Better Estimates of Genetic Covariance Matrices by “Bending” Using Penalized Maximum Likelihood

Karin Meyer^{*,1} and Mark Kirkpatrick[†]

^{*}*Animal Genetics and Breeding Unit, University of New England, Armidale, New South Wales 2351, Australia and*

[†]*Section of Integrative Biology, University of Texas, Austin, Texas 78712*

Manuscript received December 20, 2009

Accepted for publication April 25, 2010

ABSTRACT

Obtaining accurate estimates of the genetic covariance matrix Σ_G for multivariate data is a fundamental task in quantitative genetics and important for both evolutionary biologists and plant or animal breeders. Classical methods for estimating Σ_G are well known to suffer from substantial sampling errors; importantly, its leading eigenvalues are systematically overestimated. This article proposes a framework that exploits information in the phenotypic covariance matrix Σ_P in a new way to obtain more accurate estimates of Σ_G . The approach focuses on the “canonical heritabilities” (the eigenvalues of $\Sigma_P^{-1}\Sigma_G$), which may be estimated with more precision than those of Σ_G because Σ_P is estimated more accurately. Our method uses penalized maximum likelihood and shrinkage to reduce bias in estimates of the canonical heritabilities. This in turn can be exploited to get substantial reductions in bias for estimates of the eigenvalues of Σ_G and a reduction in sampling errors for estimates of Σ_G . Simulations show that improvements are greatest when sample sizes are small and the canonical heritabilities are closely spaced. An application to data from beef cattle demonstrates the efficacy this approach and the effect on estimates of heritabilities and correlations. Penalized estimation is recommended for multivariate analyses involving more than a few traits or problems with limited data.

QUANTITATIVE geneticists, including evolutionary biologists and plant and animal breeders, are increasingly dependent on multivariate analyses of genetic variation, for example, to understand evolutionary constraints and design efficient selection programs. New challenges arise when one moves from estimating the genetic variance of a single phenotype to the multivariate setting. An important but unresolved issue is how best to deal with sampling variation and the corresponding bias in the eigenvalues of estimates for the genetic covariance matrix, Σ_G . It is well known that estimates for the largest eigenvalues of a covariance matrix are biased upward and those for the smallest eigenvalues are biased downward (LAWLEY 1956; HAYES and HILL 1981). For genetic problems, where we need to estimate at least two covariance matrices simultaneously, this tends to be exacerbated, especially for Σ_G . In turn, this can result in invalid estimates of Σ_G , *i.e.*, estimates with negative eigenvalues, and can produce systematic errors in predictions for the response to selection.

There has been longstanding interest in “regularization” of covariance matrices, in particular for cases where the ratio between the number of observations

and the number of variables is small. Various studies recently employed such techniques for the analysis of high-dimensional, genomic data. In general, this involves a compromise between additional bias and reduced sampling variation of “improved” estimators that have less statistical risk than standard methods (BICKEL and LI 2006). For instance, various types of shrinkage estimators of covariance matrices have been suggested that counteract bias in estimates of eigenvalues by shrinking all sample eigenvalues toward their mean. Often this is equivalent to a weighted combination of the sample covariance matrix and a target matrix, assumed to have a simple structure. A common choice for the latter is an identity matrix. This yields a ridge regression type formulation (HOERL and KENNARD 1970). Numerous simulation studies in a variety of settings are available, which demonstrate that regularization can yield closer agreement between estimated and population covariance matrices, less variable estimates of model terms, or improved performance of statistical tests.

In quantitative genetic analyses, we attempt to partition observed, overall (phenotypic) covariances into their genetic and environmental components. Typically, this results in strong sampling correlations between them. Hence, while the partitioning into sources of variation and estimates of individual covariance matrices may be subject to substantial sampling variances, their sum, *i.e.*, the phenotypic covariance matrix, can generally be estimated much more accurately. This has

Supporting information is available online at <http://www.genetics.org/cgi/content/full/genetics.109.113381/DC1>.

¹*Corresponding author:* Animal Genetics and Breeding Unit, University of New England, Armidale, New South Wales 2351, Australia.
E-mail: kmeyer.agbu@gmail.com

led to suggestions to “borrow strength” from estimates of phenotypic components to estimate the genetic covariances. In particular, HAYES and HILL (1981) proposed a method termed “bending” that involved regressing the eigenvalues of the product of the genetic and the inverse of the phenotypic covariance matrix toward their mean. One objective of this procedure was to ensure that estimates of the genetic covariance matrix from an analysis of variance were positive definite. In addition, the authors showed by simulation that shrinking eigenvalues even further than needed to make all values nonnegative could improve the achieved response to selection when using the resulting estimates to derive weights for a selection index, especially for estimation based on small samples. Subsequent work demonstrated that bending could also be advantageous in more general scenarios such as indexes that included information from relatives (MEYER and HILL 1983).

Modern, mixed model (“animal model”)-based analyses to estimate genetic parameters using maximum likelihood or Bayesian methods generally constrain estimates to the parameter space, so that—at the expense of introducing some bias—estimates of covariance matrices are positive semidefinite. However, the problems arising from substantial sampling variation in multivariate analyses remain. In spite of increasing applications of such analyses in scenarios where data sets are invariably small, *e.g.*, the analysis of data from natural populations (*e.g.*, KRUEK *et al.* 2008), there has been little interest in regularization and shrinkage techniques in genetic parameter estimation, other than through the use of informative priors in a Bayesian context. Instead, suggestions for improved estimation have focused on parsimonious modeling of covariance matrices, *e.g.*, through reduced rank estimation or by imposing a known structure, such as a factor-analytic structure (KIRKPATRICK and MEYER 2004; MEYER 2009), or by fitting covariance functions for longitudinal data (KIRKPATRICK *et al.* 1990). While such methods can be highly advantageous when the underlying assumptions are at least approximately correct, data-driven methods of regularization may be preferable in other scenarios.

This article explores the scope for improved estimation of genetic covariance matrices by implementing the equivalent to bending within animal model-type analyses. We begin with a review of the underlying statistical principles (which the impatient reader might skip), examining the concept of improved estimation, its implementation via shrinkage estimators or penalized estimation, and selected applications. We then describe a penalized restricted maximum-likelihood (REML) procedure for the estimation of genetic covariance matrices that utilizes information from its phenotypic counterparts and present a simulation study demonstrating the effect of penalties on parameter estimates and their sampling properties. The article concludes

with an application to a problem relevant in genetic improvement of beef cattle and a discussion.

REVIEW: PRINCIPLES OF PENALIZED ESTIMATION

In broad terms, regularization in statistics refers to a scenario where estimation for ill-posed or overparameterized problems is improved through use of some form of additional information. Often, the latter is composed of a penalty for a deviation from a desired outcome. For example, in fitting smoothing splines a “roughness penalty” is commonly employed to place preference on simple functions (GREEN 1998). This section reviews some of the underlying principles of improved estimation of covariance matrices.

Minimizing statistical risk: A central term in estimation is that of risk, defined as expected loss, arising from the inevitable deviation of estimates from the underlying population values. Consider a set of q normally distributed variables with population covariance matrix Σ , recorded on n individuals, and estimator $\hat{\Sigma}$. Common loss functions considered in the statistical literature are the entropy (L_1) and quadratic (L_2) loss (JAMES and STEIN 1961)

$$L_1(\Sigma, \hat{\Sigma}) = \text{tr}(\Sigma^{-1}\hat{\Sigma}) - \log |\Sigma^{-1}\hat{\Sigma}| - q \quad (1)$$

and

$$L_2(\Sigma, \hat{\Sigma}) = \text{tr}(\Sigma^{-1}\hat{\Sigma} - \mathbf{I})^2 = \text{tr}(\Sigma^{-1}(\hat{\Sigma} - \Sigma))^2. \quad (2)$$

A natural estimator for Σ is a scalar multiple of the matrix of sums of squares and cross-products among the q variables, $\mathbf{S}$. In this class of estimators, the sample covariance matrix $\mathbf{S}/d$ with d the degrees of freedom, *i.e.*, the usual, unbiased estimator, minimizes the L_1 risk, while $\mathbf{S}/(d + q + 1)$ yields the minimum risk estimator under loss function L_2 (*e.g.*, HAFF 1980).

