Perils of parsimony: Properties of reduced rank estimates of genetic covariance matrices

Karin Meyer* and Mark Kirkpatrick†

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

and

† Section of Integrative Biology, 1 University Station C-0930, University of Texas, Austin, Texas 78712, USA

[1]a joint venture with NSW Agriculture

Running head: Reduced rank covariance matrices

Keywords: Genetic covariance matrix, estimation, reduced rank, bias

Corresponding author: Karin Meyer,

Animal Genetics and Breeding Unit,

University of New England,

Armidale NSW 2351,

Australia

E-mail : kmeyer@didgeridoo.une.edu.au

Phone : +61 2 6773 3331

Fax : +61 2 6773 3266

ABSTRACT

Eigenvalues and eigenvectors of covariance matrices are important statistics for multivariate problems in many applications, including quantitative genetics. Estimates of these quantities are subject to different types of bias. This paper reviews and extends the existing theory on these biases, considering a balanced one-way classification and restricted maximum likelihood estimation. Biases are due to the spread of sample roots, and from ignoring selected principal components when imposing constraints on the parameter space, to ensure positive semi-definite estimates or to estimate covariance matrices of chosen, reduced rank. In addition, it is shown that reduced rank estimators which consider only the leading eigenvalues and -vectors of the 'between group' covariance matrix may be biased due to selecting the wrong subset of principal components. In a genetic context, with groups representing families, this bias is inverse proportional to the degree of genetic relationship among family members, but is independent of sample size. Theoretical results are supplemented by a simulation study, demonstrating close agreement between predicted and observed bias for large samples. It is emphasized that the rank of the genetic covariance matrix should be chosen sufficiently large to accommodate all important genetic principal components, even though, paradoxically, this may require including a number of components with negligible eigenvalues. A strategy for rank selection in practical analyses is outlined.

INTRODUCTION

Traits of interest in quantitative genetics are seldom independent of each other. Hence, in analyses of 'complex' phenotypes it is desirable to consider all components simultaneously, in particular when considering the effects of selection and its impact on evolution (Blows and Walsh, 2008). However, analyses to estimate genetic parameters are often limited to a few traits only. This can be attributed to the burden imposed by multivariate estimation, due both to computational requirements and limitations and the need for sufficiently large data sets to support accurate estimation of the numerous parameters involved.

By and large, covariance matrices are considered to be unstructured, i.e. for q traits of interest we have $q(q+1)/2$ distinct variance and covariance components among them. In a genetic context, there are at least two covariance matrices to be determined, namely the covariance matrix due to additive genetic effects and the corresponding matrix due to residual effects. This yields $q(q+1)$ parameters to be estimated, i.e. the number of parameters increases quadratically with the number of traits considered. Recently, improvements in computing facilities together with advances in the implementation of modern inference procedures, such as residual or restricted maximum likelihood (REML), have made routine multivariate analyses involving numerous traits and large data sets feasible. In addition, availability of corresponding software, specialised towards quantitative genetic analyses fitting the so-called 'animal model', has made analyses conceptually straightforward, even for scenarios with complex pedigrees, many fixed effects, additional random effects or arbitrary patterns of missing observations.

Yet, the 'curse of dimensionality' remains. This has kindled interest in estimation imposing a structure, in particular for genetic covariance matrices; see Meyer (2007a) for a

recent review. Principal component (PC) analysis is a widely used method to summarise multivariate information, dating back as far as Hotelling (1933) and Pearson (1901) (both reprinted in Bryant and Atchley (1975)). Moreover, PCs are invaluable in reducing the dimension of analyses, i.e. the number of variables to be considered. For a set of q variables, the PCs are the q linear combinations of the variables which are independent of each other, and successively explain a maximum amount of variation. Hence, if the $m+1-$th PC explains negligible variation, PCs $m+1$ to q convey little information which is not already contained in PCs 1 to m. We can then safely ignore PCs $m+1$ to q, reducing the number of variables from q to m. In many applications, m can be considerably smaller than q.

The PCs for a set of variables are estimated from the eigen-decomposition of the corresponding covariance matrix: The eigenvectors provide the linear functions of the original variables while the corresponding eigenvalues give the amount of variance explained by each PC. PCs are usually given in descending order of the eigenvalues and the matrix of eigenvectors is scaled to be orthonormal. The latter implies that the matrix of eigenvectors (with q^2 distinct elements) is described by $q(q-1)/2$ parameters, with the remaining $q(q+1)/2$ elements defined by the orthonormality constraints. Ignoring PCs $m+1$ to q thus reduces the number of parameters to describe the covariances among the q traits to $m(2q-m+1)/2$, comprising m eigenvalues and the $m(2q-m-1)/2$ 'free' elements of the first m eigenvectors. The resulting covariance matrix, of size $q \times q$, has reduced rank m.

Closely related technically, but with a somewhat different underlying concept is factor analysis (FA). While PC analysis aims at identifying variables which explain maximum variation, FA is primarily concerned with finding the common 'factors' which cause covariances between traits. Like PCs, the predictors of such factors are independent, linear combinations of the original traits. In addition, FA allows for specific effects which

are generally assumed to be uncorrelated. More importantly, FA implies a latent variable model for the original traits – modelling each as the sum of common factors and specific effects – while PC analysis does not. In the general case, specific variances for all q traits are assumed to be non-zero and different from each other. Considering a FA model with m factors, this yields a full rank covariance matrix modelled by $q + m(2q - m + 1)/2$ parameters. This quantity cannot exceed the number in the unstructured case, $q(q + 1)/2$, which limits the maximum number of common factors which can be fitted. If specific effects are assumed to have zero variance, the FA model with m factors yields the same, reduced rank covariance structure as considering the leading m PCs only.

Hence FA models are readily incorporated in the linear mixed models commonly fitted for the estimation of genetic parameters. This enables direct estimation of the leading principal components (i.e. factors, assuming no specific effects) only of genetic covariance matrices, as proposed by Kirkpatrick and Meyer (2004). Here 'direct' refers to estimation directly from the data instead of the more customary two-step procedure which involves estimation of the complete, unstructured genetic covariance matrix before considering its eigen-decomposition. For highly correlated traits with PCs which have eigenvalues close to zero, such reduced rank analyses avoid estimation of unnecessary parameters. This not only makes more efficient use of the data, but can also reduce computational requirements markedly over those in a full rank analysis. Suitable REML algorithms have been described by Thompson *et al.* (2003), Meyer and Kirkpatrick (2005) and Meyer (2008), and Bayesian estimation via Gibbs sampling has been outlined by Los Campos and Gianola (2007).

Classic PC analysis considers a single covariance matrix at a time. Due to their orthogonality, we can alter the number of PCs fitted for a single matrix successively, i.e. estimates of the $i-$th PC are expected to remain more or less the same when increasing or reduc-

ing the number of PCs considered. For quantitative genetic analyses, however, with at least two covariance matrices estimated simultaneously, this is not necessarily the case. If genetic PCs with non-negligible eigenvalues are omitted while estimating a full rank residual covariance matrix, genetic covariances can be be partitioned into the environmental components, leading to biassed estimates. This paper examines the properties of estimates of genetic covariance matrices and their eigenvalues and -vectors from reduced rank analyses, showing that up to three sources of bias may affect estimates.

THEORETICAL CONSIDERATIONS

In the following, we present a brief review of pertinent statistical literature on multivariate estimation, focusing on problems associated with sampling variation and constraining estimates to the parameter space. Subsequently, we examine the bias arising from estimating genetic covariance matrices of reduced rank.

Bias due to sampling variance

Spread of sample roots: Consider N sets of observations for q traits from a multivariate normal distribution with population covariance matrix $\mathbf{\Sigma}$. The sample covariance matrix $\mathbf{S}$, estimated from the sums of squares and cross-products among observations, then has a central Wishart distribution. It is well known that the eigenvalues (latent roots) of such sample covariance matrix are spread further than the population values. This results in an upward bias of estimates of the largest eigenvalues and a downward bias of estimates of the smallest values while their mean is expected to be unbiased. Let λ_i denote the $i-$th eigenvalue of $\mathbf{\Sigma}$, and ℓ_i the $i-$th eigenvalue of $\mathbf{S}$. For λ_i distinct from all other roots,

Lawley (1956) gave an expected value for the corresponding sample value of

$$\mathrm{E}\left[\ell_i\right] = \lambda_i\left(1 + \frac{1}{N}\sum_{\substack{j=1\\ j\neq i}}^{q}\frac{\lambda_j}{\lambda_i - \lambda_j}\right) + \mathrm{O}\left(\frac{1}{N^2}\right) \tag{1}$$

with $\mathrm{O}(1/N^2)$ a residual term proportional to $1/N^2$. (Eq. 1) shows that the bias decreases inversely proportionally to the sample size, becoming small as q/N becomes negligible. Furthermore, bias tends to be the more pronounced the closer together population roots are. Lawley (1956) suggested that an estimate of λ_i less biassed than ℓ_i might be obtained from (Eq. 1), substituting sample for population values, as $\hat{\lambda}_i = \ell_i\left(1 - \left(\sum_{j\neq i}^{q}\ell_j/(\ell_i - \ell_j)\right)/N\right)$.

An obvious alternative to improve estimates of covariance matrices for small samples or of high dimensions is to squeeze the sample eigenvalues together. A number of improved estimators have been proposed, based on minimizing a loss or risk function. These differ in the functions utilized and resulting properties, e.g. whether the original order of eigenvalues is maintained or whether the resulting estimates are guaranteed to be non-negative; see Muirhead (1987) for a review of early work and Srivastava and Kubokawa (1999) for more recent references. Minimizing the squared error loss to determine an optimal amount of shrinkage, Ledoit and Wolf (2004) derived an estimator which regresses sample eigenvalues towards their mean and yields a weighted combination of the sample covariance matrix and an identity matrix. While this has seen diverse applications, including the analysis of high-dimensional genomic data (Schäfer and Strimmer, 2005), Daniels and Kass (2001) reported over-shrinkage of the smallest roots when eigenvalues were spread far apart, and suggested shrinking the log sample eigenvalues towards their posterior mean as an alternative.

Corresponding work has considered the properties of the simultaneous distribution of the roots of two sample covariance matrices, $\mathbf{S}_1$ and $\mathbf{S}_2$, in particular for the product $\mathbf{S}_1\mathbf{S}_2^{-1}$ or $\mathbf{S}_1(\mathbf{S}_1 + \mathbf{S}_2)^{-1}$ (e.g. Chang, 1970; Krishnaiah and Chang, 1971; Venables, 1973;

Muirhead and Verathaworn, 1985; Bilodeau and Srivastava, 1992). These matrices and their eigenvalues play an important role in the multivariate analysis of variance (MANOVA) and hypothesis testing. However, asymptotic distributions and expansions given are by and large complicated and cumbersome to evaluate, or derivations are limited to special cases. As in the single sample case, a number of estimators have been suggested which involve some form of shrinkage or truncation and minimize the quadratic or entropy loss (e.g. Dey, 1988; Loh, 1991; Srivastava and Kubokawa, 1999).

Analysis of variance: The simplest scenario for a MANOVA is the balanced one-way classification, and estimation for this case has received substantial attention in the statistical literature. A particular problem is that the standard quadratic, unbiased estimator of the between groups covariance matrix is not guaranteed to be positive semi-definite (*p.s.d.*), i.e. to have eigenvalues which are non-negative. Let $\mathbf{\Sigma}_B$ and $\mathbf{\Sigma}_W$ denote the population matrices of the between and within group covariances in a one-way classification, with s groups and n observations per group. Further, let $\mathbf{W}$ and $\mathbf{B}$ be the corresponding matrices of mean squares and cross-products (MSCP) within and between groups with expected values $\mathrm{E}[\mathbf{W}] = \mathbf{\Sigma}_W$ and $\mathrm{E}[\mathbf{B}] = \mathbf{\Sigma}_W + n\mathbf{\Sigma}_B$. Estimates of $\mathbf{\Sigma}_B$ and $\mathbf{\Sigma}_W$ are then obtained by equating $\mathbf{B}$ and $\mathbf{W}$ to their expectations.

