A NOTE ON BIAS IN REDUCED RANK ESTIMATES OF COVARIANCE MATRICES

Karin Meyer[1] **and Mark Kirkpatrick**[2]

[1]Animal Genetics and Breeding Unit[3], University of New England, Armidale, NSW 2351
[2]Section of Integrative Biology, 1 University Station C-0930, University of Texas, Austin, Texas 78712

SUMMARY

Fitting only the leading principal components allows genetic covariance matrices to be modelled parsimoniously, yielding reduced rank estimates. If principal components with non-zero variances are omitted from the model, genetic variation is moved into the covariance matrices for residuals or other random effects. The resulting bias in estimates of genetic eigen-values and -vectors is examined.

INTRODUCTION

Direct estimation of only the leading principal components (PC) of genetic covariance matrices has been proposed to model variation among numerous traits parsimoniously and make efficient use of data (Kirkpatrick and Meyer 2004). Due to their orthogonality, we can increase the number of PCs fitted successively when considering a single matrix, i.e. estimates of the i−th PC remain constant for analyses fitting $k \geq i$ PCs. For quantitative genetic analyses, however, we consider at least two covariance matrices, genetic and environmental, simultaneously. This allows genetic covariances to be partitioned into the environmental components if PCs with non-zero eigenvalues are omitted. This note examines bias in estimates of genetic eigen-values and -vectors from reduced rank (RdR) analyses.

Table 1. MANOVA table

Source	d.f.[A]	MS	E[MS]
Between	$s-1$	$\mathbf{B}$	$\mathbf{\Sigma}_W + m\mathbf{\Sigma}_B$
Within	$s(m-1)$	$\mathbf{W}$	$\mathbf{\Sigma}_W$

[A] degrees of freedom

MATERIAL AND METHODS

Consider a balanced one-way classification with s independent groups, m individuals per group and q traits recorded for each individual. Let $\mathbf{B}$ and $\mathbf{W}$ denote the matrices of mean squares and cross-products (MS) between and within groups. This gives the multivariate analysis of variance (MANOVA) as shown in Table 1. Let $\mathbf{\Sigma}_G$ and $\mathbf{\Sigma}_E$ be the genetic and environmental covariance matrices among the p traits. Assume groups represent families whose members have degree of relationship α, so that $\mathbf{\Sigma}_B = \alpha\,\mathbf{\Sigma}_G$ and $\mathbf{\Sigma}_W = (1-\alpha)\,\mathbf{\Sigma}_G + \mathbf{\Sigma}_E$. This gives $\hat{\mathbf{\Sigma}}_W = \mathbf{W}$ and $\hat{\mathbf{\Sigma}}_B = (\mathbf{B}-\mathbf{W})/m$.

It is well known that for the balanced case, restricted maximum likelihood (REML) estimators of covariance components have closed form and are identical to those from MANOVA (e.g. Corbeil and Searle 1976; Lee and Kapadia 1984). However, REML estimates are, by definition, only valid if they are within the parameter space (Harville 1977), while MANOVA estimates of $\mathbf{\Sigma}_B$ can be negative-definite, i.e. yield estimates of covariance matrices which have negative eigenvalues. As demonstrated by Hill and Thompson (1978), the probability of this happening increases rapidly with increasing number of traits, decreasing sample or group size, and if $\mathbf{W}^{-1}\mathbf{B}$ has small eigenvalues. Due to sampling variation, eigenvalues of estimated matrices are dispersed more widely than the corresponding population values, leading to overestimates of the largest and underestimates of the smallest values, while their mean is estimated without bias (Hill and Thompson 1978). Hence, Hayes and Hill (1981) suggested to improve

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

the quality of estimates of $\mathbf{\Sigma}_B$ or, equivalently, $\mathbf{\Sigma}_G$ by regressing the eigenvalues of $\mathbf{W}^{-1}\mathbf{B}$ towards their mean, leaving the corresponding eigenvectors unchanged, a procedure termed 'bending'. In particular, this has been used to ensure that estimates of $\mathbf{\Sigma}_B$ were non-negative definite, choosing a degree of shrinkage so that the smallest, modified eigenvalue of $\hat{\mathbf{\Sigma}}_B$ was equal to zero. Closely related is the method of Amemiya (1985), who used the 'non-negative part' of $\mathbf{W}^{-1}\mathbf{B}$ only to estimate $\mathbf{\Sigma}_B$, which is equivalent to setting any negative estimates of eigenvalues of $\hat{\mathbf{\Sigma}}_B$ to zero. The author showed that the resulting estimators are REML estimators, imposing non-negativity constraints. This approach can be employed to examine the bias in RdR estimates of covariance matrices, fitting selected subsets of PCs.