For Σ known, these loss functions provide a measure of how different the estimate $\hat{\Sigma}$ is from the true value, where the difference can reflect both chance (*i.e.*, sampling error) and systematic deviations (*i.e.*, bias). While estimates with minimum bias may be preferred in some circumstances, in general estimators with a low overall loss or risk are considered desirable. The “statistical” risk defined in such a general manner may not reflect the importance of errors in specific elements of $\hat{\Sigma}$ adequately in some situations, and alternative definitions may be more appropriate. These are, in principle, readily substituted where required, although they may not necessarily translate to simple penalties on the likelihood function as for the L_1 and L_2 loss. In a genetic context, an important use of estimated genetic and phenotypic covariance matrices is in the calculation of the weights for individual traits in selection indexes. The achieved response to selection on the index then depends on how closely the estimates—and thus the index weights derived from them—adhere to the true

values (HAYES and HILL 1981). While reducing the general risk, as quantified by the loss functions above (Equations 1 and 2), is likely to reduce the error in index weights and thus the discrepancy between expected and achieved response, this may not be optimal. For instance, if economic weights for individual traits in an index differ substantially, the definition of risk may need to be adapted accordingly.

Improved estimators of covariance matrices: There is a considerable body of literature on improved estimators of covariance matrices. These are generally biased, but have a lower risk than the standard, unbiased estimator (sample covariance matrix). Several studies derived the risk for a certain class of estimator and given loss function and presented estimators that “dominate” over other estimators followed by a simulation study to demonstrate their properties. Others obtained estimators using a different motivation, such as minimax or empirical Bayesian estimation; see KUBOKAWA (1999) and HOFFMANN (2000) for reviews.

Sampling variation causes the largest eigenvalues of a covariance matrix to be overestimated and the smallest eigenvalues to be underestimated, while their mean is expected to be unbiased. Hence, attention has focused on estimators that modify the eigenvalues of the sample covariance matrix while retaining the corresponding eigenvectors. The impetus for this is generally attributed to STEIN (1975), but similar suggestions can be found earlier, for instance, in LAWLEY (1956). Let $\hat{\omega}_i$ denote the i th eigenvalue of the sample covariance matrix. Stein’s proposal then consisted of an adaptive shrinking obtained by scaling each $\hat{\omega}_i$ by $d/(d - q + 1 + 2\hat{\omega}_i \sum_{j \neq i} (\hat{\omega}_i - \hat{\omega}_j)^{-1})$. The resulting estimator minimizes the entropy loss but does not preserve the order of eigenvalues or ensure nonnegativity. Hence, later work often combined this with order-preserving measures such as an “isotonizing” regression (which restores order by merging values out of line) or truncation at zero (DEY and SRINIVASAN 1985; LIN and PERLMAN 1985; YE and WANG 2009).

A simple modification scheme entails the linear shrinkage of the sample eigenvalues toward their mean. It can be shown that this yields an estimator that is a weighted combination of the sample covariance matrix and an identity matrix. Considering a quadratic loss function, LEDOIT and WOLF (2004) derived an optimal shrinkage factor $\rho \in [0, 1]$ that minimized the risk associated with the estimator $\rho\bar{\omega}\mathbf{I} + (1 - \rho)\mathbf{S}/d$ (with $\mathbf{I}$ an identity matrix and $\bar{\omega}$ the mean sample eigenvalue). DANIELS and KASS (2001) argued that, due to the nature of the quadratic loss, such an estimator could result in overshrinkage, in particular of the smallest eigenvalues, when the true eigenvalues were far apart. Instead the authors proposed an estimator derived by assuming a prior normal distribution for the eigenvalues on the logarithmic scale, approximated as $\log(\hat{\omega}_i) \propto N(\log(\omega_i), 2/n)$. This resulted in modified values of the form $\hat{\omega}_i =$

$\exp(\rho \log(\bar{\omega}) + (1 - \rho)\log(\hat{\omega}_i))$ [with $\log(\bar{\omega})$ the mean of $\log(\hat{\omega}_i)$], *i.e.*, again involved a regression toward the sample mean, but on a different scale. WARTON (2008) proposed a similar, regularized estimator of the sample correlation matrix $\hat{\mathbf{R}}, \rho\hat{\mathbf{R}} + (1 - \rho)\mathbf{I}$, and showed that this was the penalized maximum-likelihood estimator with penalty term proportional to $-\text{tr}(\mathbf{R}^{-1})$, with the corollary that the corresponding, ridge type estimator of a covariance matrix $\hat{\Sigma}, \hat{\Sigma} + \kappa\mathbf{I}$, involved a penalty proportional to $-\text{tr}(\hat{\Sigma}^{-1})$.

Other work considered shrinkage toward a more general structure. SCHÄFER and STRIMMER (2005) and SANCETTA (2008) extended the approach of LEDOIT and WOLF (2004) to different target matrices of simple structure with few parameters to be estimated, *e.g.*, a diagonal matrix with different variances or a matrix with all correlations equal. BÖHM (2008) examined estimators for multivariate time series and suggested data-driven shrinkage toward a factor-analytic structure. Shrinkage estimators of correlation matrices have been described by LIN and PERLMAN (1985), DANIELS and KASS (2001), and WARTON (2008).

More than one matrix: Few studies have addressed improved estimation for multilevel models such as appear in quantitative genetics. The simplest case, with two matrices to be estimated, is a balanced one-way classification. Let Σ_B and Σ_W denote the covariance matrices between and within groups, respectively, and $\mathbf{B}$ and $\mathbf{W}$ denote the corresponding matrices of mean squares and cross-products (MSCP). Derivations of improved estimators by and large utilized the so-called canonical decomposition of $\mathbf{B}$ and $\mathbf{W}$: For any two symmetric (real) matrices, $\mathbf{W}$ and $\mathbf{B}$, of size $q \times q$ with $\mathbf{W}$ positive-definite and $\mathbf{B}$ positive semidefinite (p.s.d.), a matrix $\mathbf{T}$ exists such that $\mathbf{T}\mathbf{T}' = \mathbf{W}$ and $\mathbf{T}\mathbf{A}\mathbf{T}' = \mathbf{B}$, with $\mathbf{A} = \text{Diag}\{\lambda_i\}$ the diagonal matrix of eigenvalues of $\mathbf{W}^{-1}\mathbf{B}$ (ANDERSON 1984).

An immediate, additional problem then is to ensure that estimates are within the parameter space, *i.e.*, are not negative definite. Due to sampling variation, the usual unbiased estimator for the between-group component, $\hat{\Sigma}_B = (\mathbf{B} - \mathbf{W})/m$ (with m the group size), has a high probability, increasing with q and decreasing sample size, of not being p.s.d., *i.e.*, to have negative eigenvalues (HILL and THOMPSON 1978; BHARGAVA and DISCH 1982). Using the canonical transformation yields $\hat{\Sigma}_B = (\mathbf{T}(\mathbf{A} - \mathbf{I})\mathbf{T}')/m$, and it is readily seen that $\hat{\Sigma}_B$ is guaranteed to be nonnegative definite by truncating the elements of $\mathbf{A}$ at a minimum of unity, *i.e.*, by replacing $\mathbf{A}$ with $\mathbf{A}_T^* = \text{Diag}\{\min(1, \lambda_i)\}$. The resulting estimator is the REML estimator (KLOTZ and PUTTER 1969; AMEMIYA 1985; ANDERSON *et al.* 1986). In a genetic context where we estimate the matrix of environmental covariances as $\hat{\Sigma}_E = \hat{\Sigma}_W - (\alpha - 1)\hat{\Sigma}_B$ (with α^{-1} the degree of relationship among group members), additional constraints may be required to ensure that $\hat{\Sigma}_E$ is within the parameter space (MEYER and KIRKPATRICK 2008).

As outlined above, HAYES and HILL (1981) suggested to bend the estimate of the genetic covariance matrix, $\hat{\Sigma}_G = \alpha \hat{\Sigma}_B$, toward the estimate of the phenotypic covariance matrix, $\hat{\Sigma}_P = \hat{\Sigma}_B + \hat{\Sigma}_W$, by regressing the eigenvalues of $\hat{\Sigma}_P$ to $\hat{\Sigma}_G$ to their mean. Their rationale for this was somewhat *ad hoc*: $\hat{\Sigma}_P$ plays a central role in computing the weights in a selection index and the main objective was to improve the properties of selection indexes based on the estimated covariance matrices. Rather than manipulating the roots of $\hat{\Sigma}_P$ directly though, HAYES and HILL (1981) modified $\mathbf{W}^{-1}\mathbf{B}$, using that for λ_i a root of $\mathbf{W}^{-1}\mathbf{B}$, $\alpha(\lambda_i - 1)/(\lambda_i - 1 + n)$ is a root of $\hat{\Sigma}_P$. Their estimator for Σ_B was then obtained by replacing Λ above by $\Lambda_B^* = \text{Diag}\{\rho\bar{\lambda} + (1 - \rho)\lambda_i\}$, with $\bar{\lambda}$ the mean of the λ_i and $\rho \in [0, 1]$ the bending factor.

