Proc. Assoc. Advmt. Anim. Breed. Genet. 20:424-427

POST-ESTIMATION PENALIZATION: MORE 'PEP' FOR ESTIMATES OF GENETIC COVARIANCE MATRICES

Karin Meyer

Animal Genetics and Breeding Unit*, University of New England, Armidale, NSW 2351

SUMMARY

Maximum likelihood estimation of genetic covariances subject to a penalty to reduce sampling variation has been shown to yield improved estimates, especially for analyses comprising many traits. However, this can increase computational requirements substantially. Similarly, penalties have been found to be beneficial in a maximum likelihood based approach for pooling results from analyses of subsets of traits. This paper examines the scope for using the latter method to apply penalties to results from multivariate analyses in a computationally undemanding post-estimation step. A simulation study is presented demonstrating that even slight changes to estimates in this way can result in 'regularized' values markedly closer to population values than standard, unpenalized estimates.

INTRODUCTION

Restricted maximum likelihood (REML) estimation of genetic covariance matrices subject to a penalty to borrow strength from their phenotypic counterparts has been shown to 'improve' estimates, i.e. to result in estimates which are, on average, closer to the population values than standard (unpenalized) estimates (Meyer and Kirkpatrick 2010; Meyer 2011b). Whilst highly appealing, penalized estimation can increase computational requirements by orders of magnitude. This may be prohibitive for multivariate analyses comprising numerous traits where penalization is likely to be most beneficial. Recently, Meyer (2013) demonstrated that penalization can also yield 'better' estimates when employing a maximum likelihood approach to combine estimates from analyses of overlapping subsets of traits to construct overall covariance matrices. This suggests that the same procedure might be used to modify estimates from a single, unpenalized multivariate analysis in a simple, computationally undemanding post-estimation penalization (PEP) step. This paper presents a simulation study examining the scope for PEP.

PENALIZING ESTIMATES

Penalized REML estimates are obtained by maximising the log likelihood ($\log \mathcal{L}$) in a multivariate analysis subject to a penalty ($\mathcal{P}$), $\log \mathcal{L} - \frac{1}{2}\psi\mathcal{P}$, with $\mathcal{P}$ a suitable function of the covariance components to be estimated and $\psi \geq 0$ the so-called tuning factor determining the stringency of penalization. For PEP, unpenalized estimates ($\psi = 0$) of covariance matrices are first obtained preforming a standard, multivariate analysis. In a second step, these are 'converted' to 'data' by forming matrices of mean squares and crossproducts corresponding to a selected simple, balanced pedigree structure from the estimates. Together with the assumed pseudo pedigree, these matrices then provide a likelihood function which again is maximised subject to a penalty. Further details are given in Meyer (2013).

MATERIAL AND METHODS

Data for 10 traits were simulated for 250 independent families of size 8, as per Bondari *et al.* (1978)'s design, sampling genetic and residual effects from appropriate multivariate Normal distributions for two sets of population parameters. For case A, all heritabilities were assumed equal to 0.4, for case B values ranged from 0.6 to 0.2, $0.2 + 0.1 \bmod(i, 5)$ for trait i. All genetic correlations were set to 0.5 and all residual values to 0.2. Phenotypic variances for the $i-$th trait were $\bmod(i, 3) + 1$.

*AGBU is a joint venture of NSW Department of Department of Primary Industries and the University of New England

This yielded canonical eigenvalues (λ_i) of 0.57 and 9×0.29 for case A and from 0.69 to 0.14 for B. A total of 250 replicates per case were carried out.

Analyses. Estimates of genetic ($\mathbf{\Sigma}_G$) and residual ($\mathbf{\Sigma}_E$) covariance matrices were obtained from multivariate REML analyses (MUV), with and without penalties. Unpenalized estimates were then modified by PEP, considering a paternal half-sib design (PHS) comprising $s = 2$ sires and $n = 2$ progeny per sire, a hierarchical full-sib design (HFS) with $s = 2$, $d = 2$ dams per sire and $n = 2$, and 2 families with Bondari's design (BON, $n = 8$) as pseudo pedigree structures. Penalties considered were

$$\mathcal{P}_\lambda = \textstyle\sum_i (\hat{\lambda}_i - \bar{\lambda})^2 \tag{1}$$

$$\mathcal{P}_\lambda^{\ell 2} = \textstyle\sum_i (\log(\hat{\lambda}_i) - \bar{\lambda}_1)^2 + (\log(1 - \hat{\lambda}_i) - \bar{\lambda}_2)^2 \tag{2}$$