Procedure. Both Hayes and Hill (1981) and Amemiya (1985) utilise the so-called canonical transformation, which simultaneously diagonalises two symmetric matrices. Steps involved are :

a) Determine a matrix $\mathbf{L}$ so that $\mathbf{L}'\mathbf{W}\mathbf{L} = \mathbf{I}$. The simplest choice is $\mathbf{L} = \mathbf{E}_W\mathbf{\Lambda}_W^{-1/2}$, with $\mathbf{\Lambda}_W$ the diagonal matrix of eigenvalues of $\mathbf{W}$ and $\mathbf{E}_W$ the matrix whose columns are its eigenvectors, i.e. $\mathbf{W} = \mathbf{E}_W\mathbf{\Lambda}_W\mathbf{E}_W'$. Alternatives are $\mathbf{L} = \mathbf{E}_W\mathbf{\Lambda}_W^{-1/2}\mathbf{E}_W'$ or $\mathbf{L} = \mathbf{U}^{-1}$, with $\mathbf{U}$ the Cholesky factor of $\mathbf{W}$.

b) Determine $\mathbf{Q} = \mathbf{L}'\mathbf{B}\mathbf{L} = \mathbf{E}_Q\mathbf{\Lambda}_Q\mathbf{E}_Q'$, with eigenvalues $\mathbf{\Lambda}_Q$ and eigenvectors $\mathbf{E}_Q$.

c) Obtain $\mathbf{P} = (\mathbf{L}')^{-1}\mathbf{E}_Q = \mathbf{\Lambda}_W^{1/2}\mathbf{E}_W'\mathbf{E}_Q$. This gives matrix $\mathbf{P}$ so that $\mathbf{P}\mathbf{P}' = \mathbf{W}$ and $\mathbf{P}\mathbf{\Lambda}_Q\mathbf{P}' = \mathbf{B}$.

Considering the first $k < q$ PCs of $\mathbf{\Sigma}_B$ only, RdR estimators are (from Amemiya 1985)

$$\hat{\mathbf{\Sigma}}_B = \left(\mathbf{P}\left(\mathbf{\Lambda}_Q^* - \mathbf{I}\right)\mathbf{P}'\right)/m$$

$$\hat{\mathbf{\Sigma}}_W = \left((s-1)\left(\mathbf{B} - m\hat{\mathbf{\Sigma}}_B\right) + s(m-1)\mathbf{W}\right)/(sm-1)$$

where $\mathbf{\Lambda}_Q^*$ is $\mathbf{\Lambda}_Q$ with diagonal elements λ_{Qi} replaced by $\lambda_{Qi}^* = 1$ for $i = k+1, \ldots, q$. If $\lambda_{Qi} > 1$ for all $i = 1, \ldots, k$, $\hat{\mathbf{\Sigma}}_B$ is positive definite with rank k.

Bias. For $k < q$, $\mathbf{\Sigma}_P = \mathbf{\Sigma}_B + \mathbf{\Sigma}_W$ the total variance and $\mathbf{p}_i$ denoting the i–th column of $\mathbf{P}$, this gives

$$\hat{\mathbf{\Sigma}}_B = \mathbf{\Sigma}_B - (1/m)\,\mathbf{\Delta} \qquad \text{or} \qquad \hat{\mathbf{\Sigma}}_G = \mathbf{\Sigma}_G - (\alpha^{-1}/m)\,\mathbf{\Delta}$$

$$\hat{\mathbf{\Sigma}}_w = \mathbf{\Sigma}_W + ((s-1)/(sm-1))\;\mathbf{\Delta} \qquad \text{or} \qquad \hat{\mathbf{\Sigma}}_E = \mathbf{\Sigma}_E + \left((s-1)/(sm-1) + (\alpha^{-1}-1)\right)\mathbf{\Delta}$$

$$\hat{\mathbf{\Sigma}}_P = \mathbf{\Sigma}_P - (1-1/m)/(sm-1)\,\mathbf{\Delta} \qquad \text{with} \qquad \mathbf{\Delta} = \mathbf{P}\left(\mathbf{\Lambda}_Q - \mathbf{\Lambda}_Q^*\right)\mathbf{P}' = \sum_{i=k+1}^{q} (\lambda_{Qi} - 1)\,\mathbf{p}_i\mathbf{p}_i'$$

Calculations. RdR estimates of $\mathbf{\Sigma}_G$ and $\mathbf{\Sigma}_E$ and their eigenvalues were obtained for two examples with a paternal half-sib design. Case 1 comprised $s = 500$ and $m = 10$ for two traits with a genetic correlation of $r_G = 0.5$. Trait 1 was assumed to have phenotypic variance of 1 and heritability (h^2) of 0.4. Variance and h^2 for trait 2 and the environmental correlation, r_E, were varied. Case 2 considered $\mathbf{\Sigma}_G$ and $\mathbf{\Sigma}_E$ for 8 traits measured by ultra-sound scanning of cattle with $s = 4000$ and $m = 4$, corresponding to an earlier analysis and simulation study (Meyer 2005).

