TO HAVE YOUR STEAK AND EAT IT : GENETIC PRINCIPAL COMPONENT ANALYSIS FOR BEEF CATTLE DATA

Karin Meyer

Animal Genetics and Breeding Unit, University of New England (a joint venture with NSW Department of Primary Industries), Armidale, NSW 2351, Australia

INTRODUCTION

Quantitative genetic analyses usually deal with several, if not many, correlated traits or effects. Generally, the matrices of covariances among these effects are considered to be 'unstructured', i.e. for k traits we have $k(k+1)/2$ distinct (co)variance components, and restrictions on estimates are imposed only to ensure that estimated matrices are positive semi-definite, i.e. do not have negative eigenvalues. In contrast, in other areas of statistics covariance matrices are often assumed to be structured. Parametric forms, such as compound symmetry or auto-regressive covariances (e.g. Jennrich and Schluchter 1986; Wolfinger 1996) are common assumptions for longitudinal or spatial data. Alternative parameterisations are based on the eigen-vectors and -values of the covariances matrices concerned. In particular, principal component (PC) analysis is widely utilised to summarise multivariate information and as a dimension reduction technique.

So far, PC analyses (PCA) for genetic (or other random) effects have by and large been carried out in 2 steps, first obtaining full rank estimates of covariance matrices, and then carrying out an eigen-decomposition of the estimate. A better approach is to estimate the PCs directly, restricting estimation to the most important components only (Kirkpatrick and Meyer 2004). This is readily accommodated within the usual linear, mixed model framework, requiring only a simple reparameterisation. This paper reviews the direct estimation of PCs, and presents an application to an analysis of carcass traits of beef cattle.

PRINCIPAL COMPONENTS

What are principal components ? The PCs of a set of k correlated effects are simply a set of k variables which are a) linear functions of the effects, b) uncorrelated with each other, and c) successively explain a maximum of variation among the k effects.

Consider a random vector $\mathbf{v}$, representing k variables, with covariance matrix $\mathbf{\Sigma}$. Let $\mathbf{\Sigma} = \mathbf{E\Lambda E}'$, with $\mathbf{E}$ the matrix of eigenvectors and $\mathbf{\Lambda}$ the diagonal matrix of eigenvalues (λ_i) of $\mathbf{\Sigma}$. The PCs are then $\mathrm{PC}_i = \mathbf{e}_i'\mathbf{v}$, where $\mathbf{e}_i$ is the $i-$th column of $\mathbf{E}$, with $\mathbf{e}_i'\mathbf{e}_i = 1$. Hence, the weights or 'loadings' for individual effects in PC_i are the elements of the corresponding eigenvector, $\mathbf{e}_i$. The variance of PC_i is given by the corresponding eigenvalue, λ_i. As shown in Figure 1 for 2 traits, the transformation to PCs can be thought of as a rotation of the data space, replacing the original co-ordinate system by the axes of the data ellipsoid.

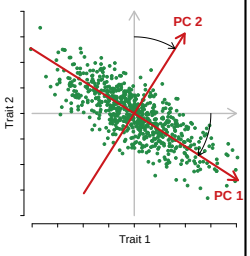

Figure 1. Two PCs

As is standard practice, assume that eigen-vectors and -values are given in descending order of the λ_i. The $i-$th PC then explains the maximum amount of variance, given PC_1 to PC_{i-1}. (e.g. Jollife 1986). For a given number of terms, $m < k$, PCs provide the expansion for which the

error of truncation is minimised. Moreover, any PCs with λ_i close to zero contribute virtually no information that is not already contained in the leading PCs. Hence, these components can be ignored, with negligible loss of information. Often, the bulk of variation is explained by the first few PCs, and m can be considerably smaller than k. This is the principle underlying the use of PCs to reduce 'dimensions'. Considering the first m PCs only to model the covariance of $\mathbf{v}$ gives

$$\mathbf{\Sigma}^{\star} = \mathbf{E}_m \mathbf{\Lambda}_m \mathbf{E}'_m = \sum_{i=1}^{m} \lambda_i \, \mathbf{e}_i \mathbf{e}'_i \tag{1}$$

where $\mathbf{E}_m$ is $\mathbf{E}$ truncated to the first m columns, and $\mathbf{\Lambda}_m$ is the corresponding, $m \times m$ submatrix of $\mathbf{\Lambda}$. Covariance matrix $\mathbf{\Sigma}^{\star}$, of size $k \times k$ has rank m, and is determined by $m(2k - m + 1)/2$ parameters, m values λ_i and $m(2k-m-1)/2$ elements of $\mathbf{E}_m$. While $\mathbf{E}_m$ has km non-zero elements, the orthogonality constraint $\mathbf{E}_m \mathbf{E}'_m = \mathbf{I}_m$ (where $\mathbf{I}_m$ denotes an identity matrix of size m) reduces the number of 'free' parameters (Kirkpatrick and Meyer 2004).