$$\mathcal{P}_\Sigma = \log|\hat{\mathbf{\Sigma}}_G| + \mathrm{tr}(\hat{\mathbf{\Sigma}}_G^{-1}\hat{\mathbf{\Sigma}}_P^0) + \log|\hat{\mathbf{\Sigma}}_E| + \mathrm{tr}(\hat{\mathbf{\Sigma}}_E^{-1}\hat{\mathbf{\Sigma}}_P^0) \tag{3}$$

$$\mathcal{P}_R = \log|\hat{\mathbf{R}}_G| + \mathrm{tr}(\hat{\mathbf{R}}_G^{-1}\hat{\mathbf{R}}_P^0) + \log|\hat{\mathbf{R}}_E| + \mathrm{tr}(\hat{\mathbf{R}}_E^{-1}\hat{\mathbf{R}}_P^0) \tag{4}$$

with $\bar{\lambda}$, $\bar{\lambda}_1$ and $\bar{\lambda}_2$ the means of estimates $\hat{\lambda}_i$, $\log(\hat{\lambda}_i)$ and $\log(1 - \hat{\lambda}_i)$, respectively, $\hat{\mathbf{\Sigma}}_P^0$ the unpenalized estimate of the phenotypic covariance matrix, $\hat{\mathbf{R}}_G$ and $\hat{\mathbf{R}}_E$ the estimates of the genetic and residual correlation matrix, and $\hat{\mathbf{R}}_P^0$ their unpenalized, phenotypic counterpart. In addition, simple 'bending' (BEN) was applied, regressing $\hat{\lambda}_i$ towards $\bar{\lambda}$, as proposed by Hayes and Hill (1981).

Degree of penalization. Tuning factors for each replicate were determined as values of ψ for which a) the sum of losses in $\hat{\mathbf{\Sigma}}_G$ and $\hat{\mathbf{\Sigma}}_E$ was smallest ("Optimum"), and b) the largest value for which the deviation (absolute value) of $\log \mathcal{L}$ from the (unpenalized) maximum did not exceed $\chi^2_{1,5\%} = 1.92$ ("$\Delta\mathcal{L}$"). In addition, fixed values selected to provide "very mild" and "mild" penalties were used, c) $\psi = 0.1$ for MUV and $\psi = 0.001$ for PEP, and d) $\psi = 1.0$ (MUV) and $\psi = 0.01$ (PEP). For BEN, regression coefficients were set to 0.98 for "very mild" and 0.90 for "mild" shrinkage.

Summary statistics. The deviation of estimated covariance matrices ($\hat{\mathbf{\Sigma}}$) for q traits from the respective population values ($\mathbf{\Sigma}$) was evaluated as the so-called entropy loss (L_1) and, with $\bar{L}_1(\cdot)$ denoting the mean over replicates and $\hat{\mathbf{\Sigma}}^\psi$ the estimate for a tuning factor of ψ, the percent reduction in average loss (PRIAL),

$$L_1(\mathbf{\Sigma}, \hat{\mathbf{\Sigma}}) = \mathrm{tr}(\mathbf{\Sigma}^{-1}\hat{\mathbf{\Sigma}}) - \log|\mathbf{\Sigma}^{-1}\hat{\mathbf{\Sigma}}| - q \qquad \text{and} \tag{5}$$

$$\mathrm{PRIAL} = 100\big[1 - \bar{L}_1\big(\mathbf{\Sigma}, \hat{\mathbf{\Sigma}}^\psi\big)/\bar{L}_1\big(\mathbf{\Sigma}, \hat{\mathbf{\Sigma}}^0\big)\big]. \tag{6}$$

RESULTS

The distribution of losses in estimates of $\mathbf{\Sigma}_G$ for case B is summarized in Figure 1. Shown on the left of each panel are losses for unpenalized estimates from standard, multivariate analyses. Penalization using the optimum tuning factor (top panel) reduced both the mean and variation in losses dramatically for all penalties and both MUV and PEP. Moreover, simple 'bending' performed similar to a penalty encouraging shrinkage of the canonical eigenvalues towards their mean. In line with previous experience with MUV for cases with a substantial spread of population canonical eigenvalues (Meyer 2011b), a penalty shrinking correlation matrices towards their phenotypic counterpart ($\mathcal{P}_R$) was most effective, with MUV yielding a PRIAL of 74% and PEP of 61%.