LOH (1991), MATHEW *et al.* (1994), SRIVASTAVA and KUBOKAWA (1999), and KUBOKAWA and TSAI (2006) considered estimation for two independent Wishart matrices, such as $\mathbf{B}$ and $\mathbf{W}$ in the one-way classification. Minimizing the sum of entropy losses, they derived different types of joint estimators, analogous to those proposed for a single matrix, and showed that improved estimators were available that had lower risk than the unbiased or REML estimators. However, no practical applications using any of these results are available. Again, these estimators involved some form of modification of the eigenvalues arising from the canonical decomposition of the two matrices, indicating that the suggestion of HAYES and HILL (1981) was well founded.

Penalized maximum-likelihood estimation: A standard method of regularization is that of penalized estimation, in particular for regression problems. In a least-squares or maximum-likelihood (ML) context, this involves a penalty that is added to the criterion to be minimized. The effect of the penalty is modulated by a so-called tuning parameter (ψ) that determines the relative importance of information from the data and the desired outcome. For (RE)ML, this replaces the objective function $\log \mathcal{L}(\boldsymbol{\theta})$ with

$$\log \mathcal{L}_P(\boldsymbol{\theta}) = \log \mathcal{L}(\boldsymbol{\theta}) - \frac{1}{2} \psi \mathcal{P}(\boldsymbol{\theta}),$$

where $\boldsymbol{\theta}$ is the vector of parameters to be estimated, $\log \mathcal{L}$ denotes the standard log likelihood, and $\mathcal{P}$ is a (nonnegative) penalty function (the factor $\frac{1}{2}$ is for algebraic consistency and could be omitted).

Let $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$ denote a simple regression model with $\mathbf{y}$, $\mathbf{b}$, and $\mathbf{e}$ the vectors of observations, regression coefficients, and residuals, respectively, and $\mathbf{X}$ the corresponding incidence matrix. A class of penalties commonly employed is that of the ℓ_p norm

$$\ell_p(\mathbf{b}) = \|\mathbf{b}\|_p = \sum_i |b_i|^p, \quad (3)$$

with b_i the i th element of $\mathbf{b}$. Different values of p are appropriate for different analyses. For $p = 0$, the penalty

is equal to the number of elements in $\mathbf{b}$ and may be employed in model selection. A value of $p = 1$ yields a least absolute shrinkage and selection operator (LASSO)-type penalty that encourages shrinkage of small (absolute value) elements of $\mathbf{b}$ to zero and thus subset selection (TIBSHIRANI 1996). A quadratic penalty ($p = 2$) produces a ridge regression formulation where all elements of $\mathbf{b}$ are shrunk proportionally (HOERL and KENNARD 1970). Noninteger values for p have been suggested, *e.g.*, the so-called “bridge,” attributed to FRANK and FRIEDMAN (1993), for values of $0 < p < 1$. Other proposals have been to combine ℓ_1 - and ℓ_2 -type penalties to form the “elastic net” (ZOU and HASTIE 2005) or to account for the correlation among predictors in the penalty term (TUTZ and ULBRICHT 2009).

In a more general framework, we may assume a certain prior distribution for the parameters to be estimated and impose a penalty that is proportional to minus the logarithmic value of the prior density. This provides a direct link to Bayesian estimation—indeed, penalized maximum-likelihood estimation has been described as “an attempt of enjoying the Bayesian fruits without paying the B-club fee” (MENG 2008, p. 1545). For instance, imposing an ℓ_1 -type penalty is equivalent to assuming a double exponential prior distribution, while an ℓ_2 penalty implies a normal prior.

Estimation of the tuning factor: A general procedure to estimate the tuning parameter ψ from the data at hand is cross-validation. This involves splitting the data into so-called training and validation sets. We then fit our model and estimate the parameters of interest for a range of values of ψ , using the training set only. For each set of estimates, the value of the (unpenalized) objective function, *e.g.*, the likelihood or the residual sum of squares, is determined using the validation set. The value of ψ that optimizes this criterion is then chosen as the best value to use for a penalized analysis of the complete data set.

In practice, multiple splits are used. A popular scheme is that of K -fold cross-validation (*e.g.*, HASTIE *et al.* 2001, Chap. 7) where the data are evenly split into K subsets, and K analyses are carried out for each value of ψ , with the i th subset treated as the validation set and the remaining $K - 1$ subsets forming the training set. The tuning parameter is then chosen on the basis of the objective function averaged across the K validation sets. Common values of K used are 5 and 10. A related technique is repeated random subsampling. For example, BICKEL and LEVINA (2008) employed a scheme using 50 “random splits” of the data, with the training set comprising a third of the data, and showed that while, due to noise, this tended to overestimate the true risk substantially, the location of the minimum tended to be identified correctly.

In a simulation setting, an alternative is to simply simulate additional data sets for validation, with the same data structure and sampling from distributions

with the same population parameters as for the “training” sets. Using one set per replicate, ROTHMAN *et al.* (2009) noted that use of such validation sets could be replaced with cross-validation without any significant changes in results (but did not give details on the cross-validation scheme referred to).

In special cases, ψ can be estimated directly. An example is that of function estimation (smoothing) using a semiparametric regression or penalized splines with a quadratic penalty. In that case, the regression can be rewritten as a mixed model with the penalized coefficients treated as random effects and the tuning parameter can be estimated “automatically” in a (RE)ML analysis from the ratio of the residual variance and the variance due to random effects; see RUPPERT *et al.* (2003, Chap. 5) for an example. FOSTER *et al.* (2009) showed that a LASSO-type penalty on effects can be imposed in a mixed model by treating these as random effects with a double-exponential distribution and that the respective variance parameters (which determine the amount of penalization) can be estimated using a ML approach.

Applications to covariance matrices: Penalized ML estimation of covariance matrices has predominantly been applied in the spirit of covariance selection (DEMPSTER 1972), *i.e.*, to encourage sparsity in the estimate of the inverse of the covariance matrix (“concentration” matrix), and mostly for problems with high dimensions and some natural ordering of variables, *e.g.*, longitudinal data. This relied on the modified Cholesky decomposition $\Sigma = \mathbf{L}\mathbf{D}\mathbf{L}'$ or $\Sigma^{-1} = \mathbf{L}'\mathbf{D}^{-1}\mathbf{L}$, with $\mathbf{L}$ a lower triangular matrix with diagonal elements of unity and $\mathbf{D}$ a diagonal matrix. For longitudinal data, the nonzero off-diagonal elements of $\mathbf{L}$ have an interpretation as (minus) the regression coefficients in an autoregressive model (POURAHMADI 1999) and can be penalized in the same fashion as coefficients in a simple regression context. Applications using both ℓ_1 - and ℓ_2 -type penalties can be found in HUANG *et al.* (2006), BICKEL and LEVINA (2008), FRIEDMAN *et al.* (2008), LEVINA *et al.* (2008), ROTHMAN *et al.* (2008), and YAP *et al.* (2009). With a different objective, WARTON (2008) considered penalized ML estimation of covariance and correlation matrices to obtain regularized estimates with a stable inverse for use in multivariate regression problems.

PENALIZED REML ESTIMATION

Consider a simple animal model for q traits, $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{g} + \mathbf{e}$ with $\mathbf{y}$, $\mathbf{b}$, $\mathbf{g}$, and $\mathbf{e}$ the vectors of observations, fixed effects, and additive genetic and residual effects, respectively, and $\mathbf{X}$ and $\mathbf{Z}$ the corresponding incidence matrices. Let Σ_G and Σ_E denote the matrices of additive genetic and residual covariances among the q traits, and let $\text{Var}(\mathbf{g}) = \Sigma_G \otimes \mathbf{A} = \mathbf{G}$ with $\mathbf{A}$ the numerator relationship matrix between individuals and $\text{Var}(\mathbf{e}) = \sum_k^+ \mathbf{R}_k = \mathbf{R}$, with $\mathbf{R}_k$ the submatrix of Σ_E corresponding to the traits recorded for the k th individual and $\sum^+$ is the direct

matrix sum. This gives $\text{Var}(\mathbf{y}) = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R} = \mathbf{V}$ and the REML log likelihood is, apart from a constant,

$$\log \mathcal{L}(\theta) = -\frac{1}{2}(\log |\mathbf{V}| + \log |\mathbf{X}'_0 \mathbf{V}^{-1} \mathbf{X}_0| + (\mathbf{y} - \mathbf{X}\mathbf{b})' \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\mathbf{b})) \quad (4)$$

with $\mathbf{X}_0$ a full-rank submatrix of $\mathbf{X}$ (*e.g.*, HARVILLE 1977).

Assuming Σ_G and Σ_E are unstructured, we have $q(q+1)$ variance parameters to be estimated. Maximization of Equation 4 with respect to the elements of Σ_G and Σ_E represents a constrained optimization problem, as estimates need to be in the parameter space, *i.e.*, cannot have negative eigenvalues. Hence implementations of REML estimation often employ a parameterization to a scale that does not require constraints (PINHEIRO and BATES 1996), *e.g.*, estimating the elements of the Cholesky factor of a covariance matrix rather than the covariance components directly.

