

# **Better estimates of genetic covariance matrices by ‘bending’ using penalized maximum likelihood**

Karin Meyer\* and Mark Kirkpatrick†

\*Animal Genetics and Breeding Unit<sup>1</sup>, University of New England, Armidale NSW 2351,  
Australia

and

† Section of Integrative Biology, 1 University Station C-0930, University of Texas, Austin,  
Texas 78712, USA

Running head: Bending over backwards

Keywords: Genetic covariance matrix, regularization, bias, sampling variance, bending

Corresponding author: Karin Meyer,  
Animal Genetics and Breeding Unit,  
University of New England,  
Armidale NSW 2351,  
Australia

E-mail : [kmeyer@didgeridoo.une.edu.au](mailto:kmeyer@didgeridoo.une.edu.au)

Phone : +61 2 6773 3331

Fax : +61 2 6773 3266

<sup>1</sup>a joint venture with NSW Agriculture

## ABSTRACT

Multivariate analyses to estimate genetic covariance matrices are generally subject to substantial sampling variation and bias in the estimates of genetic eigenvalues. The paper explores the use of regularization techniques to obtain better estimates. After a review of the underlying principles of statistical risk, shrinkage estimators and penalized maximum estimation, a restricted maximum likelihood procedure is described which implements the equivalent to ‘bending’ within ‘animal model’ type analyses. This is achieved by imposing a penalty on the deviation of the canonical eigenvalues (i.e. the eigenvalues of the product of the genetic and the inverse of the phenotypic covariance matrix) from their mean, and can be interpreted as ‘borrowing strength’ from the phenotypic covariance matrix, which is generally estimated much more accurately than the genetic covariance matrix. A simulation study demonstrates that penalized estimation can substantially reduce the average loss in estimates, i.e. the deviation of estimates from population values, even for analyses of moderate dimensions. Improvements in estimates are largest for small samples and scenarios where the population canonical eigenvalues are close together. An application to data from beef cattle is given showing the effects of regularization on estimates of heritabilities and correlations. While penalized estimation increases the computational requirements it can be recommended for multivariate analyses involving more than a few traits and problems with limited data.

## INTRODUCTION

Problems inherent in multivariate estimation of covariance components, especially for small samples or larger numbers of variables (traits), are well known. These arise predominantly from sampling variation, in particular the over-dispersion of sample eigenvalues, and are exacerbated when two or more matrices have to be considered simultaneously, as for genetic parameter estimation.

There has been longstanding interest in the ‘regularization’ of covariance matrices, in particular for cases with a the ratio between the number of observations and the number of variables is small. A variety of recent studies 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 which have less statistical risk than

30 standard methods; see BICKEL and LI (2006) for a review. For instance, various types of shrink-  
31 age estimators of covariance matrices have been suggested which counter-act upwards bias of  
32 the largest and downwards bias of the smallest eigenvalues by shrinking all sample eigenvalues  
33 towards their mean (see below for references). Often this is equivalent to a weighted combina-  
34 tion of the sample covariance matrix and a target matrix, assumed to have a simple structure.  
35 A common choice for the latter is an identity matrix. This yields a ridge regression type for-  
36 mulation (HOERL and KENNARD, 1970). Numerous simulation studies in a variety of settings  
37 are available which demonstrate that regularization can yield closer agreement between esti-  
38 mated and population covariance matrices, less variable estimates of model terms or improved  
39 performance of statistical tests.

40 In quantitative genetic analyses, we attempt to partition observed, overall (phenotypic) co-  
41 variance matrices into their genetic and environmental components. Typically, this results in  
42 strong sampling correlations between them. Hence, while the partitioning into sources of vari-  
43 ation and estimates of individual covariance matrices may be subject to substantial sampling  
44 variances, their sum, i.e. the phenotypic covariance matrix, can generally be estimated much  
45 more accurately. This has led to suggestions to ‘borrow strength’ from estimates of pheno-  
46 typic covariances matrices in estimating the genetic matrices. In particular, HAYES and HILL  
47 (1981) proposed a method termed ‘bending’ which involved regressing the eigenvalues of the  
48 product of the genetic and the inverse of the phenotypic covariance matrix towards their mean.  
49 One objective of this procedure was to ensure that estimates of the genetic covariance matrix  
50 from an analysis of variance were positive definite. In addition, the authors showed by sim-  
51 ulation that shrinking eigenvalues even further than needed to make all eigenvalues positive  
52 could improve the achieved response to selection when using the resulting estimates to derive  
53 weights for a selection index, especially for estimation based on small samples. Subsequent  
54 work demonstrated that ‘bending’ could also be advantageous in more general scenarios such  
55 as indexes which included information from relatives (MEYER and HILL, 1983).

