolution* **53**:1506.
Pletcher, S. D. and Geyer, C. J. (1999) *Genetics* **153**:825.
Pourahmadi, M. (1999) *Biometrika* **86**:677.
Pourahmadi, M., Daniels, M. J. and Park, T. (2007) *J. Multiv. Anal.* **98**:569.
Ripley, B. D. (2004) In *Methods and Models in Statistics in Honour of Professor John Nelder, FRS*. Imperial College Press, London, pp. 155–170.
Ruppert, D., Wand, M. P. and Carroll, R. J. (2003) *Semiparametric Regression*. Cambridge University Press, New York.
Schott, J. R. (1999) *Biometrika* **86**:899.
Self, S. G. and Liang, K. Y. (1987) *J. Amer. Stat. Ass.* **82**:605.
Smith, A. B., Cullis, B. R. and Thompson, R. (2001) *Biometrics* **57**:1138.
Smith, M. and Kohn, R. (2002) *J. Amer. Stat. Ass.* **97**:1141.
Steppan, S. J., Phillips, P. C. and Houle, D. (2002) *Trends Ecol. Evol.* **17**:320.
Thompson, R., Cullis, B. R., Smith, A. B. and Gilmour, A. R. (2003) *Austr. New Zeal. J. Stat.* **45**:445.
Vaida, F. and Blanchard, S. (2005) *Biometrika* **92**:351.
Verbeke, G. and Molenberghs, G. (2003) *Biometrics* **59**:254.
Wackernagel, H. (2003) *Multivariate Geostatistics : An Introduction with Applications*. Springer Verlag, New York, 3rd edition.
White, I. M. S., Thompson, R. and Brotherstone, S. (1999) *J. Dairy Sci.* **82**:632.
Wolfinger, R. D. (1993) *Comm. Stat. - Simul. Comp.* **22**:1079.
Wolfinger, R. D. (1996) *J. Agric. Biol. Env. Stat.* **1**:205.
Zimmermann, D. L. and Harville, D. A. (1991) *Biometrics* **47**:223.

ACKNOWLEDGEMENTS

This work was supported by Meat and Livestock Australia under grant BFGEN.100B.