In a quantitative genetic context, groups are families with a degree of genetic relationship of ρ (e.g. $\rho = 0.25$ for half-sib and $\rho = 0.50$ for full-sib families), and estimates of the genetic ($\mathbf{\Sigma}_G$), environmental ($\mathbf{\Sigma}_E$) and phenotypic ($\mathbf{\Sigma}_P$) covariance matrices are obtained as $\hat{\mathbf{\Sigma}}_G = \rho^{-1}\hat{\mathbf{\Sigma}}_B$, $\hat{\mathbf{\Sigma}}_E = \hat{\mathbf{\Sigma}}_W - (1-\rho)\hat{\mathbf{\Sigma}}_G$ and $\hat{\mathbf{\Sigma}}_P = \hat{\mathbf{\Sigma}}_G + \hat{\mathbf{\Sigma}}_E = \hat{\mathbf{\Sigma}}_B + \hat{\mathbf{\Sigma}}_W$, respectively. An extensive simulation study by Hill and Thompson (1978) demonstrated that the probability of obtaining non-positive definite estimates of $\mathbf{\Sigma}_B$ and thus $\mathbf{\Sigma}_G$ is high, increasing with the number of traits considered and decreasing sample size, in particular number of groups. Bhargava and Disch (1982) reported similar results, presenting an analytical method to

determine this probability. In addition, Hill and Thompson (1978) showed that this probability does not depend on the individual elements of $\boldsymbol{\Sigma}_B$ and $\boldsymbol{\Sigma}_W$, but is determined entirely by the eigenvalues of the product $\boldsymbol{\Sigma}_W^{-1}\boldsymbol{\Sigma}_B$.

Hayes and Hill (1981) considered the use of the estimated covariance matrices to derive weights in a genetic selection index. This involves the product $\hat{\boldsymbol{\Sigma}}_P^{-1}\hat{\boldsymbol{\Sigma}}_G$. The authors thus proposed to shrink the roots of this product towards their mean to reduce the effect of sampling errors, and showed in a simulation study that this could improve the achieved response to selection. Rather than manipulating the roots of $\hat{\boldsymbol{\Sigma}}_P^{-1}\hat{\boldsymbol{\Sigma}}_G$ directly, Hayes and Hill (1981) modified $\mathbf{W}^{-1}\mathbf{B}$, using that for γ_i a root of $\mathbf{W}^{-1}\mathbf{B}$, $\rho^{-1}(\gamma_i - 1)/(\gamma_i - 1 + n)$ is a root of $\hat{\boldsymbol{\Sigma}}_P^{-1}\hat{\boldsymbol{\Sigma}}_G$. This was referred to as 'bending' the matrix between groups MSCP towards the matrix of within MSCP. In particular, the authors suggested to choose the shrinkage or 'bending' factor ($0 \leq \beta \leq 1$) so that the smallest, modified root of $\hat{\boldsymbol{\Sigma}}_G$ was zero or equal to a small positive value. Similarly, if λ_i is a root of $\hat{\boldsymbol{\Sigma}}_W^{-1}\hat{\boldsymbol{\Sigma}}_B$, then $1 + n\lambda_i$ is a root of $\mathbf{W}^{-1}\mathbf{B}$ and $\rho^{-1}\lambda_i/(\lambda_i + 1)$ is a root of $\hat{\boldsymbol{\Sigma}}_P^{-1}\hat{\boldsymbol{\Sigma}}_G$, i.e. Hayes and Hill's (1981) procedure is also equivalent to squeezing the roots of $\hat{\boldsymbol{\Sigma}}_W^{-1}\hat{\boldsymbol{\Sigma}}_B$ together. Note that 'bending', as proposed originally, relates to two matrices. However, it is often used to describe the corresponding modification of the eigenvalues of a single matrix, a procedure more appropriately termed 'squeezing' by Kirkpatrick and Lofsvold (1992, Appendix B).

Earlier, Klotz and Putter (1969) derived maximum likelihood (ML) and REML estimators of $\boldsymbol{\Sigma}_B$ for the balanced one-way classification, which were constrained to be *p.s.d.*. Anderson *et al.* (1986) extended this to a *p.s.d.* ML estimator for $\boldsymbol{\Sigma}_B$ of maximum rank, while Amemiya (1985) simply considered how to modify the matrix of between components when estimates were not positive definite. In essence, all these again considered $\mathbf{W}^{-1}\mathbf{B}$ and its latent roots. Rather than applying shrinkage, however, any eigenvalues less than unity were fixed at unity, yielding a *p.s.d.* estimator of $\boldsymbol{\Sigma}_B$ of rank equal to the number of

unmodified eigenvalues. This is equivalent to setting any negative eigenvalues of $\hat{\boldsymbol{\Sigma}}_W^{-1}\hat{\boldsymbol{\Sigma}}_B$ to zero, i.e results in the nearest symmetric *p.s.d.* matrix in the Frobenius norm (Higham, 1988). For ML the divisor used to obtain $\mathbf{B}$ from the sums of squares and cross-products of group means is s, rather than $s-1$ as in the MANOVA.

In the balanced case, REML estimators without restrictions on the parameter space have closed form and are identical to their counterparts from (M)ANOVA (e.g. Corbeil and Searle, 1976; Lee and Kapadia, 1984). For normally distributed data, restricting the eigenvalues of $\mathbf{W}^{-1}\mathbf{B}$ in the one-way classification to a minimum of unity in fact is equivalent to REML estimation of $\boldsymbol{\Sigma}_B$ and $\boldsymbol{\Sigma}_W$ constrained to the parameter space (Amemiya, 1985, Appendix). Note, however, that in a genetic context where we estimate $\boldsymbol{\Sigma}_E$ as $\hat{\boldsymbol{\Sigma}}_W - (\rho^{-1}-1)\hat{\boldsymbol{\Sigma}}_B$, this does not guarantee that $\hat{\boldsymbol{\Sigma}}_E$ is positive definite (except for clones, i.e. $\rho = 1$). In contrast, 'animal model' REML analyses estimate $\boldsymbol{\Sigma}_E$ directly, constraining $\hat{\boldsymbol{\Sigma}}_E$ to have positive eigenvalues, both for full rank multivariate analyses and reduced rank estimation imposing a FA structure on $\boldsymbol{\Sigma}_G$ (Meyer and Kirkpatrick, 2005). Whilst yielding the same estimates if $\hat{\boldsymbol{\Sigma}}_W - (\rho^{-1}-1)\hat{\boldsymbol{\Sigma}}_B$ is positive definite, the two approaches are thus not strictly equivalent.

Corresponding arguments apply for more complicated scenarios. For instance, extensions to nested two-way classifications have been examined by Hill and Thompson (1978), Amemiya (1985) and Das (1996). Similarly, Calvin and Dykstra (1991, 1992) considered constrained ML and REML estimation for the class of MANOVA models in which restrictions can be formulated as non-negativity of matrices of MSCP and of pairwise differences between matrices of expected MSCP. This class includes all nested and two-factor models. This suggests that the same mechanisms are operational in analyses utilizing several types of covariances between groups simultaneously, such as animal model REML analyses to estimate genetic variances for complex pedigrees.

Bias due to reduced rank estimation

Reduced rank estimators of covariance matrices yield estimates with eigenvalues that are zero, so that their rank is less than their dimension. Amemiya (1985) proposed to constrain estimated covariance matrices between groups to be *p.s.d.* by discarding any PCs with negative eigenvalues. More recently, generalized reduced rank estimators have been suggested to reduce the number of parameters to be estimated (Kirkpatrick and Meyer, 2004; Meyer and Kirkpatrick, 2005). This section examines the bias in reduced rank estimators for a balanced one-way classification. After reviewing the canonical transformation on which estimators are based, we describe the estimators themselves and show that they subject to two sources of bias.

Canonical transformation: As outlined above, estimates of the genetic covariance matrix are obtained by estimating the between group covariance matrix, $\hat{\mathbf{\Sigma}}_B = (\mathbf{B} - \mathbf{W})/n$. Examination of the relationship between the two matrices involved, $\mathbf{B}$ and $\mathbf{W}$, is made easier by applying a transformation so that

$$\hat{\mathbf{\Sigma}}_B = \frac{1}{n}(\mathbf{B} - \mathbf{W}) = \frac{1}{n}\mathbf{T}\left(\mathbf{\Lambda}_Q - \mathbf{I}\right)\mathbf{T}' \tag{2}$$

with $\mathbf{\Lambda}_Q = \text{Diag}\left\{\lambda_{Qi}\right\}$ the matrix of eigenvalues of $\mathbf{W}^{-1}\mathbf{B}$. This formulation implies that the eigenvectors of $\hat{\mathbf{\Sigma}}_B$ are proportional to the columns of $\mathbf{T}$.

This is the so-called canonical decomposition or transformation: For any two symmetric (real) matrices, $\mathbf{W}$ and $\mathbf{B}$, of size $q \times q$ with $\mathbf{W}$ positive-definite and $\mathbf{B}$ *p.s.d.* there exists a matrix $\mathbf{T}$ such that $\mathbf{TT}' = \mathbf{W}$ and $\mathbf{TDT}' = \mathbf{B}$, with $\mathbf{D}$ diagonal (Anderson, 1984). The canonical transformation has been used to simplify various multivariate problems in quantitative genetics, such as examination of selection indices (Hayes and Hill, 1980), and to reduce computational requirements in genetic evaluation (Ducrocq and Chapuis, 1997) or estimation of genetic parameters (Meyer, 1985). In addition, it has been instrumental

in the derivation of constrained REML or ML estimators of covariance matrices, reviewed above. A detailed description of the computational steps involved can be found in Seal (1964) or Amemiya (1985).

The transformation matrix $\mathbf{T}$ is given by

$$\mathbf{T} = \mathbf{W}^{1/2}\mathbf{E}_Q \tag{3}$$

with $\mathbf{W}^{1/2}$ a matrix square root of $\mathbf{W}$ and $\mathbf{E}_Q$ the matrix of eigenvectors of

$$\mathbf{Q} = \mathbf{W}^{-1/2}\,\mathbf{B}\,(\mathbf{W}^{-1/2})' = \mathbf{E}_Q\mathbf{\Lambda}_Q\mathbf{E}_Q' \tag{4}$$

Let $\mathbf{W} = \mathbf{E}_W\mathbf{\Lambda}_W\mathbf{E}_W'$ and $\mathbf{B} = \mathbf{E}_B\mathbf{\Lambda}_B\mathbf{E}_B'$ represent the eigen-decompositions of $\mathbf{W}$ and $\mathbf{B}$. Matrix square roots are not uniquely defined. Suitable forms are $\mathbf{W}^{1/2} = \mathbf{E}_W\mathbf{\Lambda}_W^{1/2}$ or, as $\mathbf{E}_W$ is ortho-normal, $\mathbf{W}^{1/2} = \mathbf{E}_W\mathbf{\Lambda}_W^{1/2}\mathbf{E}_W'$, or the Cholesky factor of $\mathbf{W}$. For $\mathbf{W}^{1/2} = \mathbf{E}_W\mathbf{\Lambda}_W^{1/2}$, $\mathbf{Q}$ becomes

$$\mathbf{Q} = \mathbf{\Lambda}_W^{-1/2}\mathbf{E}_W'\mathbf{E}_B\mathbf{\Lambda}_B\mathbf{E}_B'\mathbf{E}_W\mathbf{\Lambda}_W^{-1/2} \tag{5}$$

Reduced rank estimators: From (Eq. 2) it follows directly that $\hat{\mathbf{\Sigma}}_B$ has negative eigenvalues if $\mathbf{Q}$ has any eigenvalues $\lambda_{Q\,i}$ less than unity. Amemiya (1985) proposed an estimator that is constrained to be *p.s.d.* which utilizes this relationship.