In practice, the optimal tuning factor is unknown and, for MUV, estimating ψ using cross-validation techniques not only imposes a considerable computational burden but also has been found to reduce PRIALs achieved, typically by at least 10-15%. Hence, selecting a value of ψ which limits the change in $\log \mathcal{L}$ from the maximum (at $\psi = 0$) has been suggested as a simple, pragmatic alternative, and has been shown to yield losses $L_1(\cdot)$ closely related to optimal values (Meyer 2011a,b). As demonstrated

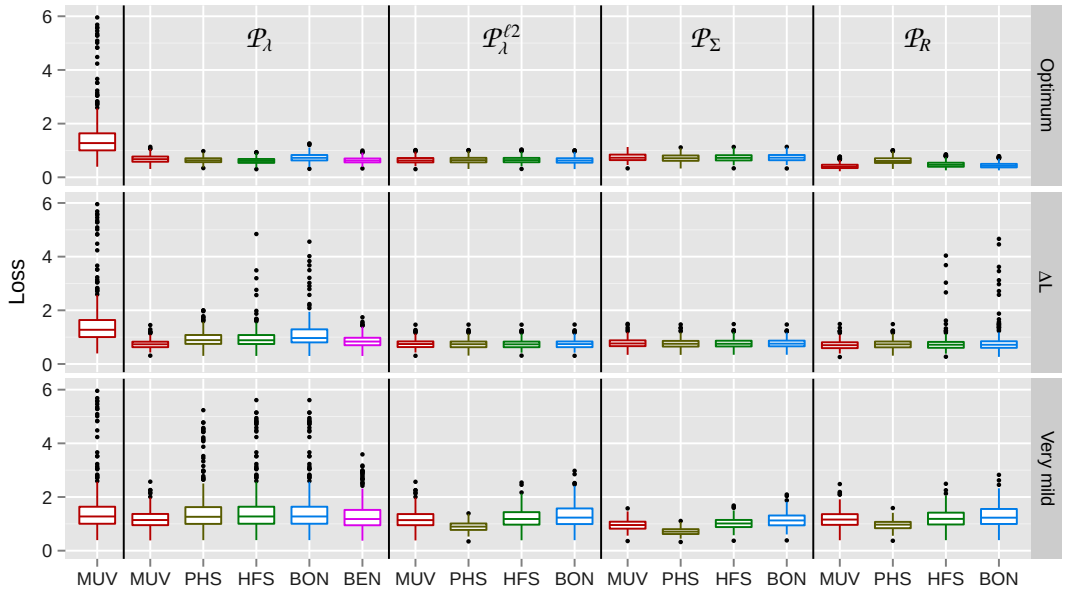

Figure 1. Distribution of entropy loss in estimates of the genetic covariance matrix for case B

in Figure 1 (middle panel), this strategy also performed well for PEP, especially for the simplest pseudo pedigree structure. For $\mathcal{P}_R$, PRIALs obtained were 55 and 53% for MUV and PEP, respectively. Limiting $\Delta\mathcal{L}$ to a value for which the change in even a single parameter was not statistically significant (at an error probability of 5%) yielded much milder penalization than for the optimum values of ψ, which resulted in average changes in $\log \mathcal{L}$ from -7.1 to -16.8. However, even such a mild penalty consistently provided substantial reductions in sampling variation and losses in estimates of the genetic covariance matrix. In contrast, whilst beneficial throughout, effects of penalization for a small, fixed value of ψ varied markedly with the type of penalty and pseudo-pedigree structure chosen.

Table 1 summarizes PRIALs and the corresponding mean change in $\log \mathcal{L}$ for selected examples. With 9 of the population canonical eigenvalues equal, stringent penalties on the λ_i, $\mathcal{P}_\lambda$ or $\mathcal{P}_\lambda^{\ell 2}$, performed best for case A, achieving optimum PRIALs (not shown) as high as 79% accompanied by changes in $\log \mathcal{L}$ around -17, with little difference between MUV and PEP. Conversely, choosing ψ on the basis of $\Delta\mathcal{L}$ was further from the optimum than for case B, but still achieved worthwhile PRIALs of more than 40% for MUV. Corresponding values for PEP were somewhat lower, but not too disconcertingly, especially as constellations of population values as for case A are uncommon in practice. Again, depending on the penalty, a fixed value of ψ resulted in substantial improvement in estimates of $\mathbf{\Sigma}_G$ for both cases, but with more fluctuations than the likelihood based choice.