A natural alternative in our context is a parameterization to the elements of the canonical decomposition of Σ_G and $\Sigma_P = \Sigma_G + \Sigma_E$: Let $\mathbf{Q}\Sigma_G\mathbf{Q}' = \Lambda$ and $\mathbf{Q}\Sigma_P\mathbf{Q}' = \mathbf{I}$ so that for $\mathbf{T} = \mathbf{Q}^{-1}$, $\mathbf{T}\mathbf{A}\mathbf{T}' = \Sigma_G$ and $\mathbf{T}\mathbf{T}' = \Sigma_P$. The elements of Λ are the eigenvalues of $\Sigma_P^{-1/2}\Sigma_G\Sigma_P^{-1/2}$ and $\mathbf{T}$ is the corresponding matrix of eigenvectors premultiplied by the matrix square root of Σ_P . This yields q parameters λ_i and q^2 elements (t_{ij}) of $\mathbf{T}$ to be estimated. Eigenvalues λ_i can be interpreted as heritabilities on the canonical scale and are thus constrained to the interval $[0, 1]$. Again, these constraints can be removed through a suitable further reparameterization, *e.g.*, by estimating $\log(-\log \lambda_i)$ instead of λ_i .

Our review above has identified that modification of the canonical eigenvalues of the “between” and “within” matrices of MSCP in a one-way classification can provide improved estimators of the corresponding covariance matrices and that manipulation of these eigenvalues is equivalent to modifying the canonical eigenvalues of Σ_G and Σ_P . Furthermore, we have shown that regressing the eigenvalues of a matrix toward their mean yields a shrinkage estimator that is a weighted combination of the matrix and a multiple of the identity matrix and that such estimators can be obtained by penalized maximum likelihood with a quadratic, ℓ_2 -type (*cf.* Equation 3) penalty. Using these findings, we propose to implement the equivalent to bending within our REML animal model analysis by replacing $\log \mathcal{L}$ in Equation 4 by

$$\log \mathcal{L}_P(\theta) = \log \mathcal{L}(\theta) - \frac{1}{2}\psi \mathcal{P}(\theta) \quad \text{with} \quad \mathcal{P}(\theta) = \sum_{i=1}^q (\lambda_i - \bar{\lambda})^2 \quad (5)$$

for $\bar{\lambda} = (\sum_{i=1}^q \lambda_i)/q$. This directly mimics the suggestion of HAYES and HILL (1981) as the quadratic penalty provides a linear shrinkage of all λ_i toward their mean. For $\psi = 0$, $\log \mathcal{L}_P$ reduces to $\log \mathcal{L}$, and for $\psi \rightarrow \infty$ Equation 5 yields a model in which all λ_i are constrained to be equal.

Equation 5 is readily extended to other types of penalties. For instance, rather than shrinking toward

the arithmetic mean $\bar{\lambda}$, we could use the geometric or harmonic mean (YE and WANG 2009). An analog of the log-posterior shrinkage estimator of DANIELS and KASS (2001) could be obtained by transforming eigenvalues λ_i to logarithmic scale. More generally, the shrinkage could be modified by applying a Box-Cox transformation, *i.e.*, by replacing λ_i with $(\lambda_i^\gamma - 1)/\gamma$ for $\gamma > 0$ or $\log \lambda_i$ for $\gamma = 0$. Alternatively, we might consider replacing the exponent of $p = 2$ with a general value or using a combination of penalties.

For the parameterization to the elements of the canonical transformation, derivatives of $\mathcal{P}(\boldsymbol{\theta})$ are straightforward, and standard REML algorithms, such as expectation-maximization or the so-called average information algorithm (see THOMPSON *et al.* 2005, for a recent review and references), are readily adapted. Derivatives required are summarized in the APPENDIX. One drawback of this parameterization, however, is that (first) derivatives of both $\boldsymbol{\Sigma}_G$ and $\boldsymbol{\Sigma}_E$ with respect to all $q(q + 1)$ parameters to be estimated are nonzero. This implies that computational requirements per iterate are increased compared to the usual implementations. However, this increase is relatively small and pales to insignificance compared to extra computations required when the tuning parameter is to be estimated using cross-validation.

SIMULATION

Method: A simulation study was carried out considering $q = 5$ traits, 11 sets of genetic parameters, and two types of penalties. Without loss of generality, population values chosen were different combinations of canonical heritabilities (λ_i). As discussed by HILL and THOMPSON (1978), these can represent a wide range of constellations of heritabilities and genetic and environmental correlations. Population values selected differed in both the average level of heritability ($\bar{\lambda}$) and the spread of the λ_i about their mean. Values for scenarios A–K are summarized in Table 1. To understand the effects of the pedigree structure, two contrasting designs were examined. Simulation I comprised a classic, balanced paternal half-sib design with 500, 200, or 100 sires with 10 progeny each. Simulation II considered 125 or 50 unrelated families, using the design of BONDARI *et al.* (1978): Each family involved records on two pairs of full-sibs in generation 1, with one male and one female per pair. In generation 2, two paternal half-sibs of different sex were mated to unrelated individuals, recording two offspring per mating. This yielded records for 8 individuals per family, which provided nine different types of covariances between relatives; see THOMPSON (1976) for a mating plan and list of covariances.

Matrices of MSCP ($\mathbf{B}$ and $\mathbf{W}$ for simulation I and the 40×40 matrix of MSCP pertaining to records for a family, 5×8 , in simulation II) for each design and set of

population values were sampled from central Wishart distributions as described by ODELL and FEIVESON (1966). Estimates of genetic and environmental covariance components were obtained using the canonical parameterization as described above and a simple derivative-free optimization procedure to locate the maximum of the (penalized) likelihood function, $\log \mathcal{L}_P$. Estimation constrained values of λ_i to the range of 0.00001–0.99999. Quadratic penalties on the deviation of the canonical heritabilities from their mean were imposed either on the eigenvalues directly or on values transformed to logarithmic scale. A range of values for ψ (0–1.8 in steps of 0.2, 2–4.5 in steps of 0.5, 5–99 in steps of 1, 100–248 in steps of 2, 250–495 in steps of 5, and 500–1000 in steps of 10; 289 in total) were considered.

For simplicity, the appropriate tuning parameter for each replicate was estimated by sampling additional validation data, as done, for instance, by LEVINA *et al.* (2008) and ROTHMAN *et al.* (2008, 2009), rather than by cross-validation. This involved sampling 100 sets of matrices of MSCP for each replicate, for the same sample size and population parameters as the training data. Let $\hat{\boldsymbol{\theta}}_r$ denote the vector of parameter estimates for a replicate and the r th value of the tuning parameter, ψ_r . For each vector of estimates $\hat{\boldsymbol{\theta}}_r$ ($r = 1, 289$) and each of the additional data sets, $\mathbf{y}_l$ ($l = 1, 100$), the value of the unpenalized likelihood, $\log \mathcal{L}(\hat{\boldsymbol{\theta}}_r, \mathbf{y}_l)$, was calculated. The estimate of the tuning parameter to be used for this replicate, $\hat{\psi}$, was then chosen as the value of ψ_r for which the average likelihood across validation sets,

$$\frac{\sum_{l=1}^{100} \log \mathcal{L}(\hat{\boldsymbol{\theta}}_r, \mathbf{y}_l)}{100},$$

was maximized. A total of 10,000 replicates were carried out for each scenario examined.

As suggested by LIN and PERLMAN (1985), the effect of penalized estimation was summarized as percentage reduction in average loss (PRIAL), calculated as

$$\frac{100[\bar{L}_1(\boldsymbol{\Sigma}_X, \hat{\boldsymbol{\Sigma}}_X^0) - \bar{L}_1(\boldsymbol{\Sigma}_X, \hat{\boldsymbol{\Sigma}}_X^\psi)]}{\bar{L}_1(\boldsymbol{\Sigma}_X, \hat{\boldsymbol{\Sigma}}_X^0)}$$

with $\hat{\boldsymbol{\Sigma}}_X^0$ the standard, unpenalized REML estimate and $\hat{\boldsymbol{\Sigma}}_X^\psi$ the penalized estimate, for $X = G, E$, and P and $\bar{L}_1(\cdot)$ the entropy loss as defined in Equation 1 averaged over replicates.