Constrained estimator: Assume that $\mathbf{Q}$ has m_0 eigenvalues $\lambda_{Q\,i}$ greater than or equal to unity, with the remaining $q - m_0$ eigenvalues less than unity. We can then think of $n\hat{\mathbf{\Sigma}}_B$ in (Eq. 2) as the sum of a *p.s.d.* matrix $\mathbf{M}$ and a negative definite matrix $\mathbf{\Delta}$:

$$n\hat{\mathbf{\Sigma}}_B = \mathbf{M}+\mathbf{\Delta} = \mathbf{T}\left(\mathbf{\Lambda}_Q^{\star} - \mathbf{I}\right)\mathbf{T}' + \mathbf{T}\left(\mathbf{\Lambda}_Q - \mathbf{\Lambda}_Q^{\star}\right)\mathbf{T}' = \sum_{i=1}^{m_0}\left(\lambda_{Q\,i} - 1\right)\mathbf{t}_i\mathbf{t}_i' + \sum_{i=m_0+1}^{q}\left(\lambda_{Q\,i} - 1\right)\mathbf{t}_i\mathbf{t}_i' \tag{6}$$

where $\mathbf{\Lambda}_Q^{\star}$ is $\mathbf{\Lambda}_Q$ with eigenvalues $m_0 + 1$ to q replaced by unity, and $\mathbf{t}_i$ the $i-$th column of $\mathbf{T}$. A natural interpretation is then to consider $\mathbf{M}$ as an estimator of $n\mathbf{\Sigma}_B$ and $\mathbf{\Delta}$ as an

estimator of $\boldsymbol{\Sigma}_W$, i.e. an estimator of $\boldsymbol{\Sigma}_B$ constrained to be *p.s.d.* is obtained by omitting $\boldsymbol{\Delta}$ (Amemiya, 1985). This gives an estimator which has rank m, with $m \leq m_0$ the number of roots of $\mathbf{Q}$ that are greater than 1:

$$\hat{\boldsymbol{\Sigma}}_B^{\star} = \frac{1}{n}\mathbf{M} = \begin{cases} \frac{1}{n}\sum_{i=1}^{m_0}\left(\lambda_{Q\,i} - 1\right)\mathbf{t}_i\mathbf{t}_i' & \text{for } m_0 > 0 \\ \mathbf{0} & \text{for } m_0 = 0 \end{cases} \tag{7}$$

The corresponding estimator for $\boldsymbol{\Sigma}_W$ is the combination of $\mathbf{W}$ and the portion of $\mathbf{B}$ not used to estimate $\boldsymbol{\Sigma}_B$, each weighted by the appropriate degrees of freedom.

$$\hat{\boldsymbol{\Sigma}}_W^{\star} = \left(s(n-1)\mathbf{W} + (s-1)(\mathbf{B} - n\hat{\boldsymbol{\Sigma}}_B^{\star})\right)/(sn-1) \tag{8}$$

It is readily shown that $(s-1)\hat{\boldsymbol{\Sigma}}_B^{\star} + (sn-1)\hat{\boldsymbol{\Sigma}}_W^{\star} = (s-1)\mathbf{B} + s(n-1)\mathbf{W}$, i.e. that the new estimators comprise the total sums of squares and cross-products.

General reduced rank estimators: Amemiya (1985) considered m and m_0 to be determined by the sample. In some circumstances, however, we may wish to select a specific value of m for a particular analysis (Kirkpatrick and Meyer, 2004). Such general reduced rank estimator of $\boldsymbol{\Sigma}_B$ is obtained by substituting m for m_0 in (Eq. 7). Note that, for simplicity of argument, we assume here that $m \leq m_0$. In practice, we may have less than m values $\lambda_{Q\,i} \geq 1$, i.e., strictly speaking, we can only choose an estimator of $\boldsymbol{\Sigma}_B$ which is *p.s.d.* and has at most a rank of m.

Bias due to rank reduction: It is well known, though often ignored, that constraining REML estimates of (co)variance components to the parameter space gives biased results. For instance, if the population value for a variance is zero, constraining estimates to be non-negative yields estimates with expectation greater than zero. Similarly, estimates of covariance matrices forced to be *p.s.d.* are biased. From (Eq. 7) and (Eq. 8), expected values

of estimators are

$$\mathrm{E}\left[\hat{\boldsymbol{\Sigma}}_B^\star\right] = \boldsymbol{\Sigma}_B - \frac{1}{n}\mathrm{E}\left[\boldsymbol{\Delta}\right] \qquad \text{and} \qquad \mathrm{E}\left[\hat{\boldsymbol{\Sigma}}_W^\star\right] = \boldsymbol{\Sigma}_W + \frac{s-1}{sn-1}\mathrm{E}\left[\boldsymbol{\Delta}\right] \tag{9}$$

i.e. are biased proportional to $\boldsymbol{\Delta}$. Amount and sign of this bias depend on how the eigenvalues of $\mathbf{Q}$ used to obtain $\hat{\boldsymbol{\Sigma}}_B^\star$ are determined.

For the constrained estimator (Amemiya, 1985), $\boldsymbol{\Delta}$ is negative definite, as $\lambda_{Q\,i} < 1$ for all $i > m_0$. Thus, from (Eq. 9), the diagonal elements of $\hat{\boldsymbol{\Sigma}}_B^\star$ are biased upwards, and those of $\hat{\boldsymbol{\Sigma}}_W^\star$ are correspondingly biased downwards. For the general, reduced rank estimator, depending on the choice of m, we may ignore principal components with non-negligible eigenvalues. In the simplest scenario, $m_0 = q$, i.e. all $\lambda_{Q\,i}$ are greater or equal to unity. Consequently, for $m < m_0$, diagonal elements of $\hat{\boldsymbol{\Sigma}}_B^\star$ and thus $\mathrm{tr}(\hat{\boldsymbol{\Sigma}}_B^\star)$ are biased downwards, while $\mathrm{tr}(\hat{\boldsymbol{\Sigma}}_W^\star)$ is biased upwards (with $\mathrm{tr}(\cdot)$ denoting the trace operator, i.e. the sum of the diagonal elements of a matrix). As the trace of a matrix is equal to the sum of its eigenvalues, this implies that the estimated eigenvalues are biased.

The general, reduced rank estimator is inconsistent if the chosen rank m is less than the true rank of $\boldsymbol{\Sigma}_B$ (Remadi and Amemiya, 1994), i.e. does not converge to its true value as sample size increases. This is perhaps not surprising, since it is the nature the reduced rank estimator that it discards part of the quantity that it seeks to estimate. A heuristic argument for this inconsistency is as follows. From (Eq. 9), the bias in $\hat{\boldsymbol{\Sigma}}_B^\star$ is

$$\frac{1}{n}\mathrm{E}\left[\boldsymbol{\Delta}\right] = \frac{1}{n}\mathrm{E}\left[\sum_{i=m+1}^{q}\left(\lambda_{Q\,i}-1\right)\mathbf{t}_i\mathbf{t}_i'\right] = \sum_{i=m+1}^{q}\mathrm{E}\left[\lambda_i\mathbf{t}_i\mathbf{t}_i'\right] \tag{10}$$

(using that $\lambda_{Q\,i} = 1 + n\lambda_i$). As λ_i is an eigenvalue of $\hat{\boldsymbol{\Sigma}}_W^{-1}\hat{\boldsymbol{\Sigma}}_B$, we surmise that $\mathrm{E}\left[\lambda_i\right]$ is approximately equal to an eigenvalue of $\boldsymbol{\Sigma}_W^{-1}\boldsymbol{\Sigma}_B$, which is a property of the population and thus independent of the sample. Further, as $\mathbf{t}_i$ is proportional to an eigenvector of $\hat{\boldsymbol{\Sigma}}_B^\star$, we expect $\mathrm{E}\left[\lambda_i\,\mathbf{t}_i\mathbf{t}_i'\right]$ to be approximately independent of the sample size. Our conclusion

is that, for fixed m, the bias in $\hat{\boldsymbol{\Sigma}}_B^\star$ does not decline as the sample size increases. In contrast, the constrained estimator of Ammemiya (1985), does not have this problem of inconsistency, as m is not fixed but rather converges on q as $n \to \infty$ and $s \to \infty$, causing $\boldsymbol{\Delta}$ to vanish.

For genetic analyses and $\boldsymbol{\Delta}^\star = \sum_{i=m+1}^{q} \mathrm{E}\left[\lambda_i \mathbf{t}_i \mathbf{t}_i'\right]$,

$$\mathrm{E}\left[\hat{\boldsymbol{\Sigma}}_G^\star\right] = \boldsymbol{\Sigma}_G - \rho^{-1}\boldsymbol{\Delta}^\star \qquad \mathrm{E}\left[\hat{\boldsymbol{\Sigma}}_E^\star\right] = \boldsymbol{\Sigma}_E + \left(\rho^{-1} - \frac{n-1}{sn-1}\right)\boldsymbol{\Delta}^\star \qquad \text{and}$$

$$\mathrm{E}\left[\hat{\boldsymbol{\Sigma}}_P^\star\right] = \boldsymbol{\Sigma}_P - \frac{n-1}{sn-1}\boldsymbol{\Delta}^\star \quad (11)$$

Hence, the bias in both $\hat{\boldsymbol{\Sigma}}_G^\star$ and $\hat{\boldsymbol{\Sigma}}_E^\star$ is inverse proportional to the degree of relationship among family members. For $\hat{\boldsymbol{\Sigma}}_E^\star$ this is tempered by a downward bias proportional to the group size, which results in a slight downward bias in the estimate of the phenotypic covariance matrix.

Bias due to subset selection: In addition, reduced rank estimators of $\boldsymbol{\Sigma}_B$ can suffer from another source of bias, introduced when the estimation procedure retains – or 'picks up' – one or more of the PCs of $\boldsymbol{\Sigma}_B$ belonging to the subset that should have been discarded, based on the size of the corresponding eigenvalues.

To gain further insight, consider the scenario where genetic and environmental eigenvectors are collinear. Let $\boldsymbol{\Sigma}_G = \mathbf{E}\boldsymbol{\Lambda}_G\mathbf{E}'$ with $\mathbf{E}$ the matrix of eigenvectors and $\boldsymbol{\Lambda}_G = \mathrm{Diag}\{\lambda_{Gi}\}$ the diagonal matrix of genetic eigenvalues, in descending order. Further, let $\boldsymbol{\Sigma}_E = \mathbf{E}\boldsymbol{\Lambda}_E\mathbf{E}'$, with $\boldsymbol{\Lambda}_E = \mathrm{Diag}\{\lambda_{Ei}\}$ the matrix of environmental eigenvalues, in appropriate order. Assume an infinitely large sample, so that $\mathbf{W}$ and $\mathbf{B}$ are equal to their population values, i.e. $\mathbf{W} = \mathbf{E}(\boldsymbol{\Lambda}_E + (1-\rho)\boldsymbol{\Lambda}_G)\mathbf{E}'$ and $\mathbf{B} = \mathbf{E}(\boldsymbol{\Lambda}_E + (1+\rho(n-1))\boldsymbol{\Lambda}_G)\mathbf{E}'$. Consequently, $\mathbf{Q}$ (Eq. 4) is a diagonal matrix with elements $\lambda_{Qi} = (\lambda_{Ei} + (1+(n-1)\rho)\lambda_{Gi})/(\lambda_{Ei} + (1-\rho)\lambda_{Gi})$. These are the eigenvalues of $\mathbf{Q}$ and the corresponding matrix of eigenvectors is an identity matrix. However, unless the diagonal elements of $\mathbf{Q}$ are in strictly descending order, the sequence of both eigenvalues and eigenvectors is changed when arranging λ_{Qi} in descending order

of magnitude. This results in $\mathbf{E}_Q$ being equal to an identity matrix with permuted columns. Only if $\mathbf{E}_Q$ is equal to a non-permuted identity matrix are the leading PCs – eigenvalues and corresponding eigenvectors – of $\mathbf{\Sigma}_G$ estimated correctly, regardless of the order of fit, and only then is the bias in $\hat{\mathbf{\Sigma}}_G$ proportional to the smallest eigenvalues λ_{Gm+1} to λ_{Gq}. When increasing the rank of $\hat{\mathbf{\Sigma}}_G$ by one, this results in an increase in tr($\hat{\mathbf{\Sigma}}_G$) equal to the estimate of the last genetic eigenvalue fitted. In all other scenarios, the permutations will cause us to 'pick up' the wrong PC in at least some reduced rank analyses. Generalizations are difficult, but in essence, we estimate the first $k \leq m$ PCs correctly from an analysis of rank m, if the subset of m PCs considered comprises all PCs from 1 to k, i.e. if the first m columns of $\mathbf{E}_Q$ include all k elementary vectors $\boldsymbol{\epsilon}_j$ for $j = 1, k$ (with $\boldsymbol{\epsilon}_j$ a vector of length q with a single non-zero element of unity in position j).