56 Modern, mixed model (‘animal model’) based analyses to estimate genetic parameters using  
57 maximum likelihood or Bayesian methods generally constrain estimates to the parameter space,  
58 so that estimates of covariance matrices are positive semi-definite. However, the problems aris-  
59 ing from substantial sampling variation in multivariate analyses remain. In spite of increasing  
60 applications of such analyses in scenarios where data sets are invariably small, e.g. the analy-  
61 sis of data from natural populations (e.g. KRUIK *et al.*, 2008), there has been little interest in

62 regularization and shrinkage techniques in genetic parameter estimation, other than through the  
63 use of priors in a Bayesian context. Instead, suggestions for improved estimation have focused  
64 on parsimonious modelling of covariance matrices, e.g. through reduced rank estimation or  
65 by imposing a known structure, such as a factor-analytic structure (KIRKPATRICK and MEYER,  
66 2004; MEYER, 2009) or by fitting covariance functions for longitudinal data (KIRKPATRICK *et al.*,  
67 1990). While such methods can be highly advantageous when the underlying assumptions are,  
68 at least approximately, correct, data driven methods of regularization may be preferable in other  
69 scenarios.

70 This paper explores the scope for improved estimation of genetic covariance matrices by im-  
71 plementing ‘bending’ within ‘animal model’ type analyses. After a review of pertinent statis-  
72 tical literature, we describe a penalized restricted maximum likelihood (REML) procedure for  
73 estimation and present a simulation study demonstrating the effect of penalties on parameter  
74 estimates and their sampling properties. In addition, an application to data from beef cattle is  
75 shown.

## 76 **REVIEW: PRINCIPLES OF PENALIZED ESTIMATION**

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

### 83 **Minimizing statistical risk**

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

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

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

89 The quadratic loss is proportional to the mean square error. One property of  $L_2$  is that it imposes  
90 a heavier penalty for over- than for underestimates.

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

## 95 Improved estimators of covariance matrices

96 There is a considerable body of literature on ‘improved’ estimators of covariance matrices.  
97 These are generally biased, but have a lower risk than the standard, unbiased estimator (sample  
98 covariance matrix). Several studies derived the risk for a certain class of estimator and given  
99 loss function, and presented estimators which ‘dominate’ over other estimators followed by  
100 a simulation study to demonstrate their properties. Others obtained estimators using a differ-  
101 ent motivation, such as minimax or empirical Bayesian estimation; see KUBOKAWA (1999) and  
102 HOFFMANN (2000) for reviews.

103 Sampling variation causes the largest eigenvalues of a covariance matrix to be overestimated  
104 and the smallest eigenvalues to be underestimated, while their mean is expected to be unbi-  
105 ased. Hence, attention has focused on estimators which modify the eigenvalues of the sam-  
106 ple covariance matrix whilst retaining the corresponding eigenvectors. The impetus for this  
107 is generally attributed to STEIN (1975), but similar suggestions can be found earlier, for in-  
108 stance, in LAWLEY (1956). Let  $\hat{\omega}_i$  denote the  $i$ -th eigenvalue of the sample covariance ma-  
109 trix. Stein’s proposal then consisted of an adaptive shrinking obtained by scaling each  $\hat{\omega}_i$  by  
110  $d/(d - q + 1 + 2\hat{\omega}_i \sum_{j \neq i} (\hat{\omega}_i - \hat{\omega}_j)^{-1})$ . The resulting estimator minimizes the entropy loss but does  
111 not preserve the order of eigenvalues nor ensure non-negativity. Hence, later work often com-  
112 bined this with order preserving measures such as an ‘isotonizing’ regression (which restores  
113 order by merging values out of line) or truncation at zero (DEY and SRINIVASAN, 1985; LIN and  
114 PERLMAN, 1985; YE and WANG, 2009).