Results: The effect of sampling variation and penalization on estimates of canonical heritabilities is illustrated in Figures 1 and 2. As expected from theory (*e.g.*, LAWLEY 1956), bias in unpenalized estimates increased markedly with decreasing spread in the population values and decreasing sample size. Patterns for both designs were similar. For scenarios with equal population

TABLE 1

Reduction in average loss (PRIAL, in percent), for estimates of the genetic (Σ_G), error (Σ_E), and phenotypic (Σ_P) covariance matrix together with mean entropy loss ($\times 100$) in unpenalized REML estimates of Σ_G ($\bar{L}_1(\hat{\Sigma}_G, \Sigma_G)$), the proportion of replicates (W , in percent) for which penalized estimation increased the loss in Σ_G , and the mean tuning factor (ψ), for different constellations ($A, \dots, K$) of population values (λ_b , canonical heritabilities) and penalties on the original (ORG) or logarithmic (LOG) scale

	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>	<i>G</i>	<i>H</i>	<i>I</i>	<i>J</i>	<i>K</i>
λ_1	0.40	0.50	0.60	0.70	0.80	0.90	0.20	0.30	0.60	0.50	0.90
λ_2	0.40	0.45	0.50	0.55	0.30	0.50	0.20	0.25	0.10	0.20	0.30
λ_3	0.40	0.40	0.40	0.40	0.30	0.30	0.20	0.20	0.10	0.15	0.10
λ_4	0.40	0.35	0.30	0.25	0.30	0.20	0.20	0.15	0.10	0.10	0.10
λ_5	0.40	0.30	0.20	0.10	0.30	0.10	0.20	0.10	0.10	0.05	0.10
500 sires											
$\bar{L}_1(\hat{\Sigma}_G^0, \Sigma_G)$	13	12	14	24	14	26	29	42	114	165	69
ORG											
Σ_G	79	52	26	27	17	15	86	53	25	38	14
Σ_E	83	48	25	24	24	68	72	35	15	12	45
Σ_P	8	5	2	2	0	0	2	1	1	0	0
W	0	0	11	27	2	18	0	4	2	10	8
ψ	984.7	307.0	68.1	30.4	29.0	12.4	991.6	293.6	23.5	43.2	7.8
LOG											
Σ_G	91	51	24	28	39	34	93	50	74	74	63
Σ_E	86	49	24	12	20	37	77	34	18	16	29
Σ_P	8	5	2	1	1	0	2	1	0	0	0
W	0	1	20	29	1	23	0	13	1	20	2
ψ	769.6	47.1	8.4	1.9	10.2	1.9	369.0	10.4	2.7	2.3	1.7
200 sires											
$\bar{L}_1(\hat{\Sigma}_G^0, \Sigma_G)$	31	32	39	109	37	110	114	217	531	497	357
ORG											
Σ_G	91	72	48	59	40	34	94	83	32	49	17
Σ_E	86	67	43	40	41	68	76	51	21	16	42
Σ_P	8	6	4	2	0	-1	2	2	0	0	-1
W	0	0	3	22	0	11	0	1	1	4	3
ψ	921.3	385.8	79.0	32.9	40.6	13.0	962.8	363.4	22.7	49.2	5.3
LOG											
Σ_G	91	71	46	59	56	67	95	83	89	83	85
Σ_E	87	68	44	31	37	38	77	52	27	27	36
Σ_P	8	6	4	2	1	0	2	2	1	1	0
W	0	0	9	35	0	19	0	3	0	10	2
ψ	388.2	65.2	10.2	2.7	7.6	2.0	169.2	15.2	1.9	2.3	1.3
100 sires											
$\bar{L}_1(\hat{\Sigma}_G^0, \Sigma_G)$	72	80	123	295	108	317	466	590	930	829	702
ORG											
Σ_G	93	84	75	73	70	47	96	88	35	57	13
Σ_E	88	79	64	64	60	61	76	61	31	29	37
Σ_P	8	7	5	3	1	-1	2	2	1	1	-1
W	0	0	1	13	0	6	0	0	2	3	4
ψ	760.9	445.3	112.4	38.3	58.5	16.0	854.4	422.9	25.6	59.7	4.5
LOG											
Σ_G	93	84	73	74	76	79	98	92	90	84	87
Σ_E	88	79	66	58	52	43	78	63	38	38	32
Σ_P	8	7	5	3	2	0	3	2	1	1	-1
W	0	0	3	28	0	12	0	1	0	4	1
ψ	194.4	72.5	13.0	3.8	6.7	2.1	92.4	23.0	1.7	2.4	1.2

Balanced paternal half-sib design is shown with 500, 200, or 100 sires and 10 progeny per sire.

values (A and G), penalization dramatically reduced the bias in estimates with the small remaining bias in the same direction as for unpenalized estimation. For the

other cases penalization appeared to overcompensate somewhat, resulting in a bias in the opposite direction for the extreme values (λ_1 and λ_5), *i.e.*, yielded estimates

FIGURE 1.—Mean estimates of canonical heritabilities, as deviation from population values (in percent) together with plus or minus one empirical standard deviation (vertical bars) for standard REML analyses ($\square$), and penalties on eigenvalues on the original ($\circ$) and logarithmic (Δ) scale, for cases A, B, and C and a paternal half-sib design with 500 or 200 sires.

of the largest values that were biased downward and estimates of the smallest values that were biased upward.

Imposing a penalty on the logarithmic scale tended to give estimates of λ_1 that were less biased than for penalization on the original scale but, in turn, yielded larger upward bias in estimates of λ_5 in most cases. Overshrinkage, in particular of the smallest eigenvalues, when population values are far apart has been observed previously and has been attributed to the nature of the quadratic penalty used (DANIELS and KASS 2001). While the upward bias due to penalization in the estimate of the smallest eigenvalue for case H (and similarly cases D and F) may look disconcertingly large (see Figure 2), it should borne in mind that the corresponding population value was 0.1 so that, in absolute terms, an upward bias of 30–60% was still relatively small and, as discussed below, penalized estimation resulted in a substantial reduction in loss in $\hat{\Sigma}_G$.

Penalization has the potential to reduce sampling variation for variables on the original scale dramatically. Furthermore, it has to be emphasized that the systematic bias in estimates of the heritabilities on the canonical scale does not automatically translate into a corresponding bias in estimates of the usual genetic parameters. Figure 3 shows the effect of penalized estimation on estimated heritabilities and genetic correlations and their sampling variances for scenarios A and D (Bondari's design with 125 families). For case A, there were very few replicates with estimates at the boundary of the parameter space, and all parameters were thus estimated without notable bias while variation in estimates was reduced by orders of magnitude. For case D, with a substantial spread in population heritabilities, however, penalized estimation caused the largest heritabilities to

be biased downward and the smallest values to be biased upward, corresponding to the overshrinkage of eigenvalues evident in Figure 2. The reduction in sampling variances achieved for this scenario was less (but still substantial for most parameters) and largest for correlations between traits with low heritabilities.

Penalization had very little effect on the estimates of eigenvectors, with only a slight increase in the average angle between true and estimated vectors apparent. Hence, the reduction in risk achieved, summarized in Tables 1 and 2, is a direct reflection of the effects of penalties on the estimates of the canonical heritabilities. Average tuning factors decreased with increasing spread in the population eigenvalues and were markedly lower for penalties applied on the logarithmic scale. For most scenarios, values for $\hat{\psi}$ tended to increase with decreasing sample size; *i.e.*, as to be expected, more severe penalties needed to be imposed as fewer data were available and sampling variation increased. The main exception to this pattern was for the cases with identical population values (A and G). For these, $\hat{\psi}$ was equal to the maximum value considered (1000) for the more replicates the larger the sample size. This implies that the average value was somewhat distorted by the upper limit on $\hat{\psi}$ imposed, with the decrease reflecting more variability for smaller samples.

Risks for $\hat{\Sigma}_G$ were largest for scenarios with a wide spread of roots. For reasonable sample sizes, penalized estimation reduced the average loss in $\hat{\Sigma}_G$ throughout, with reductions increasing as the spread in population values decreased. Penalties on the logarithmic scale appeared most advantageous for scenarios with one large eigenvalue and the remaining values close together (E, I, J, and K). For constellations with a large

FIGURE 2.—Mean estimates of canonical heritabilities, as deviation from population values (in percent) together with plus or minus one empirical standard deviation (vertical bars; truncated for case I) for standard REML analyses ($\square$), and penalties on eigenvalues on the original ($\circ$) and logarithmic (Δ) scale, for Bondari's design with 125 families.

spread in λ_i (D and, to a lesser extent, F) penalization increased the loss in $\hat{\Sigma}_G$ in up to a third of replicates; on average, though, there was a reduction in risk.

Relatively small PRIALs for $\hat{\Sigma}_G$ for cases with the largest population value close to unity (F and K) reflected, in part at least, the effects of constraints on the parameter space that decreased the scope for penalization to reduce risk. While constraints biased the average of $\bar{\lambda}$ across replicates only slightly (depending on the scenario, by $<4\%$ up- or downward), effects for individual replicates may have been larger, resulting in attempts to penalize deviations from a less appropriate estimate of the mean than we may wish for. Additional simulations (not shown here) yielded a higher PRIAL for cases D , F , I , J , and K when replacing $\bar{\lambda}$ (original scale) in Equation 5 with the corresponding harmonic mean.