Factor analysis. Closely related to PCA, but with a different emphasis, is factor analysis (FA). While PCA is predominantly concerned with identifying variables explaining maximum variation, FA is about attributing covariances between effects to common factors. In FA, we fit a model $\mathbf{v} = \mathbf{F}_m \mathbf{z} + \boldsymbol{\epsilon}$ to our vector of random effects, with m latent variables $\mathbf{z} \sim N(\mathbf{0}, \mathbf{I}_m)$ and errors $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{\Psi})$. Factor loadings are given by $\mathbf{F}_m = \mathbf{E}_m \mathbf{\Lambda}_m^{1/2}$, i.e. eigen-vectors $\mathbf{e}_i$ scaled by $\sqrt{\lambda_i}$, and $\mathbf{\Psi} = \text{Diag}\{\sigma^2_{\epsilon\, i}\}$ is the matrix of specific variances. This gives covariance matrix

$$\mathbf{\Sigma}^{\bullet} = \mathbf{F}_m \mathbf{F}'_m + \mathbf{\Psi} = \mathbf{E}_m \mathbf{\Lambda}_m \mathbf{E}'_m + \mathbf{\Psi} = \sum_{i=1}^{m} \lambda_i \, \mathbf{e}_i \mathbf{e}'_i + \text{Diag}\{\sigma^2_{\epsilon\, i}\} \tag{2}$$

Generally, $\mathbf{\Sigma}^{\bullet}$ has rank k and involves $m(2k - m + 1)/2 + k$ parameters. This implies that m needs to be sufficiently smaller than k, so that the total number of parameters does not exceed $k(k+1)/2$. If all $\sigma_{\epsilon\, i}$ are assumed to be zero, $\mathbf{\Sigma}^{\bullet}$ reduces to $\mathbf{\Sigma}^{\star}$. A typical application for FA models is the study of genotype × environment interactions; see Smith *et al.* (2001) for an example.

Principal components in the mixed model. Consider the usual linear, mixed model

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon} \tag{3}$$

with $\mathbf{y}$, $\mathbf{b}$, $\mathbf{u}$ and $\boldsymbol{\epsilon}$ the vectors of observations, fixed effects, random effects and residuals, and $\mathbf{X}$ and $\mathbf{Z}$ the respective incidence matrices. Model (3) is general, and encompasses multiple random effects, and standard multi-trait as well as random regression analyses. For simplicity of presentation, however, let (3) represent a simple animal model, with $\mathbf{u}$ the vector of additive genetic effects for k traits and N animals, and $\text{Var}(\mathbf{u}) = \mathbf{\Sigma} \otimes \mathbf{A}$. '$\otimes$' denotes the direct matrix product, and $\mathbf{A}$ is the numerator relationship matrix between animals. Reparameterise (3) to

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\,(\mathbf{Q} \otimes \mathbf{I}_N)\,(\mathbf{Q}^{-1} \otimes \mathbf{I}_N)\mathbf{u} + \boldsymbol{\epsilon} = \mathbf{X}\mathbf{b} + \mathbf{Z}^{\star}\mathbf{u}^{\star} + \boldsymbol{\epsilon} \tag{4}$$

For $\mathbf{Q}$=$\mathbf{E}$, (3) and (4) are equivalent models and, for $\mathbf{\Sigma} = \mathbf{E}\mathbf{\Lambda}\mathbf{E}'$ as above, $\mathbf{u}^{\star}_j = \mathbf{E}'\mathbf{u}_j$ is the vector of genetic PCs for animal j (with $\mathbf{u}_j$ the subvector of $\mathbf{u}$ for animal j). Choosing $\mathbf{Q} = \mathbf{E}_m$, (4) becomes a model fitting the first m PCs only, and $\mathbf{u}^{\star}_j$ has length m. Estimates of breeding values (EBVs) for the original traits from such reduced model are readily obtained as $\hat{\mathbf{u}}_j = \mathbf{E}\hat{\mathbf{u}}^{\star}_j$.