Example: For illustration, say we have $q = 3$ and $\mathbf{Q}$ with diagonal elements $\lambda_{Q1} < \lambda_{Q2} < \lambda_{Q3}$. Rearranging eigenvalues and -vectors in descending magnitude of the λ_{Qi} then gives $\mathbf{\Lambda}_Q$ with elements λ_{Q3} in position $(1,1)$, λ_{Q2} in position $(2,2)$ and λ_{Q1} in position $(3,3)$, while the columns of $\mathbf{E}_Q$ are equal to the corresponding elementary vectors, i.e. $\mathbf{E}_Q$ has non-zero elements of unity in position (1,3), (2,2) and (3,1). Let $\mathbf{e}_i$ denote the $i-$th column of $\mathbf{E}$. With $\mathbf{T} = \mathbf{E}\mathbf{\Lambda}_W^{1/2}\mathbf{\Lambda}_Q$, for $m = 1$ this gives (from Eq. 7)

$$\mathrm{E}\left[\hat{\mathbf{\Sigma}}_B^{\star}\right] = \frac{1}{n}\begin{pmatrix}\mathbf{e}_1 & \mathbf{e}_2 & \mathbf{e}_3\end{pmatrix}\begin{pmatrix}\sqrt{\lambda_{W1}} & 0 & 0\\ 0 & \sqrt{\lambda_{W2}} & 0\\ 0 & 0 & \sqrt{\lambda_{W3}}\end{pmatrix}\begin{pmatrix}0 & 0 & 1\\ 0 & 1 & 0\\ 1 & 0 & 0\end{pmatrix}\begin{pmatrix}\lambda_{Q3}-1 & 0 & 0\\ 0 & 0 & 0\\ 0 & 0 & 0\end{pmatrix}$$
$$\begin{pmatrix}0 & 0 & 1\\ 0 & 1 & 0\\ 1 & 0 & 0\end{pmatrix}\begin{pmatrix}\sqrt{\lambda_{W1}} & 0 & 0\\ 0 & \sqrt{\lambda_{W2}} & 0\\ 0 & 0 & \sqrt{\lambda_{W3}}\end{pmatrix}\begin{pmatrix}\mathbf{e}_1'\\ \mathbf{e}_2'\\ \mathbf{e}_3'\end{pmatrix} = \rho\lambda_{G3}\mathbf{e}_3\mathbf{e}_3'$$

i.e. we would obtain an estimate of λ_{G1} equal to the true value for λ_{G3}, with the corresponding estimated eigenvector at an angle of 90° to the true first eigenvector. In other

words, our estimate of the first PC would be equal to the third PC. Similarly, for $m = 2$ estimates $\hat{\lambda}_{G1}$ and $\hat{\lambda}_{G2}$ would be equal to λ_{G2} and λ_{G3}, respectively, with both estimated eigenvectors orthogonal to the true vectors. The sum of estimated genetic eigenvalues would increase from $\text{tr}(\hat{\mathbf{\Sigma}}_G) = \lambda_{G3}$ for $m = 1$, to $\text{tr}(\hat{\mathbf{\Sigma}}_G) = \lambda_{G2} + \lambda_{G3}$ for $m = 2$. Only for $m = 3$ would we estimate λ_{G1} correctly. Hence, we would severely underestimate $\mathbf{\Sigma}_G$ in both reduced rank analyses, and would be likely to conclude that all q PCs were required to model $\mathbf{\Sigma}_G$ adequately. Paradoxically, this would be independent of the size of λ_{Q3}, i.e. holding even if the last genetic eigenvalue were close to zero and thus explained negligible variation.

In a more general scenario, genetic and environmental eigen-vectors are not the same, $\mathbf{E}_G \neq \mathbf{E}_E$, and the effect of permutations is reduced. To investigate this situation, it is convenient to describe the orientation of a set of eigenvectors in terms of the angles by which they deviate from the axes (the so-called "Givens angles", e.g. Pinheiro and Bates (1996)). It is easy to see that for $q = 2$ a single angle describes the orientation of the first eigenvector relative to the X-axis and, simultaneously, the orientation of the second eigenvector relative to the Y-axis. For an arbitrary number of dimensions, $q(q-1)/2$ angles are required to describe the orientation of all the eigenvectors. Any orthonormal matrix, such as a matrix of eigenvectors, can be written in terms of these angles:

$$\mathbf{E} = \prod_{i=1}^{q} \prod_{j=i+1}^{q} \mathbf{R}\left(\alpha_{ij}\right)$$

where α_{ij} is the angle of rotation in the plane defined by the $i-th$ and $j-th$ axes, and $\mathbf{R}$ is the corresponding rotation matrix. $\mathbf{R}(\alpha_{ij})$ has diagonal elements $r_{ii} = r_{jj} = \cos(\alpha_{ij})$ and $r_{kk} = 1$ for all $k \neq i, j$, off-diagonal elements $r_{ij} = -\sin(\alpha_{ij})$ and $r_{ji} = \sin(\alpha_{ij})$, and all other elements are zero. When all angles α_{ij} are 0, the eigenvectors coincide with the axes and all traits are uncorrelated. A useful property of the rotation matrices is that

$\mathbf{R}(\alpha_{ij})'\mathbf{R}(\beta_{ij}) = \mathbf{R}(\alpha_{ij} - \beta_{ij})'$. This means that the product of $\mathbf{E}_W'\mathbf{E}_B$ in (Eq. 5) depends only on the difference in the corresponding angles between the matrices $\mathbf{E}_W$ and $\mathbf{E}_B$.

This can be used to examine how a disparity in the orientation of the genetic and environmental eigenvectors affects reduced rank estimation for the case of $q = 2$. Let α_W and α_B denote the single angles describing the orientation of $\mathbf{E}_W$ and $\mathbf{E}_B$, respectively, and let $\delta = 2(\alpha_W - \alpha_B)$. For λ_{Wi} and λ_{Bi} the eigenvalues of $\mathbf{W}$ and $\mathbf{B}$, this gives

$$\mathbf{Q} = \frac{1}{2}\begin{pmatrix} \Big(\lambda_{B1} + \lambda_{B2} + cos(\delta)\ (\lambda_{B1} - \lambda_{B2})\Big)/\lambda_{W1} & sin(\delta)\ (\lambda_{B1} - \lambda_{B2})\ /\ \sqrt{\lambda_{W1}\lambda_{W2}} \\ sin(\delta)\ (\lambda_{B1} - \lambda_{B2})\ /\ \sqrt{\lambda_{W1}\lambda_{W2}} & \Big(\lambda_{B1} + \lambda_{B2} - cos(\delta)\ (\lambda_{B1} - \lambda_{B2})\Big)/\lambda_{W2} \end{pmatrix}$$

As discussed above, for equal angles ($\alpha_B = \alpha_W$), $\mathbf{Q}$ is a diagonal matrix with elements $\lambda_{Bi}/\lambda_{Wi}$, and any bias is determined by the relative magnitude of these ratios. Similarly, $\mathbf{Q}$ is diagonal for equal roots of $\mathbf{B}$, $\lambda_{B1} = \lambda_{B2}$. Any difference in angles then does affect the estimate of λ_{G1} from a reduced rank ($m = 1$) analysis, though the estimated direction of the corresponding eigenvector may be distorted if $\lambda_{W1} \neq \lambda_{W2}$.

For practical analyses, the effects of sampling variation may modify the bias observed or yield biased estimates when population values would not predict so. This is illustrated in Figure 1, which shows the distribution of estimates of λ_{G1} and the angle (θ_1) between the corresponding true and estimated eigenvectors for two traits with equal phenotypic variances (of $\sigma_P^2 = 100$) and heritabilities (of $h^2 = 0.4$). Due to the equality of σ_P^2 and h^2, the first eigenvector of both $\mathbf{\Sigma}_G$ and $\mathbf{\Sigma}_E$ (or, equivalently, $\mathbf{\Sigma}_B$ and $\mathbf{\Sigma}_W$) is characterized by an angle of 45°, i.e. $\alpha_B = \alpha_W$, which is independent from the level of correlations. Assuming a genetic correlation of $r_G = 0.6$, population values for the genetic roots are $\lambda_{G1} = \sigma_P^2 h^2 (1 + r_G) = 64$ and $\lambda_{G2} = \sigma_P^2 h^2 (1 - r_G) = 16$. For an environmental correlation of $r_E = 0.58$ and this gives $\lambda_{Q1} = 1.672$ and $\lambda_{Q2} = 1.645$. For such population values ($\lambda_{Q1} > \lambda_{Q2}$) we expect to estimate the first PC correctly in a reduced rank analysis fitting

this component only ($m = 1$). However, with considerable sampling variation, we 'pick up' the wrong PC in a substantial number of cases. This results in a bimodal frequency distribution for $\hat{\lambda}_{G1}$, with modes close to true population values for λ_{G1} and λ_{G2}.

MATERIAL AND METHODS

Simulation

A simulation study was conducted to examine the joint effects of sampling variation, constraints on the parameter space and bias due to reduced rank estimation on estimates of genetic PCs, contrasting sample results with values predicted from the population parameters.

Samples: The simulation assumed data with a paternal half-sib structure, comprising s sire families of size n. Simulations were carried out by sampling matrices of between and within family MSCP, **B** and **W**, from central Wishart distributions with respective degrees of freedom of $s-1$ and $s(n-1)$, as described by Odell and Feiveson (1966). Combinations considered were $s = 5000$ with $n = 40$ (referred to as S5000N40) for a very large sample, $s = 1000$ with $n = 20$ or $n = 6$ (S1000N20 or S1000N6) for (moderately) large samples, $s = 200$ with $n = 20$ or $n = 6$ (S200N20 or S200N6) for moderately small samples, and $s = 100$ with $n = 20$ or $n = 6$ (S100N20 or S100N6) for a small sample size. A total of 10 000 replicates were carried out for each combination of population values and sample size.

Population values: The number of traits considered was $q = 6$ throughout. Population values for genetic and environmental covariance matrices were parameterised in terms of their eigenvalues and Givens angles (α_G and α_E) to determine the corresponding eigenvectors. Six different scenarios were considered, chosen to represent different spreads of

eigenvalues and thus rates of decline in genetic roots, and ratios of genetic to environmental roots, with the absolute values not important. Genetic eigenvalues for case A were 53, 52, 51, 49, 48 and 47, i.e. represented a scenario with very similar eigenvalues. For case B, population roots were moderately spread, with values for λ_{Gi} of 100, 75, 50, 35, 25, and 15. Case C, E and F assumed values of 175, 50, 35, 20, 15 and 5, i.e. a fairly wide spread, and case D, with values of 220, 45, 15, 10, 7 and 3, mimicked an even more extreme spread.

This yielded $\mathbf{\Sigma}_G$ of true rank $m^\star = q$, and the same $tr(\mathbf{\Sigma}_G) = 300$ and average eigenvalue of 50 for all scenarios. For cases A, B, C and D, eigenvalues of $\mathbf{\Sigma}_E$ were simulated as twice the value of their genetic counterparts, which yielded equal λ_{Qi} $(i = 1, q)$. For cases E and F, $\lambda_{Ei} = 2\lambda_{Gi} - 5$ and $\lambda_{Ei} = 5\lambda_{Gi} - 5$ were used. This was aimed at creating constellations where population values for λ_{Qi} were different and in an order which was likely to cause reduced rank estimates to pick up the wrong subset of principal components, while still allowing sampling variation to have an effect. Again, the choice of actual values was somewhat arbitrary and was made after considering a range of possibilities. In addition, a scenario where $\mathbf{\Sigma}_G$ had a true rank of $m^\star = q/2 = 3$ was simulated for all six constellations, by setting $\lambda_{Gi} = 0$ for $i = 4, 6$.

Motivated by the analytic results, we also examined the effects of the orientation of the eigenvectors of $\mathbf{\Sigma}_G$ and $\mathbf{\Sigma}_E$. As noted earlier, with q variables $q(q-1)/2$ angles are needed to describe the orientation of a set of eigenvectors. For simplicity, we assumed all these angles were equal to α_G for the genetic eigenvectors. For cases A, B, C, and D we set $\alpha_G = 0°$, which is equivalent to all genetic correlations being zero. For cases E and F, we took $\alpha_G = 45°$, meaning that the genetic principal components point in directions that lie between the axes. All angles for the environmental eigenvectors were also assumed equal. These were expressed as differences relative to the genetic eigenvectors; thus $\alpha_E = 0°$ means that the environmental and genetic eigenvectors coincide. Values ranging

from $\alpha_E = -45$ to $\alpha_E = 90°$ were used to parameterise $\boldsymbol{\Sigma}_E$.