Again, the pattern of results for the two designs was comparable, suggesting that bending is just as effective in a complex pedigree as in the paternal half-sib design it was originally suggested for. Values of PRIAL for the same sample size (100 sires and 125 families in simulations I and II, respectively) were generally smaller for Bondari's design. This was accompanied by smaller values for $\bar{L}_1(\hat{\Sigma}_G^0, \Sigma_G)$; *i.e.*, with numerous covariances between relatives the same number of observations provided more information so that the effects of sampling variations were less and penalized estimation had somewhat less impact.

APPLICATION

Application of the procedure suggested is illustrated with data for carcass measurements of beef cattle. This is

a typical example of traits considered in livestock improvement schemes that are difficult and expensive to record but play a major role in breeding programs. Data were collected from abattoirs under a meat quality research project and have been analyzed previously; see REVERTER *et al.* (2000) for details.

A total of six traits recorded on up to 1796 animals were considered, with 916, 1524, 1796, 1796, 1784, and 1672 records for traits 1–6 (retail beef yield, percentage intramuscular fat, rump fat depth, carcass weight, rib fat depth, and eye muscle area), respectively. Only 44% of individuals had all six traits recorded. All records were preadjusted for differences in age at slaughter or carcass weight as described in REVERTER *et al.* (2000). Animals in the data were the progeny of 130 sires and 932 dams, and no parents had records themselves. Adding pedigree information yielded an additional 3105 animals to be included, *i.e.*, a total of 4901 in the analysis. Data and pedigrees are available in the [supporting information, File S1](#).

The model of analysis was a simple animal model, fitting animals' additive genetic effects as random effects. The only fixed effects fitted were those of "contemporary groups" (CG) that represented a combination of herd of origin, sex of animal, date of slaughter, abattoir, finishing regime, and target market subclasses, with up to 282 levels per trait. Estimates of genetic and environmental covariance matrices were obtained by REML, using an "average information" algorithm followed by a derivative-free search to ensure the maximum of the likelihood had been located with reasonable accuracy. Both a standard multivariate analysis and analyses imposing a penalty on the squared deviation of the canonical heritabilities from their mean as described above were carried out.

FIGURE 3.—Box-and-whisker plot for nonpenalized (light shading) and penalized (dark shading) estimates of heritabilities (left column) and genetic correlations (right column) for case A (top row) and case D (bottom row), for Bondari's design with 125 families and penalties on untransformed canonical heritabilities.

The tuning parameter ψ was estimated using 10-fold cross-validation, as described above. This required splitting the data into 10 subsets. To avoid problems arising from dividing small CG subclasses in doing so, data were split by sequentially assigning all animals in a CG (for trait 4) to a subset, processing CGs in order of size. For example, subset 1 consisted of records on all animals in CGs 1, 11, 21, ..., subset 2 comprised animals in CGs 2, 12, 22, and so forth. For each of the 10 folds, the i th subset was designated the validation set and the remaining 9 subsets were merged to form the training data. Estimates of covariance components were then obtained from the training data for a range of tuning parameters, and the corresponding values for the (unpenalized) REML likelihood of these estimates in the validation data were calculated. Initially values of $\psi = 0, 1, 2, \dots, 20$ and $\psi = 25, 30, \dots, 100$ were considered, and, in a second pass, all values between 20 and 35 in steps of 1 were evaluated; *i.e.*, cross-validation involved 48×10 analyses and likelihood evaluations. The value of ψ then chosen as "best" was that for which the average of the 10 validation set likelihood values was highest.

Results: Estimates of canonical heritabilities from a standard, unpenalized analysis together with their approximate standard errors (derived from the inverse of the average information matrix at convergence) were $0.89 \pm 0.14, 0.54 \pm 0.10, 0.38 \pm 0.09, 0.24 \pm 0.09, 0.14 \pm 0.07$, and 0.03 ± 0.05 , with a mean of 0.37. Conducting a simulation study, corresponding to simulation II above with 125 families, for six traits (measured on all individuals) with canonical heritabilities of 0.8, 0.5, 0.4, 0.3, 0.2, and 0.1 yielded an average estimate of the tuning parameter of 34, with PRIAL of 19% for $\hat{\Sigma}_G$ and

of 39% for $\hat{\Sigma}_E$, respectively. In line with these results, cross-validation yielded an estimate for the tuning parameter of $\hat{\psi} = 30$. Corresponding estimates of canonical heritabilities from the penalized analysis were $0.69 \pm 0.11, 0.50 \pm 0.09, 0.38 \pm 0.08, 0.27 \pm 0.08, 0.17 \pm 0.07$, and 0.05 ± 0.05 , with a mean of 0.34. The likelihood for this set of estimates was reduced by 1.32 compared to the value from the unpenalized analysis; *i.e.*, penalization for such relatively mild penalty did not decrease the likelihood significantly even though the estimate of λ_1 was reduced by $>20\%$.

Resulting estimates of heritabilities (on the original scale) and correlations from the two analyses are contrasted in Figure 4. On the whole, there was good agreement between analyses with most penalized estimates well within the range of plus/minus one standard deviation from their unpenalized counterparts. Penalization reduced the estimates of the higher heritabilities and slightly increased the lowest value. In addition, it tended to reduce the magnitude of higher (absolute value) estimates of genetic correlations somewhat. Reassuringly, changes were largest for trait 1, the trait with the smallest number of records. In particular, an estimate of the environmental correlation between traits 1 and 4 was reduced from 0.82 to 0.63.

DISCUSSION

Accurate multivariate estimation of genetic covariance matrices is a longstanding problem. Mixed-model-based estimation, considering more than just a few traits and fitting the so-called animal model to accommodate complex pedigrees, has become feasible on a routine basis, due to advancement in computing facilities and

TABLE 2

Reduction in average loss (PRIAL, in percent), for estimates of the genetic (Σ_G), error (Σ_E), and phenotypic (Σ_P) covariance matrix, together with mean entropy loss ($\times 100$) in unpenalized REML estimates of $\hat{\Sigma}_G$ ($\bar{L}_1(\hat{\Sigma}_G, \Sigma_G)$), the proportion of replicates (W , in percent) for which penalized estimation increased the loss in Σ_G , and the mean tuning factor (ψ), for different constellations ($A, \dots, K$) of population values (see Table 1) and penalties on the original (ORG) or logarithmic (LOG) scale

	A	B	C	D	E	F	G	H	I	J	K
125 families											
$\bar{L}_1(\hat{\Sigma}_G^0, \Sigma_G)$	23	26	36	125	33	148	158	279	611	555	439
ORG											
Σ_G	86	63	45	65	41	60	94	83	40	55	44
Σ_E	72	51	29	19	11	32	63	43	11	15	33
Σ_P	8	6	3	2	1	1	2	2	0	1	1
W	0	1	13	32	3	19	0	2	8	10	9
ψ	949.8	326.1	71.7	32.4	38.2	22.5	948.8	323.5	25.4	46.9	16.7
LOG											
Σ_G	88	62	40	69	55	73	96	85	90	81	87
Σ_E	74	51	26	10	22	13	65	44	17	19	13
Σ_P	8	6	3	1	2	1	3	2	1	1	1
W	0	2	21	32	2	21	0	6	1	10	2
ψ	521.8	69.5	10.6	2.3	8.9	2.0	184.1	24.4	1.9	2.5	1.3
50 families											
$\bar{L}_1(\hat{\Sigma}_G^0, \Sigma_G)$	90	103	183	382	169	452	704	791	1102	1021	868
ORG											
Σ_G	91	85	80	76	78	70	96	90	47	65	44
Σ_E	75	65	46	34	27	60	64	53	19	26	63
Σ_P	9	7	5	3	2	1	2	2	1	1	1
W	0	0	6	25	1	14	0	1	7	5	8
ψ	865.7	411.5	84.9	34.0	35.7	20.0	851.8	407.6	26.4	55.0	13.6
LOG											
Σ_G	92	85	77	74	82	81	98	93	88	81	86
Σ_E	76	66	47	27	39	25	65	55	22	28	17
Σ_P	8	7	5	3	4	2	3	2	1	1	1
W	0	0	12	30	2	14	0	1	1	5	2
ψ	283.9	121.6	20.9	5.3	7.9	2.4	103.7	52.1	2.1	4.4	1.3

Bondari's design with 125 or 50 families is shown.

improvements in software available. However, problems associated with substantial sampling variation inherent in multivariate estimation, especially for relatively small data sets, remain. In particular, the fact that large eigenvalues tend to be biased upward while small eigenvalues tend to be biased downward is generally given little consideration. Emphasis on unbiased estimation of breeding values has fostered a corresponding preference for unbiased methods of estimation, often ignoring the fact that standard methods such as REML are biased as they require estimates to be within the parameter space, *i.e.*, constrain estimates of covariance matrices to be positive (semi-) definite.