115 A simple modification scheme entails the linear shrinkage of the sample eigenvalues towards  
 116 their mean. It can be shown that this yields an estimator which is a weighted combination  
 117 of the sample covariance matrix and an identity matrix. Considering a quadratic loss func-  
 118 tion, LEDOIT and WOLF (2004) derived an optimal shrinkage factor  $\rho \in [0, 1]$  which minimized  
 119 the risk associated with the estimator  $\rho\bar{\omega}\mathbf{I} + (1 - \rho)\mathbf{S}/d$ , with  $\mathbf{I}$  an identity matrix and  $\bar{\omega}$   
 120 the mean (sample) eigenvalue. DANIELS and KASS (2001) argued that, due to the nature of  
 121 the quadratic loss, such estimator could result in over-shrinkage, in particular of the smallest  
 122 eigenvalues, when the true eigenvalues were far apart. Instead the authors proposed an esti-  
 123 mator derived by assuming a prior normal distribution for the eigenvalues on the logarithmic  
 124 scale, approximated as  $\log(\hat{\omega}_i) \propto N(\log(\omega_i), 2/n)$ . This resulted in modified values of the form  
 125  $\tilde{\omega}_i = \exp(\rho \log(\bar{\omega}) + (1 - \rho) \log(\hat{\omega}_i))$  (with  $\log(\bar{\omega})$  the mean of  $\log(\hat{\omega}_i)$ , i.e. again involved a re-  
 126 gression towards the sample mean, but on a different scale. WARTON (2008) proposed a similar,  
 127 regularized estimator of the sample correlation matrix  $\hat{\mathbf{R}}$ ,  $\rho\hat{\mathbf{R}} + (1 - \rho)\mathbf{I}$ , and showed that this  
 128 was the penalized maximum likelihood estimator with penalty term proportional to  $-\text{tr}(\mathbf{R}^{-1})$ ,  
 129 with the corollary that the corresponding, ridge type estimator of a covariance matrix  $\Sigma$ ,  $\hat{\Sigma} + \kappa\mathbf{I}$ ,  
 130 involved a penalty proportional to  $-\text{tr}(\Sigma^{-1})$ .

131 Other work considered shrinkage towards a more general structure. SCHÄFER and STRIMMER  
 132 (2005) and SANCETTA (2008) extended the approach of LEDOIT and WOLF (2004) to different  
 133 target matrices of simple structure with few parameters to be estimated, e.g. a diagonal ma-  
 134 trix with different variances or a matrix with all correlations equal. BÖHM (2008) examined  
 135 estimators for multivariate time series and suggested data driven shrinkage towards a factor-  
 136 analytic structure. Shrinkage estimators of correlation matrices have been described by LIN and  
 137 PERLMAN (1985), DANIELS and KASS (2001) and WARTON (2008).

138 **More than one matrix:** Few studies have addressed improved estimation for multi-level mod-  
 139 els. The simplest case, with two matrices to be estimated, is a balanced one-way classification.  
 140 Let  $\Sigma_B$  and  $\Sigma_W$  denote the covariance matrices between and within groups, respectively, and  $\mathbf{B}$   
 141 and  $\mathbf{W}$  denote the corresponding matrices of mean squares and cross-products (MSCP). Deriva-  
 142 tions of improved estimators by and large utilized the so-called canonical decomposition of  $\mathbf{B}$   
 143 and  $\mathbf{W}$ : For any two symmetric (real) matrices,  $\mathbf{W}$  and  $\mathbf{B}$ , of size  $q \times q$  with  $\mathbf{W}$  positive-definite  
 144 and  $\mathbf{B}$  positive semi-definite (*p.s.d.*), there exists a matrix  $\mathbf{T}$  such that  $\mathbf{T}\mathbf{T}' = \mathbf{W}$  and  $\mathbf{T}\mathbf{A}\mathbf{T}' = \mathbf{B}$ ,  
 145 with  $\mathbf{A} = \text{Diag}\{\lambda_i\}$  the diagonal matrix of eigenvalues of  $\mathbf{W}^{-1}\mathbf{B}$  (ANDERSON, 1984).