Analyses

Estimation: For each replicate, REML estimates of $\boldsymbol{\Sigma}_G$ and $\boldsymbol{\Sigma}_E$ were obtained for $\hat{\boldsymbol{\Sigma}}_G$ of rank $m = 1, q$ and $\hat{\boldsymbol{\Sigma}}_E$ of rank q, using Ameniya's (1985) procedure to obtain estimates of $\boldsymbol{\Sigma}_B$ and $\boldsymbol{\Sigma}_W$ (see (Eq. 7) and (Eq. 8) above) and derive $\hat{\boldsymbol{\Sigma}}_G = 4\hat{\boldsymbol{\Sigma}}_B^{\star}$ and $\hat{\boldsymbol{\Sigma}}_E = \hat{\boldsymbol{\Sigma}}_W^{\star} - 3\hat{\boldsymbol{\Sigma}}_B^{\star}$. If this yielded an estimate of $\boldsymbol{\Sigma}_E$ which was not positive definite, a derivative-free search strategy was employed to maximize the REML log likelihood

$$\log \mathcal{L} = const. - \frac{1}{2}\Bigg((s-1)\left(\log|\hat{\boldsymbol{\Sigma}}_E + (n-1)\rho\hat{\boldsymbol{\Sigma}}_G| + \operatorname{tr}\left((\hat{\boldsymbol{\Sigma}}_E + (n-1)\rho\hat{\boldsymbol{\Sigma}}_G)^{-1}\mathbf{B}\right)\right) + s(n-1)\left(\log|\hat{\boldsymbol{\Sigma}}_E + (1-\rho)\hat{\boldsymbol{\Sigma}}_G| + \operatorname{tr}\left((\hat{\boldsymbol{\Sigma}}_E + (1-\rho)\hat{\boldsymbol{\Sigma}}_G)^{-1}\mathbf{W}\right)\right)\Bigg) \quad (12)$$

constraining $\hat{\boldsymbol{\Sigma}}_E$ to have full rank and $\hat{\boldsymbol{\Sigma}}_G$ to have rank no larger than m. While computationally more demanding than a maximization procedure using derivatives of $\log \mathcal{L}$, this was chosen for its ease of implementation.

Summary statistics: Bias in estimates of $\boldsymbol{\Sigma}_G$ was quantified by considering the estimated PCs, i.e. both eigenvalues and eigenvectors. Rather than examining the $m(2q - m - 1)/2$ individual elements or Givens angles of the eigenvectors of $\hat{\boldsymbol{\Sigma}}_G$, differences between estimates and population values were summarized by considering the m angles between estimated and corresponding population vectors. For $\mathbf{e}_i$ the $i-$th eigenvector and $\hat{\mathbf{e}}_i$ its estimate, the angle (in °) is

$$\theta_i = (180/\pi)\arccos\left|\mathbf{e}_i'\hat{\mathbf{e}}_i\right|$$

Taking the absolute value of the inner product $\mathbf{e}_i'\hat{\mathbf{e}}_i$ projects all angles to the first quadrant, i.e. θ_i have a value between 0 and 90°.

The effect of sampling and bias on the estimates of covariance matrices was summarized by considering the Frobenius norm of the matrix difference, $\|\hat{\mathbf{\Sigma}}_X - \mathbf{\Sigma}_X\|_F$ ($X = G, E$) (For a matrix $\mathbf{M}$ with elements m_{ij} the Frobenius norm is $\|\mathbf{M}\|_F = \sqrt{\sum_i \sum_j m_{ij}^2}$ (Golub and Van Loan, 1996)). For all statistics, means and sampling deviation across replicates were calculated. In addition, mean square errors (MSE) were obtained as the average (over replicates) of squared deviations of individual parameters estimates from the corresponding population values.

Predicted values for estimates of $\lambda_{G\,i}$ and corresponding θ_i were obtained from the eigendecomposition of $\hat{\mathbf{\Sigma}}_G^\star$ (Eq. 11), evaluated for the population values and given rank m.

Rank tests: For each sample, the rank of $\hat{\mathbf{\Sigma}}_G$, i.e. the number of eigenvalues significantly different from zero, was determined using several procedures. Where applicable, an error probability of 5% was used. Based on (Eq. 12), likelihood ratio tests (LRT) were carried out, comparing minus twice the difference in $\log \mathcal{L}$ for $m = 2, q$ and $\log \mathcal{L}$ for $m - 1$ against a χ^2 criterion with $q - m + 1$ degrees of freedom. In addition, the LRT described by Anderson *et al.* (1986) to test the hypothesis that the estimate of $\mathbf{\Sigma}_B$ from the balanced one-way classification had rank of at most m versus the alternative that it had rank great than m, was applied. Quantiles of the distribution of the test statistic, which does not follow a χ^2 distribution, have been tabulated by Amemiya *et al.* (1990) and Kuriki (1993). Let $q^\star$ denote the number of eigenvalues of $\mathbf{W}^{-1}\mathbf{B}$, $\lambda_{Q\,i}$, greater than unity. The test criterion is then (from Amemiya *et al.*, 1990)

$$\Upsilon = (sn - 1) \sum_{i=m+1}^{q^\star} \log\Big((s-1)\lambda_{Q\,i} + s(n-1)\Big) - (s-1) \sum_{i=m+1}^{q^\star} \log(\lambda_{Q\,i}) \\ - (q^\star - m)\log(sn - 1)$$

If Υ exceeds the tabulated value for a probability of 0.95 and rank difference $q - m$, the alternative hypothesis is accepted, with an error probability of 5%; see Hine and Blows

(2006) for a detailed description.

Further, the Akaike information criterion (AIC), adjusted for small sample size, and Bayesian information criterion (BIC) for each analysis were obtained as (Burnham and Anderson, 2004; Wolfinger, 1996)

$$\text{AIC} = -2\log\mathcal{L} + 2p\left(1 + \frac{p+1}{qsn - p - 1}\right) \quad \text{and} \quad \text{BIC} = -2\log\mathcal{L} + p\log\Big(q\,(sn-1)\Big)$$

with $p = (m(2q - m + 1) + q(q + 1))/2$ the number of parameters for an analysis estimating $\boldsymbol{\Sigma}_G$ of rank m. For $q = 6$ and $m = 1, 6$, values were $p = 27, 32, 36, 39, 41$ and 42, respectively. The rank of $\hat{\boldsymbol{\Sigma}}_G$ was then chosen as the value of m with the smallest value of the information criterion.

RESULTS

This section describes results of the simulation study. To begin with, we examine the bias in reduced rank estimates, showing that, in practice, reduced rank estimates of $\boldsymbol{\Sigma}_G$ are subject to bias from up to three sources, namely due to the spread of sample roots, due to constraints imposed to ensure that estimates are within the parameter space, and due to 'picking up' the wrong PC. We then demonstrate that the latter can dominate MSE, so that omitting PCs with negligible eigenvalues is not always as advantageous as we might hope. Finally, we examine the scope for likelihood based tests to determine the rank of $\hat{\boldsymbol{\Sigma}}_G$ correctly.

Bias

Figure 2 summarizes mean estimates of genetic eigenvalues from full rank analyses, for cases A and B with $\alpha_E = \alpha_G = 0$, and a number of sample sizes. Such angles yield

population covariance matrices that are diagonal. However, as emphasized by Hill and Thompson (1978), this can be thought of as representing a variety of non-diagonal constellations which have the same eigenvalues. For the case of equal roots, the authors presented a number of combinations of genetic and phenotypic correlations and heritabilities which resulted in the same eigenvalues as diagonal population values (Table 3 in Hill and Thompson, 1978). Results clearly show the upwards bias of the largest, and corresponding downwards bias of the smallest sample roots. This effect is the more pronounced the more similar the population roots are, and, for case A, increases dramatically with decreasing sample size. Estimates of the corresponding eigenvalues of $\hat{\mathbf{\Sigma}}_E$ exhibited an analogous pattern (not shown).

Mean estimates of the first genetic eigenvalue and the direction of the corresponding eigenvector are contrasted in Figure 3 to their predicted values, for analyses restricting $\hat{\mathbf{\Sigma}}_G$ to rank $m = 1$ and allowing it to have full rank ($m = 6$) for cases A, B, C and D. Graphs do not include expected values for $\alpha_E = 0$ as for these cases all population roots λ_{Qi} are the same, which makes the order in which PCs are 'picked up' arbitrary. Inspecting the (log) profile likelihood (at population values) for individual λ_{Gi}, this is exemplified by a horizontal portion of the curve, i.e. there is no clear maximum. For case A, the predicted, downwards bias in $\hat{\lambda}_{G1}$ due to reduced rank estimation ($m = 1$) is small, for all values of α_E. Mean estimates, however, are consistently higher than predicted and higher than the population values, even for a relatively large sample involving 1000 sires with 20 progeny each. There is little difference between reduced and full rank estimates, i.e. this reflects the bias due to the spread in sampling roots which, as demonstrated above, is largest when the true eigenvalues are close together. Even at full rank, estimates of the direction of first genetic eigenvector differ markedly from the population direction, indicating that the wrong PC is 'picked up' in a substantial number of replicates. Due to the closeness

of the population roots, however, this has little effect on the corresponding estimates of eigenvalues. For cases B and C, there is good agreement between predicted and observed bias for the larger sample size. For the small sample in case B, there is again a marked upward bias in $\hat{\lambda}_{G1}$ due to the spread in sample roots.

For case C, however, estimates for the small sample are less than expected, increasingly so as the population values for α_E increase. This can be attributed to constraining the estimate of $\boldsymbol{\Sigma}_E$ to be positive definite. This causes genetic variation to be partitioned into the environmental components, counter-acting the upwards bias in $\hat{\lambda}_{G1}$ due to dispersion in sample roots. For this scenario (Case C, S200N6), the proportion of samples in which such constraints were required increased from 3% for $\alpha_E = 10°$ to 48% for $\alpha_E = 90°$. A similar pattern is evident for case D, with even higher proportions of samples needing $\hat{\boldsymbol{\Sigma}}_E$ to be constrained, so that a downward bias for $\hat{\lambda}_{G1}$ was notable for $m = 1$ at higher values of α_E, even for the large sample size.

Corresponding results for case E and various orders of fit are shown in Figure 4. For this constellation of population values, we expect a complete reversal in the order of λ_{Qi} at equal genetic and environmental angles ($\alpha_E = 0$). Hence, for orders of fit of $m = 1, 2, 4$ and 5, predicted values of $\hat{\lambda}_{G1}$ are $\lambda_{G6} = 5$, $\lambda_{G5} = 15$, $\lambda_{G3} = 35$ and $\lambda_{G2} = 50$, respectively. Correspondingly, we expect to find an angle between true and estimated eigenvector of 90° for all reduced rank analyses. As graphs show, predicted bias decreases rapidly with increasing difference of angle α_E from α_G. Due to sampling variation, the probability of having equal angles in a sample is exceedingly small, and mean estimates for the first PC at $\alpha_E = 0$ are thus much less biased than expected, in particular for the direction of the first eigenvector. Moreover, the deviation from predicted values for similar genetic and environmental eigenvectors appears to increase with the number of genetic PCs fitted. As above, simulation results are again modulated by the effects of constraining $\hat{\boldsymbol{\Sigma}}_G$ to the

parameter space, in particular for S200N6 and large differences between α_E and α_G.

Mean square error

Figure 5 illustrates the combined effects of bias and sampling variation on estimates of λ_{G1} and Σ_G, for Σ_G with true rank of $m^\star = 3$. Shown are the MSE for $\hat{\lambda}_{G1}$ and the mean error, i.e. the average of $\|\hat{\Sigma}_G - \Sigma_G\|_F$ over replicates, for Σ_G. As demonstrated above, bias in $\hat{\lambda}_{G1}$ is largest for $\alpha_G = \alpha_E$ and large samples, when Σ_G is estimated at less than true rank ($m < m^\star$). Observed MSEs for case E follow a similar pattern, indicating that the bias dominates over any reduction in sampling variances due to a reduction in the number of parameters estimated. Overall there appears to be relatively little increase in MSE when attempting to fit more PCs for Σ_G than the true rank.

Results may, to some extent at least, reflect that population values considered implied moderate to moderately highheritabilities. Additional simulations (not presented here) demonstrated some advantages of estimating Σ_G with rank $m < m^\star$, or bigger penalties for fitting too many PCs at low levels of heritabilities. For case F and a small sample size we do observe a reduction in MSE of $\hat{\lambda}_{G1}$ for analyses fitting less than three genetic PCs. Case F differs from case E only by much higher environmental covariances. A similar pattern, in particular a notable reduction in MSE for reduced rank analyses when Σ_W was considerably larger than Σ_B, has been reported by Remadi and Amemiya (1994).