Regularization: Our review has shown that trading additional bias against a lower statistical risk in the estimation of covariance matrices is a well-established practice. The literature ranges from theoretical studies, which are predominantly focused on establishing that certain classes of improved estimators dominate over others, to applications that demonstrate that using

regularized estimates of covariance matrices in regression problems, discriminant analyses, or portfolio estimation results in more reliable estimates or statistical tests. In a quantitative genetic context, an early form of regularization—though not labeled as such—has been suggested in the form of bending and has been shown to improve the achieved response to selection based on indexes derived using regularized estimates of genetic covariance matrices (HAYES and HILL 1981).

We propose to implement the equivalent to bending in REML analyses fitting an animal model by penalizing the corresponding log likelihood, with the penalty term proportional to the sum of squared deviations of the canonical heritabilities from their mean. Our simulation results demonstrate the statistical risks associated with standard REML estimates of covariance matrices and show that these can be dramatically reduced using penalized estimation. On a relative scale, penalization is most effective when the population eigenvalues are close together, which is the scenario when sampling

FIGURE 4.—Estimates of genetic parameters (h_i^2 , heritability for trait i ; r_{ij} , correlation between traits i and j) for beef cattle example from standard (solid symbols) and penalized (open symbols) analyses ($\circ$, heritability; $\square$, genetic correlation; Δ , environmental correlation; vertical bars show range of one standard deviation on either side of estimates from standard analyses).

variances in estimates of eigenvalues are largest. However, mean risks increase considerably as the true roots are spread farther apart, so that a proportionally much smaller reduction for these cases can still represent a substantial decrease in absolute values.

Analyses examining the eigenvalues of estimated genetic covariance matrices usually show that a substantial proportion of the total genetic variance is explained by the leading principal components, with genetic eigenvalues declining in an approximately exponential fashion (KIRKPATRICK 2009). Corresponding canonical heritabilities have not been examined in a methodical fashion. While the pattern in genetic eigenvalues does not imply that the eigenvalues of $\Sigma_P^{-1}\Sigma_G$ follow suit, our applied example suggests that practical cases with a relatively large spread may not be unusual.

Clearly, an alternative to penalized (RE)ML is Bayesian estimation where regularization is implicit through the prior distributions specified. While such analyses have become a standard in quantitative genetics, uninformative priors appear to be used more often than not, at least in variance component estimation; *i.e.*, “only lip service is paid to the Bayesian paradigm” (THOMPSON *et al.* 2005, p. 1471). This demonstrates that specification of suitable prior distributions or of the associated hyperparameters is often not all that straightforward. Hence penalized REML may provide an easier option in practice.

Tuning parameter: While penalized estimation is appealing, it requires a decision on the strength of the penalty to be applied, *i.e.*, identification of a suitable tuning parameter, ψ . As outlined, cross-validation (K -fold or using random splits) is a widely used technique to determine this quantity from the data at hand and is applicable in our case. However, it is a laborious procedure that can increase computational requirements of penalized estimation by orders of magnitude, compared to standard, nonpenalized estimation. Furthermore, for data with a genetic family structure and usually affected by one or more cross-classified fixed

effects, representative subsampling can be challenging, especially for small samples. There appears to be a paucity of literature addressing this problem or examining other strategies such as bootstrapping.

An alternative may be to choose a penalty factor *a priori*, analogous to the choice for the degree of confidence to be made when specifying prior distributions in Bayesian estimation. Earlier, HAYES and HILL (1981) suggested bending on sample size alone. Our results indicate that this choice should consider pedigree structure and the spread in canonical heritabilities in addition. Of course the latter is unknown, and we again need to have some prior knowledge, perhaps on the basis of literature results, or try to assess this from the estimates obtained from a standard, nonpenalized analysis. Further work is required to see whether suitable rules of thumb can be established. Results dating back to JAMES and STEIN (1961) show that for a quadratic penalty, any amount of shrinkage ($\psi > 0$) in a certain range reduces the mean square error, with the range determined by the (co)variances of the standard estimator and the shrinkage target; see, for instance, OPGEN-RHEIN and STRIMMER (2007) for a lucid outline. This suggests that a simple approach to exploit at least some of the benefits of penalized multivariate estimation may be to impose a relatively mild penalty, chosen on the basis of the number of traits considered, sample size, and data structure and any information we may have on the spread of the eigenvalues. For this, “mild” might be defined as a value of ψ for which the unpenalized likelihood corresponding to the penalized estimates differs does not differ substantially from the maximum of $\log \mathcal{L}$ for $\psi = 0$.

Open questions: The purpose of this article has been to introduce the concept and demonstrate the utility of regularized estimation for the estimation of genetic parameters, considering multivariate, animal model-type analyses and (restricted) maximum likelihood. A number of problems remain for further research. These

include technical issues, such as extensions to models with additional random effects and random regression analyses, the best algorithm to be used to maximize the penalized likelihood, and which cross-validation strategy to use. In addition, there are more fundamental questions to be addressed, such as the most appropriate definition of risk for particular scenarios or the use of alternative prior distributions and penalties.

CONCLUSIONS

Penalized estimation of genetic covariance matrices can reduce the deviation of estimates from population values by reducing the inherent overdispersion in sample roots, producing “better” estimates. It is recommended for multivariate analyses comprising more than a few traits and small samples to make the best possible use of limited and precious data.

The Animal Genetics and Breeding Unit, University of New England, is a joint venture with Industries & Investment, New South Wales. This work was supported by Meat and Livestock Australia under grant BFGEN.100B (to K.M.); M.K. is grateful for support from National Science Foundation grant DEB-0819901 and the Miller Institute for Basic Research.