With penalties designed to shrink both $\mathbf{\Sigma}_G$ and $\mathbf{\Sigma}_E$, a similar pattern of improvements was observed for estimates of $\mathbf{\Sigma}_E$ though PRIALs obtained were considerably lower, ranging from 14 to 28% for case A and 10 to 20 % for case B when selecting the tuning factor on the basis of $\Delta\mathcal{L}$. Corresponding values for estimates of $\mathbf{\Sigma}_P$ were small throughout, ranging from 0 to 3%.

DISCUSSION

Estimates of covariance components from multivariate analyses comprised of more than a few traits are subject to substantial sampling variation. Regularization can reduce this dramatically and thus yield estimates closer to the population values and, ultimately, result in better predictions of genetic merit and increased response to selection, in particular if weights for selection indices need

Table 1. Percentage reduction in average loss (PRIAL) and corresponding mean change in log likelihood ($\log \mathcal{L}$) for estimates of the genetic covariance matrix imposing different penalties.

Case	Value	Tune	$\mathcal{P}_\lambda$			$\mathcal{P}_\lambda^{\ell 2}$		$\mathcal{P}_\Sigma$		$\mathcal{P}_R$	
			MUV	PHS	BEN	MUV	PHS	MUV	PHS	MUV	PHS
A	PRIAL	$\Delta\mathcal{L}$	46	32	38	45	43	37	38	40	41
		mild	18	8	21	20	52	37	51	19	45
	$\log \mathcal{L}$	$\Delta\mathcal{L}$	-1.88	-1.89	-1.91	-1.88	-1.88	-1.83	-1.86	-1.89	-1.86
		mild	-0.19	-0.15	-0.50	-0.24	-3.02	-1.71	-11.58	-0.27	-2.34
B	PRIAL	$\Delta\mathcal{L}$	53	41	46	53	53	51	52	55	53
		mild	47	19	40	48	60	55	45	44	58
	$\log \mathcal{L}$	$\Delta\mathcal{L}$	-1.87	-1.88	-1.91	-1.86	-1.88	-1.84	-1.86	-1.87	-1.88
		mild	-0.81	-0.20	-1.02	-0.92	-7.12	-3.96	-13.88	-0.57	-3.16

to be derived from these estimates. REML estimation subject to a penalty provides such improved estimates but, while desirable, can be computationally demanding and accurate estimation of the optimum tuning factor remains problematic. Hence we propose a two-step procedure as alternative, in which standard, unpenalized estimates are modified post-estimation applying a mild penalty.

A suitable choice of the tuning factor may be based on limiting the change in $\log \mathcal{L}$ from the maximum to a relatively small value. For a limit corresponding to the significance threshold in a likelihood ratio test for one parameter, results showed reductions in loss in the range of 30 to 50%, and, except for a penalty on canonical eigenvalues on the original scale ($\mathcal{P}_\lambda$), differences to values for a penalized multivariate analyses were small. For an animal model with only two sources of variation, choosing a paternal half-sib design as pseudo-pedigree structure generally performed best.

REML estimates of covariance components are biased due to constraints on the parameter space. Improvements in estimates due to penalization generally come at the price of additional bias. While a mild penalty may not fully exploit the scope for reducing losses, the impact of penalization is not linear and such strategy can thus achieve a substantial proportion of the potential benefits at little effort. In addition, mild penalization will keep the extra bias created small and often result in estimates of individual components barely changed from unpenalized values.

CONCLUSIONS

Post-estimation penalization of multivariate estimates of covariance matrices using a likelihood approach teamed with a mild penalty can yield substantially improved estimates. It is recommended for routine analyses involving more than a few traits.

ACKNOWLEDGEMENTS

This work was supported by Meat and Livestock Australia under grant B.BFG.0050.

REFERENCES

Bondari K., Willham R.L. and Freeman A.E. (1978) *J. Anim. Sci.* **47**:358.
Hayes J.F. and Hill W.G. (1981) *Biometrics* **37**:483.
Meyer K. (2011a) *Proc. Ass. Advan. Anim. Breed. Genet.* **19**:83.
Meyer K. (2011b) *Genet. Sel. Evol.* **43**:39.
Meyer K. (2013) *J. Anim. Breed. Genet.* **130**:270.
Meyer K. and Kirkpatrick M. (2010) *Genetics* **185**:1097.