CHEVERUD REVISITED: SCOPE FOR JOINT MODELLING OF GENETIC AND ENVIRONMENTAL COVARIANCE MATRICES

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

[1]Animal Genetics and Breeding Unit*, University of New England, Armidale, NSW 2351
[2]Section of Integrative Biology, University of Texas, Austin, Texas 78712

SUMMARY

Multivariate estimation fitting a common structure to estimates of genetic and environmental covariance matrices is examined in a simple simulation study. It is shown that such parsimonious estimation can considerably reduce sampling variation. However, if the assumption of similarity in structure does not hold at least approximately, bias in estimates of the genetic covariance matrix can be substantial. For small samples and more than a few traits, structured estimation is likely to reduce mean square error even if bias is quite large. Hence such models should be used cautiously.

INTRODUCTION

Accurate estimation of genetic covariances and correlations is inherently problematic as it requires substantial numbers of records on pairs of close relatives for all traits of interest, and as it may impose a considerable computational burden. Examining literature results, Cheverud (1988) found that estimates of genetic correlations for sets of traits such as body measurements are often very similar to their phenotypic counterparts. Others reported corresponding patterns for different natural populations (Roff 1995, 1996), in plants (Waitt and Levin 1998) and livestock (Koots and Gibson 1996; Kominakis 2003). Cheverud's suggestion to substitute estimates of phenotypic for genetic correlations, in particular when sample sizes are small or pedigree information is limited, has met with justifiable criticism (Willis et al. 1991; Kruuk et al. 2008). However, the idea of 'borrowing strength' from the phenotypic covariance matrix in estimating genetic covariances is appealing.

As multivariate analyses involving more than a few traits have become computationally feasible, there has been increasing interest in 'structured' estimation. A modern, mixed model based analogue to Cheverud's proposal might be to estimate genetic and phenotypic or environmental covariance matrices, imposing a common structure on the two matrices. This paper examines three alternatives to do so and their impact on estimates and their sampling properties.

MATERIAL AND METHODS

Structured estimation. Consider a multivariate analysis of q traits, fitting a simple animal model. Let $\mathbf{\Sigma}_G$ and $\mathbf{\Sigma}_E$ denote the covariance matrices for additive genetic and residual effects, respectively.
Unstructured. In most multivariate analyses, we assume covariance matrices are 'unstructured' (US), i.e. we describe the $q(q+1)/2$ distinct elements of each matrix by the corresponding number of parameters. A common parameterisation is to the elements of the Cholesky factor of a matrix.
Common correlation. To fit a common correlation (CORR) matrix, $\mathbf{R}$, we model $\mathbf{\Sigma}_G = \mathbf{S}_G\mathbf{R}\mathbf{S}_G$ and $\mathbf{\Sigma}_E = \mathbf{S}_E\mathbf{R}\mathbf{S}_E$, with $\mathbf{S}_G$ and $\mathbf{S}_E$ the diagonal matrices of genetic and residual standard deviations.
Common principal components. Fitting common principal components (CPC), we assume that both matrices have the same eigenvectors, $\mathbf{V}$, but different eigenvalues, i.e. $\mathbf{\Sigma}_G = \mathbf{V}\mathbf{\Lambda}_G\mathbf{V}'$ and $\mathbf{\Sigma}_E = \mathbf{V}\mathbf{\Lambda}_E\mathbf{V}'$ with $\mathbf{\Lambda}_G$ and $\mathbf{\Lambda}_E$ the diagonal matrices of genetic and residual eigenvalues.
Common GARP model. A related, common structure is obtained by modelling $\mathbf{\Sigma}_G = \mathbf{U}\mathbf{D}_G\mathbf{U}'$ and $\mathbf{\Sigma}_E = \mathbf{U}\mathbf{D}_E\mathbf{U}'$, with $\mathbf{U}$ a unitary, lower triangular matrix. The non-zero off-diagonal elements of $\mathbf{U}$ have an interpretation as regression coefficients in an auto-regressive model, hence the acronym