LITERATURE CITED

- AMEMIYA, Y., 1985 What should be done when an estimated between-group covariance matrix is not nonnegative definite? *Am. Stat.* **39**: 112–117.
- 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*, Ed. 2. Wiley, New York.
- 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.
- BICKEL, P. J., and E. LEVINA, 2008 Regularized estimation of large covariance matrices. *Ann. Stat.* **36**: 199–227.
- BICKEL, P. J., and B. LI, 2006 Regularization in statistics. *Test* **15**: 271–303.
- BÖHM, H., 2008 Shrinkage methods for multivariate spectral analysis. Ph.D. Thesis, Catholic University, Louvain, Belgium.
- BONDARI, K., R. L. WILLHAM and A. E. FREEMAN, 1978 Estimates of direct and maternal genetic correlations for pupa weight and family size of *Tribolium*. *J. Anim. Sci.* **47**: 358–365.
- DANIELS, M. J., and R. E. KASS, 2001 Shrinkage estimators for covariance matrices. *Biometrics* **57**: 1173–1184.
- DEMPSTER, A. P., 1972 Covariance selection. *Biometrics* **28**: 157–175.
- DEY, D., and C. SRINIVASAN, 1985 Estimation of a covariance matrix under Stein's loss. *Ann. Stat.* **13**: 1581–1591.
- FOSTER, S. D., A. P. VERBYLA and W. S. PITCHFORD, 2009 Estimation, prediction and inference for the LASSO random effects model. *Aust. N. Z. J. Stat.* **51**: 43–61.
- FRANK, I. E., and J. H. FRIEDMAN, 1993 A statistical view of some chemometrics regression tools. *Technometrics* **35**: 109–135.
- FRIEDMAN, J., T. HASTIE and R. TIBSHIRANI, 2008 Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**: 432–441.
- GREEN, P. J., 1998 Penalized likelihood, pp. 578–586 in *Encyclopedia of Statistical Sciences*, Vol. 2. John Wiley & Sons, New York.
- HAFF, L. R., 1980 Empirical Bayes estimation of the multivariate normal covariance matrix. *Ann. Stat.* **8**: 586–597.
- HARVILLE, D. A., 1977 Maximum likelihood approaches to variance component estimation and related problems. *J. Am. Stat. Assoc.* **72**: 320–338.
- HASTIE, T., R. TIBSHIRANI and J. FRIEDMAN, 2001 *The Elements of Statistical Learning* (Springer Series in Statistics). Springer-Verlag, New York.
- 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.
- HILL, W. G., and R. THOMPSON, 1978 Probabilities of non-positive definite between-group or genetic covariance matrices. *Biometrics* **34**: 429–439.
- HOERL, A. E., and R. W. KENNARD, 1970 Ridge regression: applications to nonorthogonal problems. *Technometrics* **12**: 69–82.
- HOFFMANN, K., 2000 Stein estimation—a review. *Stat. Papers* **41**: 127–158.
- HUANG, J. Z., N. LIU, M. POURAHMADI and L. LIU, 2006 Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika* **93**: 85–98.
- JAMES, W., and C. STEIN, 1961 Estimation with quadratic loss, pp. 361–379 in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistical Problems*, Vol. 1, edited by J. NEYMAN. University of California Press, Berkeley, CA.
- KIRKPATRICK, M., 2009 Patterns of quantitative genetic variation in multiple dimensions. *Genetica* **136**: 271–284.
- KIRKPATRICK, M., and K. MEYER, 2004 Direct estimation of genetic principal components: simplified analysis of complex phenotypes. *Genetics* **168**: 2295–2306.
- KIRKPATRICK, M., D. LOFSVOLD and M. BULMER, 1990 Analysis of the inheritance, selection and evolution of growth trajectories. *Genetics* **124**: 979–993.
- 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.
- KRUUK, L. E. B., J. SLATE and A. J. WILSON, 2008 New answers for old questions: the evolutionary quantitative genetics of wild animal populations. *Annu. Rev. Ecol. Evol. Syst.* **39**: 525–548.
- KUBOKAWA, T., 1999 Shrinkage and modification techniques in estimation of variance and the related problems: a review. *Commun. Stat. Theory Methods* **28**: 613–650.
- KUBOKAWA, T., and M. T. TSAI, 2006 Estimation of covariance matrices in fixed and mixed effects linear models. *J. Multivar. Anal.* **97**: 2242–2261.
- 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. Multivar. Anal.* **88**: 365–411.
- LEVINA, E., A. J. ROTHMAN and J. ZHU, 2008 Sparse estimation of large covariance matrices via a nested Lasso penalty. *Ann. Appl. Stat.* **2**: 245–263.
- LIN, S. P., and M. D. PERLMAN, 1985 A Monte Carlo comparison of four estimators of a covariance matrix, pp. 411–428 in *Multivariate Analysis*, Vol. 6, edited by P. R. KRISHNAISH. North-Holland, Amsterdam.
- LOH, W. L., 1991 Estimating covariance matrices. *Ann. Stat.* **19**: 283–296.
- MATHEW, T., A. NIYOGI and B. K. SINHA, 1994 Improved nonnegative estimation of variance components in balanced multivariate mixed models. *J. Multivar. Anal.* **51**: 83–101.
- MENG, X. L., 2008 Who cares if it is a white cat or a black cat? Discussion: “One-step sparse estimates in nonconcave penalized likelihood models” [Ann. Statist. **36** (2008), 1509–1533] by H. Zou and R. Li. *Ann. Stat.* **36**: 1542–1552.
- MEYER, K., 2009 Factor-analytic models for genotype x environment type problems and structured covariance matrices. *Genet. Sel. Evol.* **41**: 21.
- MEYER, K., and W. G. HILL, 1983 A note on the effects of sampling errors on the accuracy of genetic selection indices. *Z. Tierzucht Zuechtungsbiol.* **100**: 27–32.
- MEYER, K., and M. KIRKPATRICK, 2008 Perils of parsimony: properties of reduced rank estimates of genetic covariances. *Genetics* **108**: 1153–1166.
- ODELL, P. L., and A. H. FEIVSON, 1966 A numerical procedure to generate a sample covariance matrix. *J. Am. Stat. Assoc.* **61**: 199–203.
- OPGEN-RHEIN, R., and K. STRIMMER, 2007 Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach. *Stat. Appl. Genet. Mol. Biol.* **6**: 9.

- PINHEIRO, J. C., and D. M. BATES, 1996 Unconstrained parameterizations for variance-covariance matrices. *Stat. Comput.* **6**: 289–296.
- POURAHMADI, M., 1999 Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation. *Biometrika* **86**: 677–690.
- REVERTER, A., D. J. JOHNSTON, H.-U. GRASER, M. L. WOLCOTT and W. H. UPTON, 2000 Genetic analyses of live-animal ultrasound and abattoir carcass traits in Australian Angus and Hereford cattle. *J. Anim. Sci.* **78**: 1786–1795.
- ROTHMAN, A. J., P. J. BICKEL, E. LEVINA and J. ZHU, 2008 Sparse permutation invariant covariance estimation. *Electron. J. Stat.* **2**: 494–515.
- ROTHMAN, A. J., E. LEVINA and J. ZHU, 2009 Generalized thresholding of large covariance matrices. *J. Am. Stat. Assoc.* **104**: 177–186.
- RUPPERT, D., M. P. WAND and R. J. CARROLL, 2003 *Semiparametric Regression*. Cambridge University Press, New York.
- SANCETTA, A., 2008 Sample covariance shrinkage for high dimensional dependent data. *J. Multivar. Anal.* **99**: 949–967.
- 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.
- SRIVASTAVA, M. S., and T. KUBOKAWA, 1999 Improved non-negative estimation of multivariate components of variance. *Ann. Stat.* **27**: 2008–2032.
- STEIN, C., 1975 Estimation of a covariance matrix. Reitz lecture. 39th Annual Meeting of the Institute of Mathematical Statistics, Atlanta.
- THOMPSON, R., 1976 The estimation of maternal genetic variance. *Biometrics* **32**: 903–917.
- THOMPSON, R., S. BROTHERSTONE and I. M. S. WHITE, 2005 Estimation of quantitative genetic parameters. *Philos. Trans. R. Soc. B* **360**: 1469–1477.
- TIBSHIRANI, R., 1996 Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B* **58**: 267–288.
- TUTZ, G., and J. ULBRICHT, 2009 Penalized regression with correlation-based penalty. *Stat. Comput.* **19**: 239–253.
- WARTON, D. I., 2008 Penalized normal likelihood and ridge regularization of correlation and covariance matrices. *J. Am. Stat. Assoc.* **103**: 340–349.
- YAP, J. S., J. FAN and R. WU, 2009 Nonparametric modeling of longitudinal covariance structure in functional mapping of quantitative trait loci. *Biometrics* **65**: 1068–1077.
- YE, R. D., and S. G. WANG, 2009 Improved estimation of the covariance matrix under Stein's loss. *Stat. Probab. Lett.* **79**: 715–721.
- ZOU, H., and T. HASTIE, 2005 Regularization and variable selection via the elastic net. *J. R. Stat. Soc. B* **67**: 301–320.

Communicating editor: J. WAKEFIELD

APPENDIX

Let Δ_{ij} represent a $q \times q$ matrix with ij th element of unity and zero otherwise. The nonzero derivatives needed to adapt a standard REML algorithm to the canonical parameterization and penalized estimation for $\mathcal{P} = \sum_{i=1}^q (\lambda_i - \bar{\lambda})^2$ are given in the following.

First derivatives:

$$\begin{aligned}\frac{\partial \Sigma_A}{\partial \lambda_i} &= \mathbf{T} \Delta_{ii} \mathbf{T} & \frac{\partial \Sigma_A}{\partial t_{ij}} &= \Delta_{ij} \Lambda \mathbf{T} + \mathbf{T} \Lambda \Delta_{ij} \\ \frac{\partial \Sigma_E}{\partial \lambda_i} &= -\mathbf{T} \Delta_{ii} \mathbf{T} & \frac{\partial \Sigma_E}{\partial t_{ij}} &= \Delta_{ij} (\mathbf{I} - \Lambda) \mathbf{T} + \mathbf{T} (\mathbf{I} - \Lambda) \Delta_{ij} \\ \frac{\partial \mathcal{P}}{\partial \lambda_i} &= 2(\lambda_i - \bar{\lambda}).\end{aligned}$$

Second derivatives:

$$\begin{aligned}\frac{\partial^2 \Sigma_A}{\partial t_{ij} \partial \lambda_k} &= \Delta_{ij} \Delta_{kk} \mathbf{T} + \mathbf{T} \Delta_{kk} \Delta_{ij} & \frac{\partial^2 \Sigma_A}{\partial t_{ij} \partial t_{kl}} &= \Delta_{ij} \Lambda \Delta_{kl} + \Delta_{kl} \Lambda \Delta_{ij} \\ \frac{\partial^2 \Sigma_E}{\partial t_{ij} \partial \lambda_k} &= -(\Delta_{ij} \Delta_{kk} \mathbf{T} + \mathbf{T} \Delta_{kk} \Delta_{ij}) & \frac{\partial^2 \Sigma_E}{\partial t_{ij} \partial t_{kl}} &= \Delta_{ij} (\mathbf{I} - \Lambda) \Delta_{kl} + \Delta_{kl} (\mathbf{I} - \Lambda) \Delta_{ij} \\ \frac{\partial^2 \mathcal{P}}{\partial \lambda_i \partial \lambda_j} &= 2 \left(\delta_{ij} - \frac{1}{q} \right) & \text{with } \delta_{ij} &= \begin{cases} 1 & \text{for } i = j \\ 0 & \text{for } i \neq j. \end{cases}\end{aligned}$$