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

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

167 LOH (1991), MATHEW *et al.* (1994), SRIVASTAVA and KUBOKAWA (1999) and KUBOKAWA and TSAI  
 168 (2006) considered estimation for two independent Wishart matrices, such as  $\mathbf{B}$  and  $\mathbf{W}$  in the  
 169 one-way classification. Minimizing the sum of entropy losses, they derived different types of  
 170 joint estimators, analogous to those proposed for a single matrix, and showed that improved  
 171 estimators were available which had lower risk than the unbiased or REML estimators. How-  
 172 ever, no practical applications using any of these results are available. Again, these estimators  
 173 involved some form of modification of the eigenvalues arising from the canonical decompo-  
 174 sition of the two matrices, indicating that the suggestion of HAYES and HILL (1981) was well  
 175 founded.

## 176 Penalized maximum likelihood estimation

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

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

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

186 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,  
 187 regression coefficients and residuals, respectively, and  $\mathbf{X}$  the corresponding incidence matrix.  
 188 A class of penalties commonly employed is that of the  $\ell_p$  norm

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

189 with  $b_i$  the  $i$ -th element of  $\mathbf{b}$ . Different values of  $p$  are appropriate for different analyses.  
 190 For  $p = 0$ , the penalty is equal to the number of elements in  $\mathbf{b}$  and may be employed in  
 191 model selection. A value of  $p = 1$  yields a LASSO (least absolute shrinkage and selection  
 192 operator) type penalty which encourages shrinkage of small (absolute value) elements of  $\mathbf{b}$  to  
 193 zero and thus subset selection (TIBSHIRANI, 1996). A quadratic penalty ( $p = 2$ ) produces a ridge  
 194 regression formulation where all elements of  $\mathbf{b}$  are shrunk proportionally (HOERL and KENNARD,  
 195 1970). Non-integer values for  $p$  have been suggested, e.g. the so-called 'bridge', attributed to  
 196 FRANK and FRIEDMAN (1993), for values of  $0 < p < 1$ . Other proposals have been to combine  
 197  $\ell_1$  and  $\ell_2$  type penalties to form the 'elastic net' (ZOU and HASTIE, 2005), or to account for the  
 198 correlation among predictors in the penalty term (TUTZ and ULBRICHT, 2009).

199 In a more general framework, we may assume a certain prior distribution for the parameters to  
 200 be estimated and impose a penalty which is proportional to minus the logarithmic value of the

201 prior density. This provides a direct link to Bayesian estimation – indeed, penalized maximum  
202 likelihood estimation has been described as “an attempt of enjoying the Bayesian fruits without  
203 paying the B-club fee” (MENG, 2008). For instance, imposing an  $\ell_1$  type penalty is equivalent  
204 to assuming a double exponential prior distribution, while an  $\ell_2$  penalty implies a normal prior.

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

212 In practice, multiple splits are used. A popular scheme is that of  $K$ -fold cross-validation (e.g.  
213 HASTIE *et al.*, 2001, Chapter 7) where the data is evenly split into  $K$  subsets, and  $K$  analyses  
214 are carried out for each value of  $\psi$ , with the  $i$ -th subset treated as the validation set and the  
215 remaining  $K - 1$  subsets forming the training set. The tuning parameter is then chosen based  
216 on the objective function averaged across the  $K$  validation sets. Common values of  $K$  used  
217 are 5 or 10. A related technique is repeated random sub-sampling. For example, BICKEL and  
218 LEVINA (2008) employed a scheme using 50 ‘random splits’ of the data, with the training set  
219 comprising a third of the data.

220 In special cases,  $\psi$  can be estimated directly. An example is that of function estimation (smooth-  
221 ing) using a semi-parametric regression or penalized splines with a quadratic penalty. In that  
222 case, the regression can be rewritten as a mixed model with the penalized coefficients treated as  
223 random effects and the tuning parameter can be estimated ‘automatically’ in a (RE)ML analysis  
224 from the ratio of the residual variance and the variance due to random effects; see RUPPERT *et al.*  
225 (2003, Chapter 5) for an example. FOSTER *et al.* (2009) showed that a LASSO type penalty on  
226 effects can be imposed in a mixed model by treating these as random effects with a double ex-  
227 ponential distribution, and that the respective variance parameters (which determine the amount  
228 of penalization) can be estimated using a ML approach.