Reduced rank estimation. Standard, mixed model based methods to estimate covariance components can be applied to estimate eigen-vectors and -values directly. The main difference to multi-variate analyses under model (3) is that the parameters to be estimated are part of the incidence matrices of random effects, i.e. that $\partial \mathbf{Z}^{\star} / \partial q_{rs} = \mathbf{Z}\,(\mathbf{D}_{rs} \otimes \mathbf{I}_N)$, with q_{rs} the $rs-$th element of $\mathbf{Q}$ and $\mathbf{D}_{rs}$ a matrix whose $rs-$th element is unity and zero otherwise.

Restricted maximum likelihood (REML) estimation for model (4) has been considered in detail by Meyer and Kirkpatrick (2005). Choosing $\mathbf{Q} = \mathbf{E}_m\mathbf{\Lambda}^{1/2}$, allows the elements of $\mathbf{E}_m$ which are determined by orthogonality constraints on $\mathbf{E}$ to be obtained by solving a ssmall system of linear equations. Estimates of λ_i are then obtained simply by calculating the norms of the respective columns of $\mathbf{Q}$. This gives $\mathrm{Var}(\mathbf{u}^\star) = \mathbf{I}_{mN}$ and the REML log likelihood ($\log \mathcal{L}$) is

$$-2\log \mathcal{L} = const. + \log|\mathbf{A}| + \log|\mathbf{R}| + \log|\mathbf{C}| + \mathbf{y}'\mathbf{P}\mathbf{y} \quad (5)$$

with $\mathbf{R} = \mathrm{Var}(\mathbf{e})$, $\mathbf{C}$ the coefficient matrix in the mixed model equations for (4), and $\mathbf{y}'\mathbf{P}\mathbf{y}$ a weighted sum of squares of residuals. The likelihood is invariant to orthogonal transformations applied to $\mathbf{Q}$. Consider the Cholesky decomposition $\mathbf{\Sigma} = \mathbf{L}\mathbf{L}'$. A singular value decomposition of the Cholesky factor gives $\mathbf{L} = \mathbf{E}\mathbf{\Lambda}^{1/2}\mathbf{T}'$ with $\mathbf{T}\mathbf{T}' = \mathbf{I}$, i.e. the left singular vectors of $\mathbf{L}$ are equal to the eigen-vectors of $\mathbf{\Sigma}$ and the singular values are equal to $\sqrt{\lambda_i}$ (Harville 1997). Hence, choosing $\mathbf{T}'$ as transformation yields $\mathbf{Q} = \mathbf{L}$. In other words, we can obtain estimates of the first m eigen-vectors and -values of a covariance matrix, by estimating the non-zero elements of the first m columns of its Cholesky factor instead. Rotating the parameter space in this way automatically accounts for constraints on the number of parameters - there are $m(2k - m + 1)/2$ non-zero elements in the first m columns of $\mathbf{L}$ (Smith *et al.* 2001).

REML algorithms. The likelihood ($\log \mathcal{L}$) can be maximised using common optimisation techniques. In particular, the so-called 'average information' (AI) algorithm (Gilmour *et al.* 1995) is widely used. Meyer and Kirkpatrick (2005) describe an AI-REML procedure for reduced rank estimation via the leading PCs, using automatic differentiation of $\mathbf{C}$ to obtain first derivatives of $\log \mathcal{L}$. Similar calculations are involved in estimating the parameters of a FA model, but in addition to fitting $\mathbf{u}^*$ of length mN, we need to fit an extra random effect with kN levels to model the specific variances $\sigma^2_{\epsilon i}$. Thompson *et al.* (2003) outline an AI algorithm for this case which involves inversion of $\mathbf{C}$ in each iterate and assumes $\mathbf{A} = \mathbf{I}$.