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

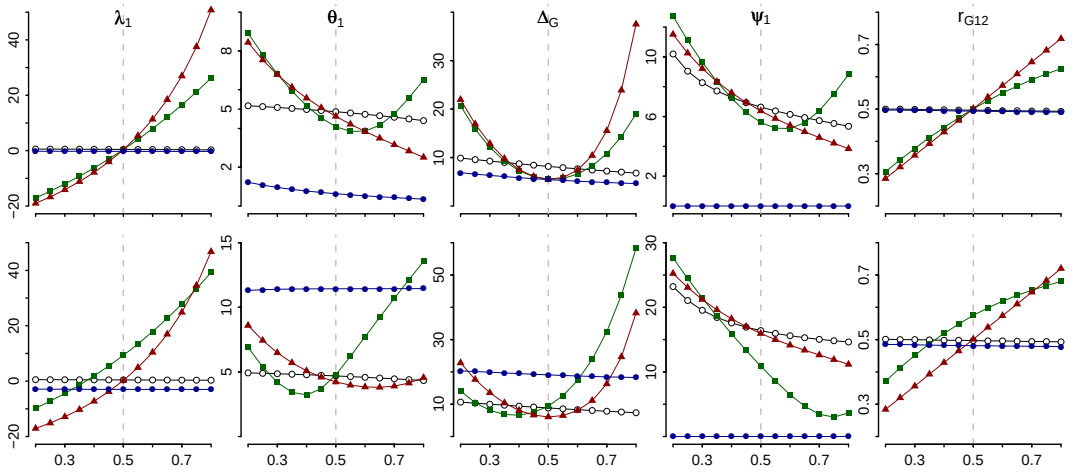

Figure 1. Mean estimates (see text for definitions) for example I, for equal (top row) or unequal (bottom row) heritabilities and r_E = 0.2 to 0.8 (○ US, ▲ CORR, ■ GARP, • CPC).

GARP, standing for generalised auto-regressive parameters (Pourahmadi et al. 2007). Correspondingly, the elements of the diagonal matrices $\mathbf{D}_G$ and $\mathbf{D}_E$ represent the 'innovation' variances, i.e. for variable i the conditional variance given variables 1 to $i-1$.

Parsimony. Each of the 3 structured models reduces the number of parameters to be estimated by $q(q-1)/2$, i.e. from $p = q(q+1)$ in the US case (for $\mathbf{\Sigma}_G$ and $\mathbf{\Sigma}_E$) to $p = q(q+3)/2$.

Simulation. Behaviour of restricted maximum likelihood (REML) estimates for the 4 different parameterisations was examined considering a simple, balanced paternal half-sib design (s sires with n progeny each). This involved sampling of the matrices of mean squares between and within families from appropriate central Wishart distributions, performing 10 000 replicates for each scenario considered. Maximisation of the likelihood, constraining both $\hat{\mathbf{\Sigma}}_G$ and $\hat{\mathbf{\Sigma}}_E$ to be positive definite, was carried out using a Method of Scoring algorithm combined with a derivative-free search.

Summary statistics. Means over replicates were calculated for estimates of genetic correlations ($r_{G\,ij}$), the eigenvalues of $\hat{\mathbf{\Sigma}}_G$ (λ_i), the log likelihood ($\log \mathcal{L}$), and (for $\mathbf{V}_X = \{\mathbf{v}_{X\,i}\}$ in $\mathbf{\Sigma}_X = \mathbf{V}_X\mathbf{\Lambda}_X\mathbf{V}'_X$)

- the angle between i–th eigenvectors of $\mathbf{\Sigma}_G$ and $\hat{\mathbf{\Sigma}}_G$: $\theta_i = (180/\pi)\arccos|\hat{\mathbf{v}}'_{G\,i}\mathbf{v}_{G\,i}|$
- the angle between i–th eigenvectors of $\hat{\mathbf{\Sigma}}_G$ and $\hat{\mathbf{\Sigma}}_E$: $\psi_i = (180/\pi)\arccos|\hat{\mathbf{v}}'_{G\,i}\hat{\mathbf{v}}_{E\,i}|$
- the 'quadratic loss' in $\hat{\mathbf{\Sigma}}_G$: $\Delta_G = \mathrm{tr}(\hat{\mathbf{\Sigma}}_G\mathbf{\Sigma}_G^{-1} - \mathbf{I})^2$
- the mean squared difference in $\hat{r}_{G\,ij}$ and $\hat{r}_{E\,ij}$ (in %): $\Delta_R = \sum_{i=1}^{q}\sum_{j=i+1}^{q}(\hat{r}_{G\,ij} - \hat{r}_{E\,ij})^2/(q(q-1)/2)$
- the 'adjusted' Akaike information criterion: $\mathrm{AIC} = -2\log\mathcal{L} + 2p(1 + \frac{p+1}{qsn-p-1})$

Example I. Example I comprised $q = 2$ traits with a genetic correlation of $r_{G\,12} = 0.5$ and equal phenotypic variances ($\sigma^2_P = 100$), for a moderate sample size ($s = 500$ with $n = 8$). Environmental correlations considered were $r_{E\,12} = 0.2$ to 0.8. Scenario A assumed heritabilities for both traits were equal ($h^2_1 = h^2_2 = 0.3$), while scenario B involved different values ($h^2_1 = 0.36, h^2_2 = 0.24$).