RESULTS

Estimates of the genetic (λ_{Gi}) and environmental (λ_{Ei}) eigenvalues for case 1 are shown in Figure 1 together with the angle, θ_G, between estimates of the first genetic eigenvector and the first axis of the coordinate system. For equal h^2 and variances, $\theta_G = 45°$, and $\lambda_{Qi} = \lambda_{Bi}/\lambda_{Wi}$. As long as $r_E \leq r_G$, λ_{G1}

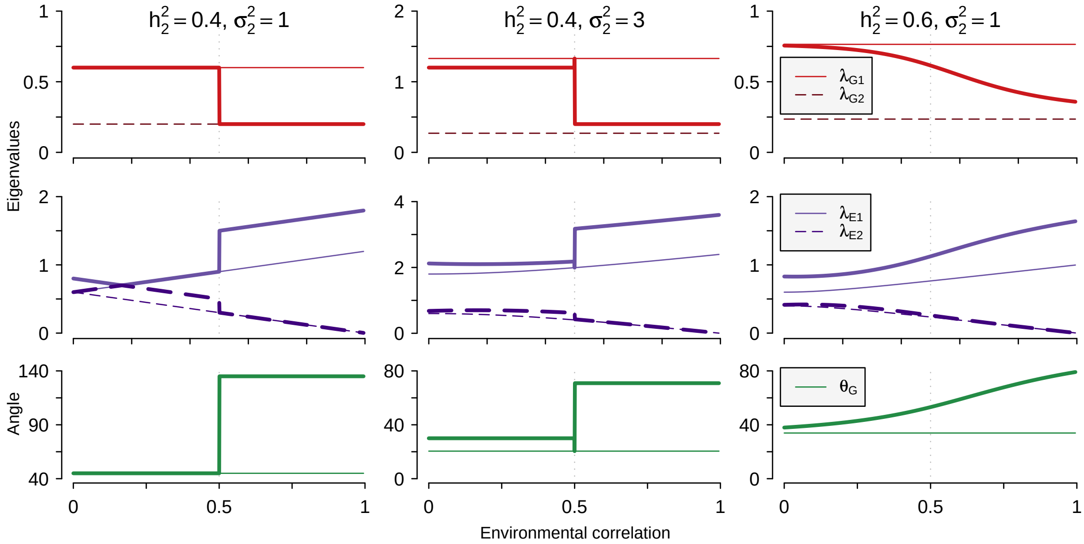

Figure 1. Estimates of eigenvalues and angles (in °) from reduced (thick lines) and full (thin lines) rank analyses.

and θ_G are estimated correctly when fitting the first PC only, while $\hat{\lambda}_{E2} = \lambda_{E2} + \lambda_{G2}$. For $r_E > r_G$, the order of the ratio of eigenvalues is reversed, i.e. $\lambda_{Q1} = \lambda_{B2}/\lambda_{W2}$. This causes the RdR estimate of the first PC to 'pick' up the second PC instead, so that $\hat{\lambda}_{G1} = \lambda_{G2}$, $\hat{\theta}_G = \theta_G + 90°$ and $\hat{\lambda}_{E1} = \lambda_{E1} + \lambda_{G1}$. Inspection of the profile likelihood for λ_{G1} and θ_G identified a saddle-point at the correct values for this scenario. A similar pattern emerges for equal h^2 but different variances. However, $\hat{\lambda}_{G1} = \lambda_{G1} - \delta$ for $r_E < r_G$ and $\hat{\lambda}_{G1} = \lambda_{G2} + \delta$ for $r_E > r_G$, while RdR estimates of λ_{G1}, λ_{E1} and θ_G are unbiased for $r_G = r_E$. For different h^2, bias in estimates changes less abruptly with r_E.