Rank

An important question in conjunction with reduced rank estimation is whether we can reliably identify the number of PCs that need to be fitted. Figure 6 summarizes the proportion of replicates with different numbers of sample roots λ_{Qi} greater than unity

together with the rank determined on the basis of LRTs and information criteria. As expected, due to the sample spread in the roots of $\mathbf{Q}$ (see Hill and Thompson, 1978), sampling $\mathbf{\Sigma}_G$ at full rank yielded a substantial number of replicates with less than q eigenvalues of $\mathbf{Q}$ greater than unity. For case C, this occurred even for a large sample size, and increasingly as genetic and environmental eigenvectors deviated in direction. Conversely, for population values of $\mathbf{\Sigma}_G$ with rank $m^\star = 3$, more than 3 roots of $\mathbf{Q}$ exceeded unity in most samples. Similarly, Remadi and Amemiya (1994) found a much greater frequency of samples with more eigenvalues $\lambda_{Qi} > 1$ than the true rank of $\mathbf{\Sigma}_B$, when $\mathbf{\Sigma}_B$ did not have full rank.

Rank tests performed reasonably well at the larger sample size, except for case C and larger values of α_E. The standard LRT used ignores the fact that hypotheses tested involved parameter values at the boundary of the parameter space, and are thus expected to be too conservative (Self and Liang, 1987), while the test of Anderson *et al.* (1986) should account for the resulting, non-standard conditions. However, for both LRTs, the proportion of samples classified at a rank other than that simulated frequently exceeded the nominal error rate of 5%. For a small sample, rank tests underestimated the true rank of $\mathbf{\Sigma}_G$ in a substantial number of cases. In particular, BIC proved very stringent and seldom provided a correct estimate, while AIC and LRTs yielded comparable results. Corresponding results were obtained for the other cases (not shown). For a simulation with similarly small samples, Hine and Blows (2006) also reported a substantial frequency of underestimates of the rank of $\mathbf{\Sigma}_G$, based on the Y criterion. Tests applied rely on asymptotic large sample properties. Not surprisingly this appears not to hold all that well for the smaller samples.

DISCUSSION

Principal components are an important tool for multivariate statistical analyses. Their utility, however, comes at a price, as estimates of both eigenvalues and eigenvectors are subject to bias. It is well known, that sampling variation causes the leading eigenvalues to be overestimated and the smallest eigenvalues to be underestimated. One consequence is that estimated genetic covariance matrices can lie outside the parameter space (Hill and Thompson, 1978). Reduced rank estimators have been proposed initially to constrain estimates to be *p.s.d.* (Amemiya, 1985), but can be generalised to estimators of given maximum rank m. Both are biased, with magnitude and sign of the bias depending on the choice of rank, and the PCs of $\hat{\mathbf{\Sigma}}_G$ which have been discarded.

Importance of bias: Our results have shown that the bias in reduced rank estimates of genetic PCs can be large. It has to be emphasized, however, that by considering the first genetic eigenvalue and reduced rank analyses estimating $\mathbf{\Sigma}_G$ at well below its true rank, we have concentrated – for the purpose of illustration – on the worst possible scenario. In practice, we are often interested in the first two to three genetic PCs; as shown bias in the leading genetic PCs declines rapidly as the number of PCs fitted increases. Examining a number of multivariate estimates from the literature, Kirkpatrick (2008) postulated that the 'effective number of dimensions' of $\mathbf{\Sigma}_G$ is generally less than two, and showed that the ratio of λ_{G1} to λ_{G2} is often in the vicinity of 5:1. This is synonymous with a large proportion of the genetic variance explained by the leading PCs and a fairly wide spread of eigenvalues. Hence, we might expect a substantial proportion of applications to fall somewhere in the range spanned by cases C and D considered in the simulation study. Simulation results further showed that the effects of sampling variation on the spread of eigenvalues can be pronounced, especially for small samples, and that this can counteract

the bias due to estimating Σ_G of rank $m < m^\star$. The bias due to reduced rank estimation we are likely to encounter in practice could thus be substantially less than demonstrated here.

As shown, biases are affected not only by the spread of sample roots, but also by the relative orientation of the genetic and environmental eigenvectors. This issue has not been has not been examined for practical data sets. Cheverud (1988) reported that genetic and phenotypic correlations are often similar, which implies similarity of genetic and environmental correlations. Common eigenvectors for two matrices generate the same correlation structure if the corresponding eigenvalues are proportional. For cases with similar genetic and environmental correlations, this suggests that genetic and environmental eigenvectors may have directions which are also similar, i.e. that differences in α_G and α_E are small. If so, this will tend to exacerbate the problems of bias due to reduced rank estimation.

This study has been motivated by results from large scale, reduced rank multivariate 'animal model' REML analyses of data from beef cattle, where estimated genetic eigenvalues changed dramatically with the number of genetic PCs fitted (Meyer, 2005, 2007b). In particular, it was observed that the eigenvalue for the last PC fitted tended to be underestimated. Our findings clearly demonstrate that this was due to the inherent bias in estimates when imposing rank constraints. Moreover, they explain the paradox that some PCs with apparently negligible eigenvalues needed to be fitted to obtain reduced rank estimates of the genetic covariance matrix which adequately described the dispersion structure in the data. With most relationships in the data due to paternal half-sibs, substituting estimates from a full rank analysis for population values, Meyer and Kirkpatrick (2007) showed that (Eq. 11) provided a good prediction of estimates of genetic PCs obtained from the practical, reduced rank analyses.

Implications for reduced rank estimation: Biases have been examined for the case of a balanced one-way classification, under the surmise that similar mechanisms affect REML estimates of the genetic covariance matrix with reduced rank in more general cases. The surprising result in has been that the bias in reduced rank estimates of $\mathbf{\Sigma}_G$ can be markedly higher than expected from simply ignoring PCs $m+1$ to $m^\star$. As outlined above, the underlying mechanism for this additional bias is that the REML estimator 'picks up' the wrong subset of PCs, with some rotations of estimates determined by the differences in orientation of genetic and environmental eigenvectors and the spread in genetic eigenvalues. This 'extra' bias then tends to dominate the MSE, so that even for small samples, MSEs are often not reduced for analyses fitting less than $m^\star$ PCs. Comparing convergence rates for reduced rank analyses, Meyer (2008) found that severe underfitting of the rank of $\mathbf{\Sigma}_G$ tended to increase total computational requirements of REML analyses, in spite of markedly reduced requirements for individual iterates. This suggests that biases may also affect the topography of the likelihood surface, making standard numerical maximization techniques less effective.

The idea of fitting only as many genetic PCs or, equivalently, parameters to be estimated, as can be supported by the data (Kirkpatrick and Meyer, 2004; Blows, 2007) is appealing, in particular for small data sets. However, without further qualifications it is only applicable in specific cases, and thus needs to be utilized cautiously. A 'safer' recommendation is to ensure that we do not underfit $\mathbf{\Sigma}_G$, i.e. to fit sufficient genetic PCs to capture all important genetic PCs. If this does comprise PCs with negligible eigenvalues, we may omit these in subsequent applications, e.g. when obtaining breeding values based on the estimated genetic covariance matrix. In other words, the optimal number of genetic PCs considered for genetic evaluation may differ from that for variance component estimation.

Selecting the rank of $\mathbf{\Sigma}_G$*:* As computational requirements of REML analyses decrease with

the number of genetic PCs fitted, Kirkpatrick and Meyer (2004) proposed a scheme which increased the rank of $\hat{\mathbf{\Sigma}}_G$ successively, until no further non-neglible eigenvalues were found. However, results from this study suggest that this may be misleading. For instance, we may fit m genetic PCs and find that the estimate of the $m-$th eigenvalue is close to zero. As we may, in fact, have 'picked up' one of the subsequent PCs, this estimate may increase substantially when increasing the number of PCs considered to $m+1$. Hence, a 'step-down' procedure to determine the rank of $\mathbf{\Sigma}_G$ to be fitted may be preferable for practical applications. If the estimate of $\lambda_{G\,m}$ from an analysis fitting m genetic PCs is markedly less than the corresponding estimate from an analysis fitting $m+1$ PCs, we may suspect that we have 'picked up' one of the subsequent PCs, i.e. one of PC $m+1$ to q instead of PC m, and that reducing the number of PCs further would be futile.

An additional criterion to monitor is the sum of estimated genetic eigenvalues, i.e. $\text{tr}(\hat{\mathbf{\Sigma}}_G)$. This may drop slightly, in particular for small samples, as we successively ignore PCs with negligible eigenvalues, due to the small eigenvalues omitted and a somewhat reduced spread in the sample roots. Ignoring PC $m+1$ with eigenvalue $\lambda_{G\,m+1}$ should yield a reduction in $\text{tr}(\hat{\mathbf{\Sigma}}_G)$, from an analysis fitting $m+1$ genetic PCs to an analysis fitting m PCs, of that amount (i.e. $\lambda_{G\,m+1}$). A larger difference then again indicates that we have encountered an analysis in which we do not estimate the correct subset of PCs, i.e. that we should fit no less than $m+1$ genetic PCs, even if $\hat{\lambda}_{G\,m+1}$ is close to zero. This should be accompanied by a significant decrease in the likelihood. For higher dimensional multivariate analyses it is good practice to carry out a series of preliminary, bivariate analyses, pooling results to obtain starting values of $\mathbf{\Sigma}_G$ and $\mathbf{\Sigma}_E$ for REML analyses considering all traits. Even if we have a more complicated pedigree structure than that in a one-way classification, substituting these for population values in (Eq. 11), may then give an indication at what order of fit for $\mathbf{\Sigma}_G$ problems might occur. In addition, such calculations may provide more

appropriate starting values for analyses with reduced rank, or suggest a permutation of trait numbers which may reduce the 'extra' bias of reduced rank estimates.

Identification of the effective dimension of the genetic covariance matrix, i.e. the number of PCs with eigenvalues greater than zero, is of considerable interest to quantitative geneticists. This includes animal breeders who would like to know which is the simplest, reduced rank model which is appropriate, and evolutionary biologist for whom matrices of less than full rank indicate constraints by nature on response to selection (Kirkpatrick and Lofsvold, 1992; Mezey and Houle, 2005; Hine and Blows, 2006; Blows and Walsh, 2008; Kirkpatrick, 2008). A number of tests for matrix rank are commonly used. Disconcertingly, simulation studies available generally show somewhat inconsistent results, both between different tests and in the ability to find the correct dimension (Jackson, 1993; Ferré, 1995; Peres-Neto *et al.*, 2005; Dray, 2007). Similarly, identification of the correct rank in our study, based on the log likelihood and information criteria, has only been moderately successful, with substantial underestimates of the true rank for smaller samples. On the one hand, LRTs are known to favour the most detailed model. On the other hand, there has been some concern that use of AIC and BIC in a random effects model violates some of the underlying assumptions (Ripley, 2004). However, LRTs and AIC yielded, by and large comparable, results, while BIC appeared to be far too stringent. Reliable identification of the dimension of $\mathbf{\Sigma}_G$ hence remains on open challenge.

CONCLUSIONS

Reduced rank estimation of genetic covariance matrices is appealing and readily accommodated in standard, mixed model based estimation procedures such as REML. However, reduced rank estimation can yield estimates biased by 'picking up' the wrong subset of

genetic principal components. It is thus important to choose the rank judiciously, i.e. sufficiently large to avoid such problems even if this does comprise a number of components with small eigenvalues.

Acknowledgments

This work was supported by Meat and Livestock Australia under grant BFGEN.100B (KM) and National Science Foundation grant EF-0328598 (MK).

References

Amemiya, Y., 1985 What should be done when an estimated between-group covariance matrix is not nonnegative definite ? Amer. Stat. **39**: 112–117.

Amemiya, Y., T. W. Anderson, and P. A. W. Lewis, 1990 Percentage points for a test of rank in multivariate components of variance. Biometrika **77**: 637–641.

Anderson, B. M., T. W. Anderson, and I. Olkin, 1986 Maximum likelihood estimators and likelihood ratio criteria in multivariate components of variance. Ann. Stat. **14**: 405–417.

Anderson, T. W., 1984 *An Introduction to Multivariate Statistical Analysis*. Wiley, New York, 2nd edition.

Bhargava, A. K., and D. Disch, 1982 Exact probabilities of obtaining estimated non-positive definite between-group covariance matrices. J. Stat. Comp. Simul. **15**: 27–32.

Bilodeau, M., and M. S. Srivastava, 1992 Estimation of the eigenvalues of $\mathbf{\Sigma}_1\mathbf{\Sigma}_2^{-1}$. J. Multiv. Anal. **41**: 1–13.

Blows, M. W., 2007 A tale of two matrices: multivariate approaches in evolutionary biology. Journal of Evolutionary Biology **20**: 1–8.

Blows, M. W., and J. B. Walsh, 2008 Spherical cows grazing in flatland: Constraints to selection and adaptation. In J. H. J. van der Werf, H.-U. Graser and C. Gondro, editors, *Adaptation and fitness in animal populations – Evolutionary and breeding perspectives on genetic resource management*. Springer Verlag, 000–000 (in press).