229 **Applications to covariance matrices:** Penalized ML estimation of covariance matrices has  
230 predominantly been applied in the spirit of covariance selection (DEMPSTER, 1972), i.e. to  
231 encourage sparsity in the estimate of inverse of the covariance matrix (‘concentration’ ma-

232 trix), and mostly for problems with high dimensions and some natural ordering of variables,  
 233 e.g. longitudinal data. This relied on the modified Cholesky decomposition  $\Sigma = \mathbf{LDL}'$  or  
 234  $\Sigma^{-1} = \mathbf{L}'\mathbf{D}^{-1}\mathbf{L}$ , with  $\mathbf{L}$  a lower triangular matrix with diagonal elements of unity and  $\mathbf{D}$  a  
 235 diagonal matrix. For longitudinal data, the non-zero off-diagonal elements of  $\mathbf{L}$  have an in-  
 236 terpretation as (minus) the regression coefficients in an auto-regressive model (POURAHMADI,  
 237 1999) and can be penalized in the same fashion as coefficients in a simple regression context.  
 238 Applications using both  $\ell_1$  and  $\ell_2$  type penalties can be found in HUANG *et al.* (2006), FRIEDMAN  
 239 *et al.* (2008), LEVINA *et al.* (2008), BICKEL and LEVINA (2008), ROTHMAN *et al.* (2008) and YAP  
 240 *et al.* (2009). With a different objective, WARTON (2008) considered penalized ML estimation  
 241 of covariance and correlation matrices to obtain regularized estimates with a stable inverse for  
 242 use in multivariate regression problems.

## 243 PENALIZED REML ESTIMATION

244 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  
 245 observations, fixed effects, additive genetic and residual effects, respectively, and  $\mathbf{X}$  and  $\mathbf{Z}$  the  
 246 corresponding incidence matrices. Let  $\Sigma_G$  and  $\Sigma_E$  denote the matrices of additive genetic and  
 247 residual covariances among the  $q$  traits, and let  $\text{Var}(\mathbf{g}) = \Sigma_G \otimes \mathbf{A} = \mathbf{G}$  with  $\mathbf{A}$  the numerator  
 248 relationship matrix between individuals and  $\text{Var}(\mathbf{e}) = \sum_k^+ \mathbf{R}_k = \mathbf{R}$ , with  $\mathbf{R}_k$  the sub-matrix of  
 249  $\Sigma_E$  corresponding to the traits recorded for the  $k$ -th individual and ‘ $\sum^+$ ’ is the direct matrix  
 250 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  
 251 constant,

$$\log \mathcal{L}(\theta) = -\frac{1}{2} \left( \log |\mathbf{V}| + \log |\mathbf{X}'_0 \mathbf{V}^{-1} \mathbf{X}_0| + (\mathbf{y} - \mathbf{X}\mathbf{b})' \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\mathbf{b}) \right) \quad (3)$$

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

253 Assuming  $\Sigma_G$  and  $\Sigma_E$  are unstructured, we have  $q(q+1)$  variance parameters to be estimated.  
 254 Maximization of (Eq. 3) with respect to the elements of  $\Sigma_G$  and  $\Sigma_E$  represents a constrained  
 255 optimisation problem, as estimates need to be in the parameter space, i.e. cannot have negative  
 256 eigenvalues. Hence implementations of REML estimation often employ a parameterisation  
 257 to a scale which does not require constraints (PINHEIRO and BATES, 1996), e.g. estimating the

258 elements of the Cholesky factor of a covariance matrix rather than the covariance components  
259 directly.

260 A natural alternative in our context is a parameterisation to the elements of the canonical de-  
261 composition of  $\Sigma_G$  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  $\mathbf{T} = \mathbf{Q}^{-1}$ ,  
262  $\mathbf{T}\Lambda\mathbf{T}' = \Sigma_G$  and  $\mathbf{T}\mathbf{T}' = \Sigma_P$ . The elements of  $\Lambda$  are the eigenvalues of  $\Sigma_P^{-1/2}\Sigma_G\Sigma_P^{-1/2}$  and  
263  $\mathbf{T}$  is the corresponding matrix of eigenvectors pre-multiplied by the matrix square root of  $\Sigma_P$ .  
264 This yields  $q$  parameters  $\lambda_i$  and  $q^2$  elements ( $t_{ij}$ ) of  $\mathbf{T}$  to be estimated. Eigenvalues  $\lambda_i$  can be  
265 interpreted as heritabilities on the canonical scale and are thus constrained to the interval  $[0, 1]$ .  
266 Again, these constraints can be removed through a suitable further reparameterisation, e.g. by  
267 estimating  $\log(-\log \lambda_i)$  instead of  $\lambda_i$ .