Recently, there has been interest in the 'parameter expanded' variant of the expectation maximisation (PX-EM) algorithm. Application to mixed models involves a reparameterisation (Foulley and van Dyk 2000) which is of the same form as that

APPLICATION

Data. Traits considered were eye muscle area (EMA), intra-muscular fat content (IMF), and fat depth at the 12/13th rib (RIB) and P8 rump site (P8) for Angus cattle, measured by ultra-sound scanning of live animals, or on the carcass (C.) of slaughtered heifers or steers. In addition, carcass weight (C.WT) and percentage retail beef yield (C.RBY) were recorded. Carcass measures were collected from abattoirs under a meat quality research project by the Australian Co-operative Research Centre for Cattle and Beef Industry; see Reverter *et al.* (2000) for details. Additional records for C.WT, C.RIB and C.P8 originated from progeny testing herds. Scan traits were recorded in the field by accredited operators, for animals from 300 to 700 days of age. Records for live heifers or steers (H.) and bulls (B.) were

Table 1. Traits analysed

Trait	Unit	No.	Mean	SD
C.WT	kg	3 780	348.9	82.8
C.RBY	%	883	67.0	3.7
C.EMA	cm^2	1 847	63.4	10.3
C.P8	mm	3 385	15.34	8.57
C.RIB	mm	2 640	9.77	4.94
C.IMF	%	1 490	4.78	2.00
H.EMA	cm^2	18 170	59.1	9.1
H.P8	mm	18 362	6.34	3.15
H.RIB	mm	18 278	4.88	2.36
H.IMF	%	14 276	4.52	2.03
B.EMA	cm^2	10 409	73.6	11.9
B.P8	mm	10 315	3.79	1.81
B.RIB	mm	10 405	3.06	1.44
B.IMF	%	7 686	2.53	1.62

treated as separate traits. Scan records utilised came from the herds of origin of animals with carcass records, selecting all records in contemporary groups (CG) which included progeny of sires of animals with carcass records. After basic edits, this yielded 121 924 records on 30 427 animals. Table 1 gives details for the 14 individual traits.

Analyses. Estimates of (co)variance matrices were obtained by REML from multivariate analyses considering all 14 traits. In addition to a 'standard', full rank analysis (F14), reduced rank analyses fitting the first $m = 3, \ldots, 11$ genetic PCs only (Fm) were carried out. The residual covariance matrix was assumed to have full rank throughout. However, carcass traits were not measured for bulls, and no records for C.RBY and C.IMF for heifers or steers with scan records were included in the data. This resulted in only 63 non-zero, residual (co)variances, and a total of $p = 102$ (F3) to $p = 168$ (F14) parameters to be estimated. Analyses were carried out using an 'average information' REML algorithm (Meyer and Kirkpatrick 2005), supplemented by derivative-free and expectation-maximisation steps, as implemented in WOMBAT (Meyer 2006b). Analyses were compared considering the maximum log likelihood ($\log \mathcal{L}$) and information criteria (AIC : Akaike, BIC : Bayesian) derived from it.

Table 2. Characteristics of analyses

	p	$\log \mathcal{L}$	-½AIC	-½BIC	$\sum_i \lambda_i$
F3	102	-407.1	-357.2	-202.4	296.5
F4	113	-152.3	-113.4	-11.8	298.0
F5	123	-82.1	-53.2	0	360.2
F6	132	-30.1	-10.2	-0.5	447.0
F7	140	-13.8	-1.9	-30.9	463.8
F8	147	-4.9	0	-62.9	466.9
F9	153	-1.1	-2.1	-94.0	469.5
F10	158	-0.3	-6.3	-122.4	469.8
F11	162	-0.1	-10.2	-145.6	469.9
F14	168	0	-16.1	-180.5	468.8

Model. Analyses fitted a simple animal model. Pedigree information up to five generations backwards was included. After 'pruning', this resulted in a total of 45 928 animals in the analysis. Animals with records were progeny of 1 024 sires and 12 727 dams. Fixed effects fitted for scan traits were contemporary groups (CG), birth type (single *vs.* twin) and a dam age class (heifers *vs.* cows). CG were defined as herd-sex-management group-date of recording subclasses, with CG subdivided further if the range of ages in a subclass exceeded 60 days. Furthermore, age at recording, nested within sex, and age of dam were fitted as a linear and quadratic covariables. C.WT records were pre-adjusted to a slaughter age of 650 days, while the other carcass traits were standardised to a C.WT of 300 kg, using the multiplicative adjustments given by Reverter *et al.* (2000). The model of analysis for carcass traits then included CG, defined as herd of origin-kill regime-sex of animal subclasses (where 'kill regime' encompassed date of kill, abattoir, finishing regime and targeted market) as the only fixed effect.