Bryant, E. H., and W. R. Atchley, 1975 *Multivariate Statistical Methods : Within-Groups Covariation*, volume 2 of *Benchmark Papers in Systematic and Evolutionary Biology*. Dowden, Hutchinson & Ross, Inc.; Halsted Press, Stroudsburg, Pa.

Burnham, K. P., and D. R. Anderson, 2004 Multimodel inference : Understanding AIC and BIC in model selection. Sociol. Meth. Res. **33**: 261–304.

Calvin, J. A., and R. L. Dykstra, 1991 Maximum likelihood estimation of a set of covariance matrices under Loewner order restrictions with applications to balanced multivariate variance components models. Ann. Stat. **19**: 850–869.

Calvin, J. A., and R. L. Dykstra, 1992 An algorithm for restricted maximum likelihood estimation in balanced multivariate variance components models. J. Stat. Comp. Simul. **40**: 233–246.

Chang, T. C., 1970 On an asymptotic representation of the distribution of the characteristic roots of $\mathbf{S}_1\mathbf{S}_2^{-1}$. Ann. Math. Stat. **41**: 440–445.

Cheverud, J. M., 1988 A comparison of genetic and phenotypic correlations. Evolution **42**: 958–968.

Corbeil, R. R., and S. Searle, 1976 A comparison of variance component estimators. Biometrics **32**: 779–791.

Daniels, M., and R. E. Kass, 2001 Shrinkage estimators for covariance matrices. Biometrics **57**: 1173–1184.

Das, K., 1996 Improved estimation of covariance matrices in balanced hierarchical multivariate variance components models. Statistics **28**: 73–83.

Dey, D. K., 1988 Simultaneous estimation of eigenvalues. Ann. Inst. Stat. Math. **40**: 137–147.

Dray, S., 2007 On the number of principal components: A test of dimensionality based on measurements of similarity between matrices. Comp. Stat. Dat. Anal. **52**: 2228–2237.

Ducrocq, V., and H. Chapuis, 1997 Generalising the use of the canonical transformation for the solution of multivariate mixed model equations. Genet. Select. Evol. **29**: 205–224.

Ferré, L., 1995 Selection of components in principal component analysis: A comparison of methods. Comp. Stat. Dat. Anal. **19**: 669–682.

Golub, G. H., and C. F. Van Loan, 1996 *Matrix Computations*. John Hopkins University Press, Baltimore, 3rd edition.

Hayes, J. F., and W. G. Hill, 1980 A reparameterisation of a genetic index to locate its sampling properties. Biometrics **36**: 237–248.

Hayes, J. F., and W. G. Hill, 1981 Modifications of estimates of parameters in the construction of genetic selection indices ('bending'). Biometrics **37**: 483–493.

Higham, N. J., 1988 Computing a nearest symmetric positive semi-definite matrix. Lin. Alg. Appl. **103**: 103–118.

Hill, W. G., and R. Thompson, 1978 Probabilities of non-positive definite between-group or genetic covariance matrices. Biometrics **34**: 429–439.

Hine, E., and M. W. Blows, 2006 Determining the effective dimensionality of the genetic variance-covariance matrix. Genetics **173**: 1135–1144.

Hotelling, H., 1933 Analysis of a complex of statistical variables into principal components. J. Educ. Psych. **24**: 417–441.

Jackson, D. A., 1993 Stopping rules in principal component analysis : a comparison of heuristic and statistical approaches. Ecology **74**: 2204–2214.

Kirkpatrick, M., 2008 Patterns of quantitative genetic variation in multiple dimensions. Genetica **00**: 000–000 (in press).

Kirkpatrick, M., and D. Lofsvold, 1992 Measuring selection and constraint in the evolution of growth. Evolution **46**: 954–971.

Kirkpatrick, M., and K. Meyer, 2004 Direct estimation of genetic principal components: Simplified analysis of complex phenotypes. Genetics **168**: 2295–2306.

Klotz, J., and J. Putter, 1969 Maximum likelihood estimation of multivariate covariance components for the balanced one-way layout. Ann. Math. Stat. **40**: 1100–1105.

Krishnaiah, P. R., and T. C. Chang, 1971 On the exact distributions of the extreme roots of the Wishart and MANOVA matrices. J. Multiv. Anal. **1**: 108–117.

Kuriki, S., 1993 One-sided test for the equality of two covariance matrices. Ann. Stat. **21**: 1379–1384.

Lawley, D. N., 1956 Tests of significance for the latent roots of covariance and correlation matrices. Biometrika **43**: 128–136.

Ledoit, O., and M. Wolf, 2004 A well-conditioned estimator for large-dimensional covariance matrices. J. Multiv. Anal. **88**: 365–411.

Lee, K. R., and C. H. Kapadia, 1984 Variance component estimators for the balanced two-way mixed model. Biometrics **40**: 507–512.

Loh, W. L., 1991 Estimating covariance matrices. Ann. Stat. **19**: 283–296.

Los Campos, G., and D. Gianola, 2007 Factor analysis models for structuring covariance matrices of additive genetic effects: a Bayesian implementation. Genet. Select. Evol. **39**: 481–94.

Meyer, K., 1985 Maximum likelihood estimation of variance components for a multivariate mixed model with equal design matrices. Biometrics **41**: 153–166.

Meyer, K., 2005 Genetic principal components for live ultra-sound scan traits of Angus cattle. Anim. Sci. **81**: 337–345.

Meyer, K., 2007a Covariance structures for quantitative genetic analyses. Proc. Ass. Advan. Anim. Breed. Genet. **17**: 142–149.

Meyer, K., 2007b Multivariate analyses of carcass traits for Angus cattle fitting reduced rank and factor-analytic models. J. Anim. Breed. Genet. **124**: 50–64.

Meyer, K., 2008 Parameter expansion for estimation of reduced rank covariance matrices. Genet. Select. Evol. **40**: 3–24.

Meyer, K., and M. Kirkpatrick, 2005 Restricted maximum likelihood estimation of genetic principal components and smoothed covariance matrices. Genet. Select. Evol. **37**: 1–30.

Meyer, K., and M. Kirkpatrick, 2007 A note on bias in reduced rank estimates of covariance matrices. Proc. Ass. Advan. Anim. Breed. Genet. **17**: 154–157.

Mezey, J. G., and D. Houle, 2005 The dimensionality of genetic variation for wing shape in Drosophila melanogaster. Evolution **59**: 1027–1038.

Muirhead, R. J., 1987 Developments in eigenvalue estimation. In A. K. Gupta, editor, *Advances in Multivariate Statistical Analysis*. Kluwer, Dordrecht, Holland, 277–288.

Muirhead, R. J., and T. Verathaworn, 1985 On estimating the latent roots of $\Sigma_1\Sigma_2^{-1}$. In P. R. Krishnaiah, editor, *Multivariate Analysis VI*. North Holland, Amsterdam, 431–477.

Odell, P. L., and A. H. Feiveson, 1966 A numerical procedure to generate a sample covariance matrix. J. Amer. Stat. Ass. **61**: 199–203.

Pearson, K., 1901 On lines and planes of closest fit to systems of points in space. Phil. Mag. **2**: 559–572.

Peres-Neto, P. R., D. A. Jackson, and K. M. Somers, 2005 How many principal components? Stopping rules for determining the number of non-trivial axes revisited. Comp. Stat. Dat. Anal. **49**: 974–997.

Pinheiro, J. C., and D. M. Bates, 1996 Unconstrained parameterizations for variance-covariance matrices. Stat. Comp. **6**: 289–296.

Remadi, S., and Y. Amemiya, 1994 Asymptotic properties of the estimators for multivariate components of variance. J. Multiv. Anal. **49**: 110–131.

Ripley, B. D., 2004 Selection among large classes of models. In *Methods and Models in Statistics in Honour of Professor John Nelder, FRS*. Imperial College Press, London, 155–170.

Schäfer, J., and K. Strimmer, 2005 A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Stat. Appl. Genet. Mol. Biol. **4**: 32.

Seal, H. L., 1964 *Multivariate Statistical Analysis for Biologists*. Methuen, London.

Self, S. G., and K. Y. Liang, 1987 Asymptotic properties of the maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. J. Amer. Stat. Ass. **82**: 605–610.

Srivastava, M. S., and T. Kubokawa, 1999 Improved non-negative estimation of multivariate components of variance. Ann. Stat. **27**: 2008–2032.

Thompson, R., B. R. Cullis, A. B. Smith, and A. R. Gilmour, 2003 A sparse implementation of the Average Information algorithm for factor analytic and reduced rank variance models. Austr. New Zeal. J. Stat. **45**: 445–459.

Venables, W. N., 1973 Computation of the null distribution of the largest or smallest latent roots of a beta matrix. J. Multiv. Anal. **3**: 125–131.

Wolfinger, R. D., 1996 Heterogeneous variance-covariance structures for repeated measures data. J. Agric. Biol. Env. Stat. **1**: 205–230.

FIGURE 1. – Estimates of the first genetic eigenvalue (λ_{G1}) and angle (θ_1) between the corresponding estimated and true eigenvector, from a reduced rank analysis fitting the first principal component only, together with respective frequency distributions.
(Shown are estimates for 2000 replicates, simulating data for 1000 sires with 6 progeny each for 2 traits; gray vertical lines indicate the population eigenvalues.)

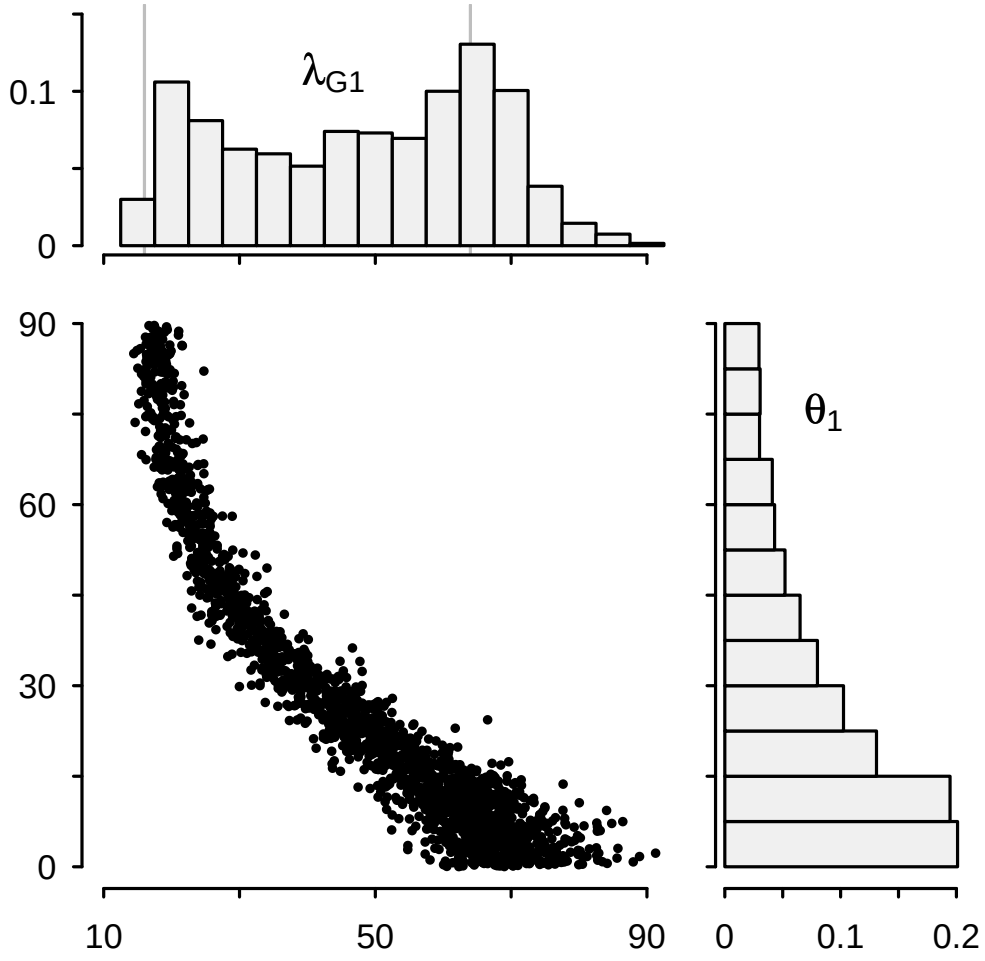

FIGURE 2. – Estimates of genetic eigenvalues (λ_i, i=1,6) from full rank analyses for low (Case A) and moderate (Case B) spread of true values (●) and different sample sizes. (▼ S5000N40, ▲ S1000N20, ■ S200N20, ◆ S100N6)