Figure 2 summarises estimates of the first 4 eigenvalues for case 2, for analyses fitting increasing numbers of PCs, F1,…,F8. With large s, bias in $\hat{\boldsymbol{\Sigma}}_P$ and its eigenvalues is negligible. For the first PC a strong downward bias in λ_{G1} and corresponding upward bias in λ_{E1} is evident until at least 4 PCs are fitted. With genetic eigenvalues of 97.9, 20.0, 13.7, 2.6, 1.8, 0.20, 0.17 and 0.01, the first

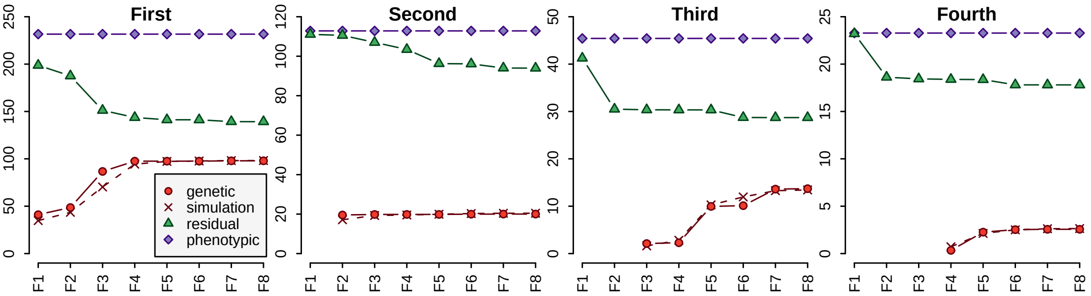

Figure 2. Estimates of the first 4 eigenvalues for analyses fitting n principal components (Fn).

4 PCs accounted for 98.4% of the total genetic variation. There is good agreement with earlier simulation results for $\lambda_{G\,i}$ (from Meyer 2005), even though the simulation assumed traits to be recorded on two distinct sets of animals and obtained estimates setting the respective environmental covariances to zero. Effects of omitting PCs on genetic eigenvectors are quantified by the angle between estimates from analysis Fi and F8, shown in Figure 3 together with corresponding means (open symbol) and empirical standard deviations from simulation. While the first two eigenvectors are estimated with little error if at least 3 PCs are fitted, estimates of the third and fourth eigenvectors deviate 80° or more for analyses F3 and F4. This is accompanied by a substantial downward bias in estimates of the corresponding eigenvalues, suggesting that we have, as observed for case 1,'picked' up one of the remaining PCs. Indeed, for $i = 3, 4, 5$ estimates of the i–th PC from analysis Fi deviated least from true PC $i + 1$.

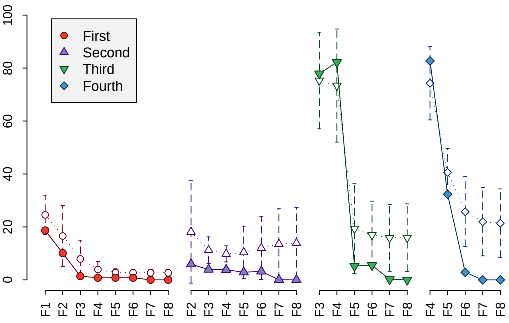

Figure 3. Angles (°) for eigenvectors 1 to 4.

DISCUSSION

Constraining the parameter space yields biased estimates of covariance components. This is well established for the non-negativity constraints commonly imposed in REML estimation, but is equally applicable to RdR estimation. The main difference is that we select the maximum rank of $\hat{\mathbf{\Sigma}}_B$ rather than a minimum value for its eigenvalues. Results show that estimates of the largest eigenvalues can be severely biased downwards if PCs explaining significant amounts of variation are ignored. Moreover, for certain constellations, there is a tendency for the estimate of the last PC fitted to 'pick' up one of the subsequent PCs instead. This implies that an estimate of $\lambda_{G\,i}$ close to zero from an analysis fitting i PCs does not necessarily indicate that i PCs suffice to model $\mathbf{\Sigma}_G$.

REFERENCES

Amemiya, Y. (1985) *Amer. Stat.* **39**:112.
Corbeil, R. R. and Searle, S. (1976) *Biometrics* **32**:779.
Harville, D. A. (1977) *J. Amer. Stat. Ass.* **72**:320.
Hayes, J. F. and Hill, W. G. (1981) *Biometrics* **37**:483.
Hill, W. G. and Thompson, R. (1978) *Biometrics* **34**:429.
Kirkpatrick, M. and Meyer, K. (2004) *Genetics* **168**:2295.
Lee, K. R. and Kapadia, C. H. (1984) *Biometrics* **40**:507.
Meyer, K. (2005) *Anim. Sci.* **81**:337.

ACKNOWLEDGEMENTS

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