Results. Values for $\log \mathcal{L}$, -½AIC and -½BIC (all given as deviation from the respective 'best' value) are summarised in Table 2, together with the sum of estimated genetic eigenvalues ($\sum_i \lambda_i$). Both a likelihood ratio test and AIC suggest that 8 PCs suffice to summarise genetic variation amongst the 14 traits considered. More conservatively, a model fitting only 5 (or 6) PCs and involving 24 (or 15) parameters less was 'best' based on BIC. However, $\sum_i \lambda_i$ did not stabilise to a consistent value until at least 8 PCs were considered, suggesting that model selection on the

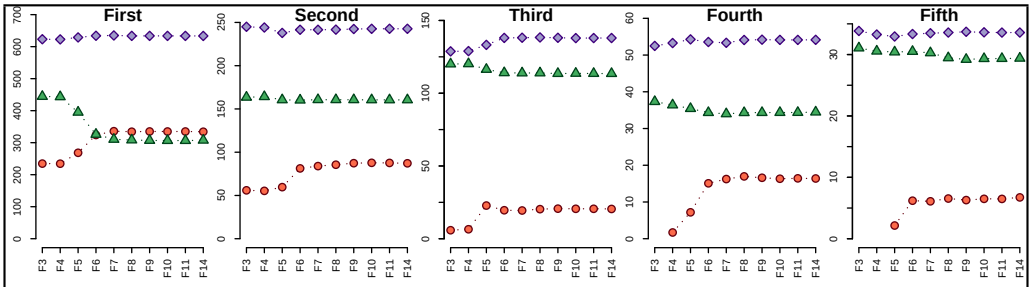

Figure 2. Estimates of the first 5 genetic (●), residual (▲) and phenotypic (♦) eigenvalues

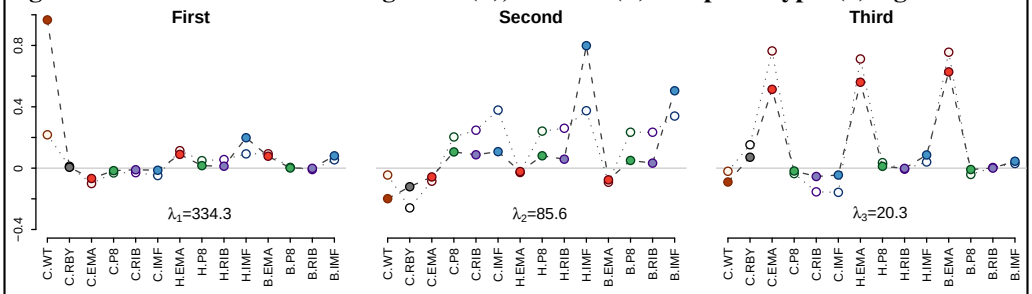

Figure 3. Loadings for first 3 principal components (● : original; ○ : standardised ×4)

basis of BIC would lead to an underestimate of total genetic variation. Estimates of the first five eigenvalues of the estimated genetic, residual and phenotypic covariance matrices for all analyses are shown in Figure 2. Fitting less than 6 PCs, clearly yields biased partitioning of variation – genetic eigenvalues are underestimated, while residual values are inflated, especially for PC_1.

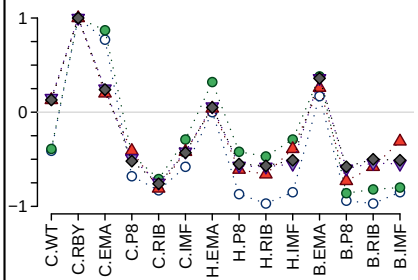

Figure 4. Genetic correlations for C. RBY (○ F3, ● F5, ▲ F7, ▼ F9, ♦ F14)

Estimated 'loadings' for individual traits for the first 3 genetic PCs from analysis F8 are shown in Figure 3. PC_1 is dominated by C.WT, the trait with the largest variance, with some positive weights for scan traits, and slight negative weights for their counterparts measured on the carcass. It explains 71.6% of genetic variation between animals for the 14 traits. In essence, PC_2, explaining 18.3% of variance, is a weighted sum of 'fatness' traits, especially when considering standardised values, i.e. loadings divided by the corresponding estimatates of genetic standard deviations. Similarly, PC_3 is essentially a weighted sum of EMA measurements. Together, PCs 1 to 3 and 1 to 5 account for 94.3% and 99.3% of variation, respectively.