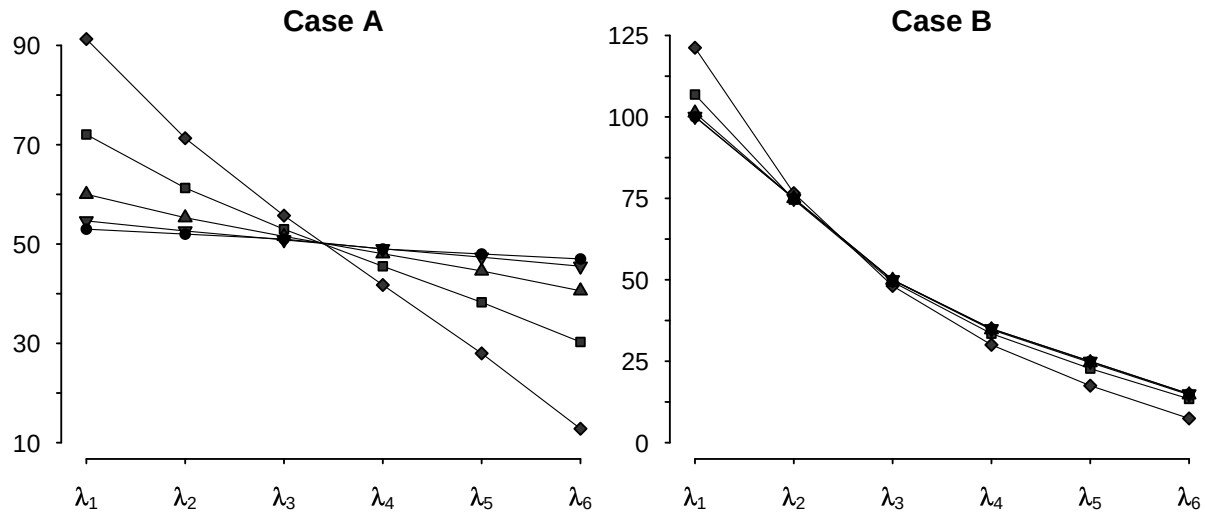

FIGURE 3. – Mean estimates of the first genetic eigenvalue (left; given as % deviation from population value) and direction of the corresponding eigenvector (right; given as deviation in ° from population vector) together with corresponding values predicted from population parameters, from analyses fitting one (Fit 1) or six (Fit 6) principal components and different environmental angles α_E.

(—— predicted values, ▲ mean estimates for S1000N20, ◆ mean estimates for S200N6; the horizontal gray line marks population eigenvalue)

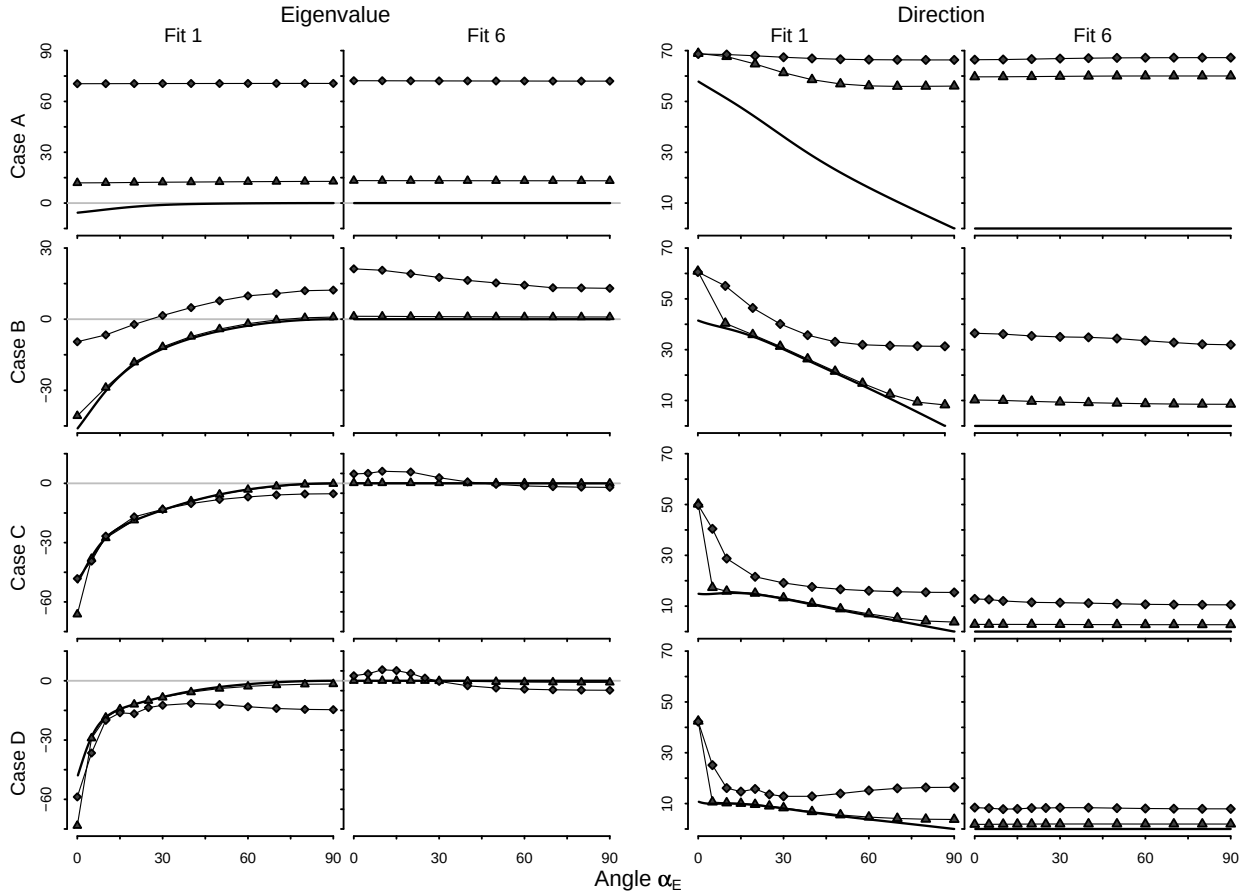

FIGURE 4. – Mean estimates of the first genetic principal component (top row: eigenvalue, given as % deviation from population value, bottom row: direction of eigenvector, given as deviation in ° from population vector) together with corresponding values predicted from population parameters, for reduced rank analyses fitting $m = 1, 2, 4, 5$ (Fit m) principal components, for case E and different environmental angles α_E.
(—— predicted values, ▲ mean estimates for S1000N20, ◆ mean estimates for S200N6; horizontal gray line marks population value for eigenvalues)

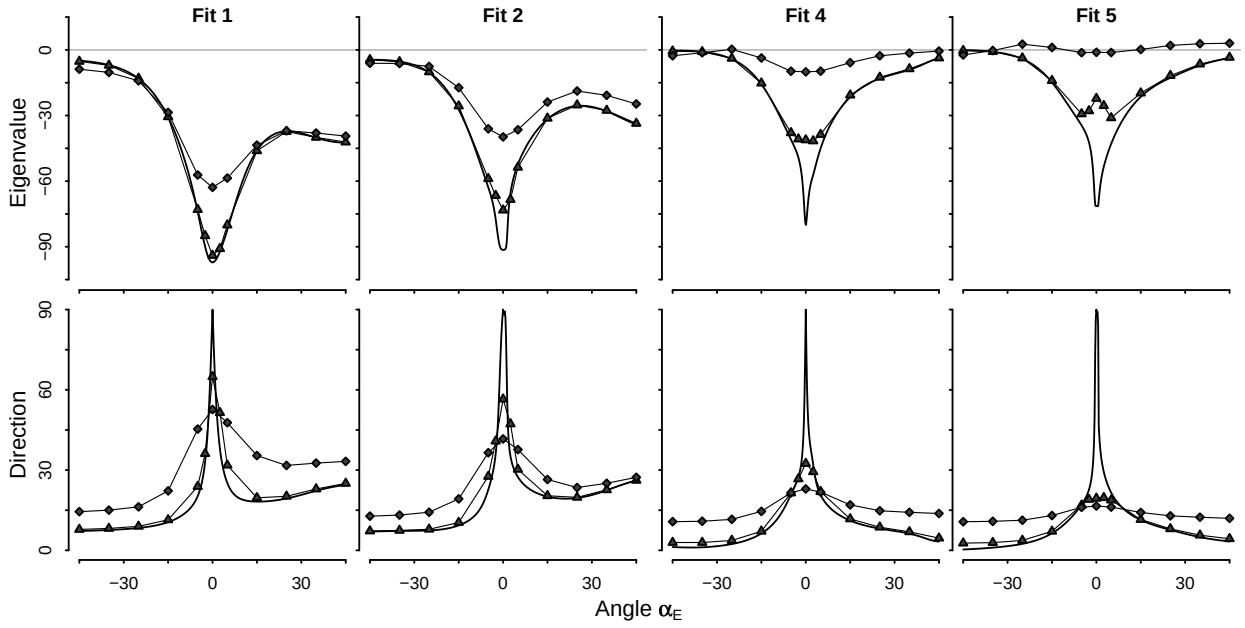

FIGURE 5. – Square root of mean square error for estimates of first genetic eigenvalue (top row), and mean error in estimates of the genetic covariance matrix (Σ_G, bottom row), for cases E and F with true rank of Σ_G of three, and different environmental angles α_E. (▲ $\alpha_E = -35°$, ● $\alpha_E = 0°$ and ◆ $\alpha_E = 35°$)

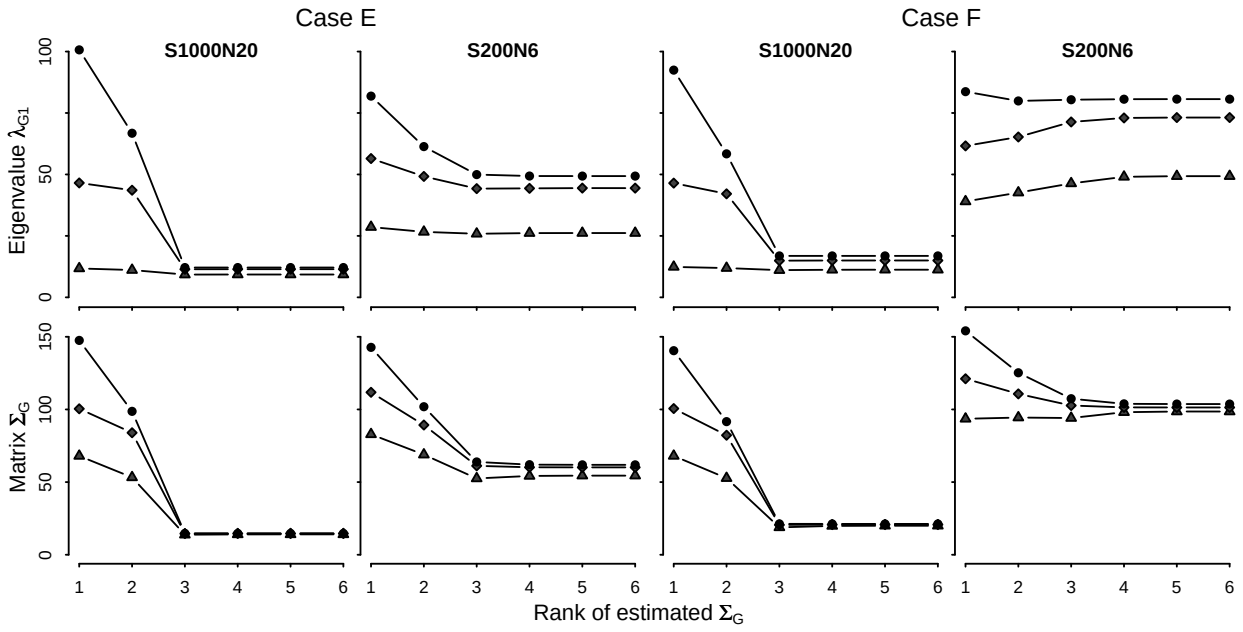

FIGURE 6. – Proportion of replicates (×100) for which the estimated genetic covariance matrix is found to have rank m, based on different test criteria (Q: Number of eigenvalues of $\mathbf{Q}$ greater than 1, AI: lowest AIC value, L: likelihood ratio test, Y: Y-score test, and BI: lowest BIC value), for population covariance matrices of rank 6 and 3, respectively.