Example II. The second example involved $q = 6$, again using equal parameters of $h^2_i = 0.33$, $r_{G\,ij} = r_{E\,ij} = 0.5$ and $\sigma^2_P = 100$ for all traits to construct population values for $\mathbf{\Sigma}_G$ and $\mathbf{\Sigma}_E$. $\mathbf{\Sigma}_E$ was then replaced by $\mathbf{T}\mathbf{\Sigma}_E\mathbf{T}'$, with $\mathbf{T} = \prod_{i<j}^{q}\mathbf{C}(\alpha)^{ij}$ and $\mathbf{C}(\alpha)^{ij}$ a rotation matrix with elements $c_{ii} = c_{jj} = \cos(\alpha)$, $c_{kk} = 1$ for $k \neq i, j$, $c_{ij} = \sin(\alpha)$, $c_{ji} = -c_{ij}$ and zero otherwise. Rotation angles from $\alpha = 0°$ to 6° (equal for all i, j) were used to generate $\mathbf{\Sigma}_G$ and $\mathbf{\Sigma}_E$ with increasingly different eigenvectors. Three sample sizes, $s = 1000, n = 20$, $s = 500, n = 10$ and $s = 250, n = 8$, were examined.

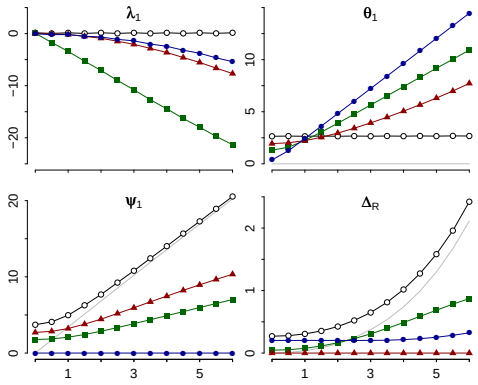

Figure 2. Means statistics for example II (s=1000, n=20; see Figure 1 for legend).

RESULTS

Means of summary statistics for example I are summarised in Figure 1. For equal heritabilities, eigenvectors of $\mathbf{\Sigma}_G$ and $\mathbf{\Sigma}_E$ are collinear regardless of the value of r_{E12} (shown along the horizontal axes). Hence CPC is the correct model throughout, and estimates of λ_1 (expressed here as % deviation from population value) and $r_{G\,12}$ for CPC and US are virtually the same. Fitting CPCs thus reduces sampling variation in the direction of the first genetic eigenvector (θ_1) substantially, and yields a consistently lower loss in $\hat{\mathbf{\Sigma}}_G$ (Δ_G) than the US model. For CORR and GARP, estimates of $r_{G\,12}$ are dominated by the population value for $r_{E\,12}$ (i.e. $\hat{r}_{G\,12}$ closely follows $r_{E\,12}$), with a corresponding bias in estimates of λ_1. The angle between the first eigenvectors of $\hat{\mathbf{\Sigma}}_G$ and $\hat{\mathbf{\Sigma}}_E$ (ψ_1) is similar to that for US, except for high values of $r_{E\,12}$. With a moderate sample size, there is a narrow range of $r_{E\,12} \neq r_{G\,12}$ ($\approx 0.35 - 0.60$) for which these structured models reduce Δ_G compared to US.

For $h_1^2 \neq h_2^2$, however, CPC is no longer the correct model, with the angle between the first eigenvectors of $\mathbf{\Sigma}_G$ and $\mathbf{\Sigma}_E$ ranging from 22.7° ($r_{E\,12} = 0.2$) to 14.6° ($r_{E\,12} = 0.8$). Fitting CPC for this case, estimates of λ_1 and $r_{G\,12}$ are little affected, but estimates of the direction of eigenvectors are heavily biased, with a correspondingly large loss Δ_G. Nevertheless, CPC appears advantageous over both CORR and GARP for larger differences between $r_{G\,12}$ and $r_{E\,12}$.

Results for example II are given in Figure 2, with different values of the rotation angle α along the horizontal axes. The population value for ψ_1 (shown as smooth gray line) increases linearly with the α used, causing estimates of θ_1 to increase similarly when fitting CPC. Again, estimates of λ_1 are relatively little biased, even if the CPC model is grossly wrong. True differences in Δ_R (gray line) increase quadratically with α. For US analyses, estimates of Δ_R are consistently larger, reflecting marked sampling variation. All three structured models underestimate differences in genetic and environmental correlations for values of α larger than $\approx 3°$.