268 Our review above has identified that modification of the canonical eigenvalues of the ‘between’  
269 and ‘within’ matrices of MSCP in a one-way classification can provide ‘improved’ estimators  
270 of the corresponding covariance matrices, and that manipulation of these eigenvalues is equiv-  
271 alent to modifying the canonical eigenvalues of  $\Sigma_G$  and  $\Sigma_P$ . Furthermore, we have shown that  
272 regressing the eigenvalues of a matrix towards their mean yields a shrinkage estimator which  
273 is a weighted combination of the matrix and a multiple of the identity matrix, and that such  
274 estimators can be obtained by penalized maximum likelihood with an  $\ell_2$  type penalty. Using  
275 these findings, we propose to implement the equivalent to ‘bending’ within our REML animal  
276 model analysis by replacing  $\log \mathcal{L}$  in (Eq. 3) by

$$\log \mathcal{L}_P(\boldsymbol{\theta}) = \log \mathcal{L}(\boldsymbol{\theta}) - \frac{1}{2} \psi \mathcal{P}(\boldsymbol{\theta}) \quad \text{with} \quad \mathcal{P}(\boldsymbol{\theta}) = \sum_{i=1}^q (\lambda_i - \bar{\lambda})^2 \quad (4)$$

277 for  $\bar{\lambda} = (\sum_{i=1}^q \lambda_i) / q$ . This directly mimics the suggestion of HAYES and HILL (1981) as the  
278 quadratic penalty provides a linear shrinkage of all  $\lambda_i$  towards their mean. For  $\psi = 0$ ,  $\log \mathcal{L}_P$   
279 reduces to  $\log \mathcal{L}$ , and for  $\psi \rightarrow \infty$  (Eq. 4) yields a model in which all  $\lambda_i$  are constrained to be  
280 equal.

281 (Eq. 4) is readily extended to other types of penalties. For instance, rather than shrinking  
282 towards the arithmetic mean  $\bar{\lambda}$ , we could use the geometric or harmonic mean (YE and WANG,  
283 2009). An analogue of the log-posterior shrinkage estimator of DANIELS and KASS (2001) could  
284 be obtained by transforming eigenvalues  $\lambda_i$  to logarithmic scale. More generally, the shrinkage  
285 could be modified by applying a Box-Cox transformation, i.e. by replacing  $\lambda_i$  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 parameterisation to the elements of the canonical transformation, derivatives of  $\mathcal{P}(\theta)$  are straightforward, and standard REML algorithms, such expectation-maximisation 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 parameterisation, however, is that (first) derivatives of both  $\Sigma_G$  and  $\Sigma_E$  with respect to all  $q(q + 1)$  parameters to be estimated are non-zero. This implies that computational requirements per iterate are increased compared to the usual implementations.

## 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 ( $\lambda_i$ ). 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  $\lambda_i$  about their mean. Values for scenarios A to K are summarised 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  $40 \times 40$  matrix of MSCP pertaining to records for a family 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

314 of genetic and environmental covariance components were obtained using the canonical param-  
 315 eterisation as described above and a simple derivative-free optimisation procedure to locate the  
 316 maximum of the (penalized) likelihood function. Estimation constrained values of  $\lambda_i$  to the  
 317 range of 0.00001 to 0.99999. Quadratic penalties on the deviation of the canonical heritabilities  
 318 from their mean were imposed either on the eigenvalues directly or on values transformed to  
 319 logarithmic scale. A range of values for  $\psi$  (0 to 1.8 in steps of 0.2, 2 to 4.5 in steps of 0.5,  
 320 5 to 99 in steps of 1, 100 to 248 in steps of 2, 250 to 495 in steps of 5 and 500 to 1000 in  
 321 steps of 10) were considered. To estimate the appropriate tuning parameter, an additional set of  
 322 100 matrices of MSCP were sampled for each replicate. The best value,  $\hat{\psi}$ , was then chosen as  
 323 the value of  $\psi$  for which the average (unpenalized) likelihood for the corresponding estimates  
 324 across these validation sets was maximised. A total of 10 000 replicates were carried out for  
 325 each scenario examined.