Figure 4 illustrates the effect of fitting increasing numbers of PCs on estimates of genetic correlations for C.RBY. If only PC_1 were fitted, the resulting covariance matrix would have rank 1, i.e. all correlations would be 1 or −1. Fitting more PCs attenuates correlation estimates. Fitting too few PCs, then tends to yield correlation estimates biased towards an absolute value of unity (F3,

F5 to some extend). Estimates from analyses F9 and F14 are virtually undishinguishable. Similar patterns were observed for the other traits. More detailed results, including full correlation matrices, are given by Meyer (2006a).

Table 3. Accuracy (%) of genetic evaluation

Trait	1	2	3	4	5	6	7	8
C.WT	68.0	67.8	73.0	73.4	73.4	73.9	74.0	74.0
C.EMA	1.2	12.5	74.0	74.0	74.3	74.5	74.7	74.7
C.IMF	6.3	69.8	69.8	77.3	82.1	82.0	82.4	84.2
C.RBY	10.0	64.5	64.5	64.2	71.1	81.9	82.5	82.7
C.P8	3.7	55.6	58.4	64.7	74.0	76.5	79.7	80.4
C.RIB	3.8	70.8	71.1	79.5	87.2	87.9	87.8	88.8

Accuracy. In genetic evaluation, live scan records are generally only used to obtain EBVs for the carcass traits. While as many as 8 genetic PCs may be required to model the covariance structure among the 14 traits adequately, less PCs may suffice to determine EBVs without great loss in accuracy. Table 3 shows the expected accuracy of evaluation for the carcass traits for a sire with 20 male and 20 female progeny, with 4 scan records each, and 5 steer progeny with all carcass traits recorded, considering increasing numbers of PCs and assuming estimates from analysis F8 are the population values. Values clearly reflect the loadings for individual traits in each PC. For instance, the EBV for C.WT is largely determined by PC_1, while C.EMA does not have a substantial weight until PC_3. Slight decreases in accuracy when adding a PC are explicable by a loading close to zero for the trait in the new PC, and an increase in sampling variances of individual PCs as more PCs are estimated. Results suggest that for breeding schemes with main emphasis on C.WT and C.EMA as few as 3 to 4 PCs may suffice, while at least 6 PCs are required if C.IMF and C.RBY are of concern.

DISCUSSION

Genetic evaluation.

CoMputational considerations.

sampling variances.

CONCLUSIONS

Meyer (2005)

REFERENCES

Foulley, J. L. and van Dyk, D. A. (2000) *Genet. Select. Evol.* **32**:143–163.
Gilmour, A. R., Thompson, R. and Cullis, B. R. (1995) *Biometrics* **51**:1440–1450.
Harville, D. A. (1997) *Matrix Algebra from a Statistician's Perspective*. Springer Verlag.
Jennrich, R. I. and Schluchter, M. D. (1986) *Biometrics* **42**:805–820.
Jollife, I. T. (1986) *Principal Component Analysis*. Springer Series in Statistics. Springer Verlag, New York.
Kirkpatrick, M. and Meyer, K. (2004) *Genetics* **168**:2295–2306.
Meyer, K. (2005) *Anim. Sci.* **81**:337–345.
Meyer, K. (2006a) *to be decided* **00**:000–000 (in preparation).
Meyer, K. (2006b) *Proceedings Eighth World Congr. Genet. Appl. Livest. Prod.* Communication No. 27–00; in preparation.
Meyer, K. and Kirkpatrick, M. (2005) *Genet. Select. Evol.* **37**:1–30.
Reverter, A., Johnston, D. J., Graser, H.-U., Wolcott, M. L. and Upton, W. H. (2000) *J. Anim. Sci.* **78**:1786–1795.
Smith, A. B., Cullis, B. R. and Thompson, R. (2001) *Biometrics* **57**:1138–1147.
Thompson, R., Cullis, B. R., Smith, A. B. and Gilmour, A. R. (2003) *Austr. New Zeal. J. Stat.* **45**:445–459.
Wolfinger, R. D. (1996) *J. Agric. Biol. Env. Stat.* **1**:205–230.

ACKNOWLEDGEMENTS

This work was supported by Meat and Livestock Australia **(www.mla.com.au)**.