Corresponding values for Δ_G together with the proportion of replicates for which each model fitted 'best', based on the value of AIC, are shown in Figure 3. With a difference of 15 parameters between US and structured models, the latter can provide estimates of $\mathbf{\Sigma}_G$ with substantially lower quadratic loss than US, especially for small samples. While CORR and GARP appeared advantageous over CPC in terms of Δ_G, model selection on the basis of AIC generally favoured CPC over the other structured models, decreasingly so as α increased. Bias in both the individual parameters and Δ_G increased faster with α for CPC than for CORR or GARP. AIC is derived from $\log \mathcal{L}$ and thus

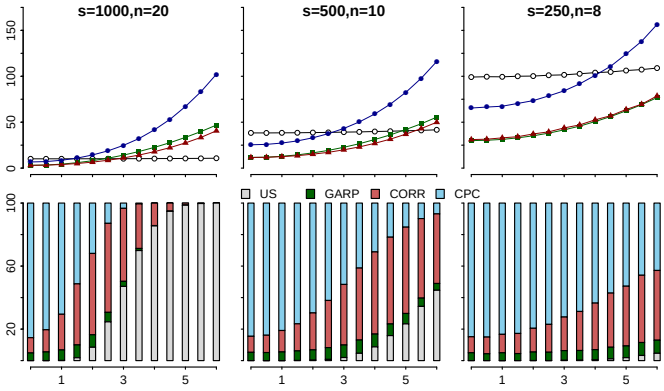

Figure 3. Estimates of Δ_G and proportion of samples (in %) for which each model fitted 'best' for example II (○ US, ▲ CORR, ■ GARP, • CPC).

dominated by $\hat{\mathbf{\Sigma}}_E$, i.e. such model selection by and large does not aim at minimising loss in $\hat{\mathbf{\Sigma}}_G$.

DISCUSSION

It has been demonstrated that joint modelling of the genetic and environmental covariance matrix in mixed model analyses is readily feasible, and can result in reduced sampling variation. This resulted in 'improved' estimates of $\mathbf{\Sigma}_G$, i.e estimates with a smaller quadratic loss than unstructured estimates in a range of scenarios. Of the three alternative parameterizations considered, none proved best overall. The common principal components model tended to yield least biased estimates of the first genetic eigenvalue. Reduction in mean square errors and loss generally comes at the price of bias in estimates. Disconcertingly, standard likelihood based model selection procedures (AIC) appeared to favour parsimonious models imposing a common structure for a range of cases where this was not the appropriate model, in spite of accounting for the number of parameters estimated. For small samples in particular, Δ_G somewhat higher than in the US case and thus potentially non-negligible bias seemed to be tolerated. Further work is necessary to determine the best strategy for model selection in practical applications.

A less rigid alternative to the assumption of a common structure may be a 'shrinkage' of the estimated genetic towards the phenotypic covariance matrix. While this does not reduce the number of parameters to be estimated, it can reduce sampling variation in $\hat{\mathbf{\Sigma}}_G$ and thus Δ_G. For instance, we could maximise $\log \mathcal{L}$ subject to a penalty which measures the divergence between $\hat{\mathbf{\Sigma}}_G$ and $\hat{\mathbf{\Sigma}}_G + \hat{\mathbf{\Sigma}}_E$. This is similar in spirit to the 'bending' procedure proposed by Hayes and Hill (1981). Preliminary analyses have been promising, showing a marked reduction of loss in $\hat{\mathbf{\Sigma}}_G$ even for mild penalties accompanied by relatively small bias.

CONCLUSION

Structured estimation provides a powerful tool to increase the accuracy of genetic parameter estimation, especially for multivariate analyses comprising more than a few traits and smaller sample sizes, and is readily implemented in a mixed model framework. However, as always, there is the trade-off between a reduction in sampling variation and bias. The utility of such analyses depends very much on the underlying assumption of a common structure to hold at least approximately – while parsimonious estimation may yield estimates with reduced loss or mean square error, this may be at the expense of substantial bias. Structured estimation appears most promising when sample sizes are small, but such models are not a substitute for using data sets of sufficient size.

ACKNOWLEDGEMENTS

This work was supported by Meat and Livestock Australia under grant BFGEN.100B (KM) and National Science Foundation grants EF-0328594 and DEB-0819901 (MK).

REFERENCES

Cheverud, J. M. (1988) *Evolution* **42**:958.
Hayes, J. F. and Hill, W. G. (1981) *Biometrics* **37**:483.
Kominakis, A. P. (2003) *J. Anim. Breed. Genet.* **120**:269.
Koots, R. and Gibson, J. P. (1996) *Genetics* **143**:1409.
Kruuk, L. E. B., Slate, J. and Wilson, A. J. (2008) *Ann. Rev. Ecol. Evol. System.* **39**:525.
Pourahmadi, M., Daniels, M. J. and Park, T. (2007) *J. Multiv. Anal.* **98**:569.
Roff, D. A. (1995) *Heredity* **74**:481.
Roff, D. A. (1996) *Evolution* **50**:1392.
Waitt, D. E. and Levin, D. A. (1998) *Heredity* **80**:310.
Willis, J. H., Coyne, J. A. and Kirkpatrick, M. (1991) *Evolution* **45**:441.