326 As suggested by LIN and PERLMAN (1985), the effect of penalized estimation was then summa-  
 327 rized as percentage reduction in average loss (PRIAL), calculated as

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

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

## 330 Results

331 The effect of sampling variation and penalization on estimates of canonical heritabilities is  
 332 illustrated in Figure 1 and Figure 2. As expected from theory (e.g. LAWLEY, 1956), bias in  
 333 unpenalized estimates increased markedly with decreasing spread in the population values and  
 334 decreasing sample size. Patterns for both designs were similar. For scenarios with equal popu-  
 335 lation values (A and G), penalization dramatically reduced the bias in estimates with the small  
 336 remaining bias in the same direction as for unpenalized estimation. For the other cases pe-  
 337 nalization appeared to overcompensate somewhat, resulting in a bias in the opposite direction  
 338 for the extreme values ( $\lambda_1$  and  $\lambda_5$ ), i.e. yielded estimates of the largest values which were bi-  
 339 ased downwards and estimates of the smallest values which were biased upwards. Imposing a  
 340 penalty on the logarithmic scale tended to give estimates of  $\lambda_1$  which were less biased than for

penalization on the original scale but, in turn, yielded larger upward bias in estimates of  $\lambda_5$  in most cases. Over-shrinkage, 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).

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 Table 1 and Table 2, is a direct reflection of the effects of penalties on the estimates of the canonical heritabilities. 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 and K). For constellations with a large spread in  $\lambda_i$  (D and F) penalization increased the loss in  $\hat{\Sigma}_G$  in up to  $\frac{2}{3}$  of replicates; on average though there was some reduction in risk throughout, except for a very small sample (50 families or 400 records) together with a penalty on the logarithmic scale (case D).

For cases with the largest population value close to unity (F and K), penalized estimation increased the average loss in  $\hat{\Sigma}_E$  while still reducing the loss in  $\hat{\Sigma}_G$ . This was associated with a substantial proportion of replicates for which  $\hat{\lambda}_1$  was close to unity, so that  $\hat{\Sigma}_E$  was almost *p.s.d.* rather than firmly positive definite. Relatively small PRIALs for  $\hat{\Sigma}_G$  for these scenarios also reflected, in part at least, the effects of constraints on the parameter space which 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 less than 4% up- or downwards), 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  $\bar{\lambda}$  for cases D, F, I, J and K when replacing  $\bar{\lambda}$  (original scale) in (Eq. 4) 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 than for 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

372 number of observations provided more information so that the effects of sampling variations  
373 were less and penalized estimation had somewhat less impact.

## 374 APPLICATION

375 Application of the procedure suggested is illustrated with data for carcass measurements of  
376 beef cattle. This is a typical example of traits considered in livestock improvement schemes  
377 which are difficult and expensive to record but play a major role in breeding programmes. Data  
378 were collected from abattoirs under a meat quality research project and have been analysed  
379 previously; see REVERTER *et al.* (2000) for details.

380 A total of 6 traits recorded on 1796 animals were considered. All individuals had records for  
381 traits 3 (rump fat depth) and 4 (carcass weight), and there were 1784, 1524, 1671 and 916  
382 records for traits 5 (rib fat depth), 6 (eye muscle area), 2 (percentage intramuscular fat) and 1  
383 (retail beef yield), respectively. Only 44% of individual had all 6 traits recorded. All records  
384 were pre-adjusted for differences in age at slaughter or carcass weight as described in REVERTER  
385 *et al.* (2000). Animals in the data were the progeny of 130 sires and 932 dams. No parents had  
386 records themselves. Adding pedigree information yielded an additional 3105 animals to be  
387 included, i.e. a total of 4901 in the analysis.

388 The model of analysis was a simple animal model, fitting animals' additive genetic effects as  
389 random effect. The only fixed effects fitted were those of 'contemporary groups' (CG) which  
390 represented a combination of herd of origin, sex of animal, date of slaughter, abattoir, finishing  
391 regime and target market subclasses, with up to 282 levels per trait. Estimates of genetic and  
392 environmental covariance matrices were obtained by REML, using an 'average information'  
393 algorithm followed by derivative-free search to ensure the maximum of the likelihood had been  
394 located with reasonable accuracy. Both a standard multivariate analysis and analyses imposing  
395 a penalty on the squared deviation of the canonical heritabilities from their mean as described  
396 above were carried out. The tuning parameter  $\psi$  was estimated using 10-fold cross-validation.  
397 To avoid problems arising from dividing small CG subclasses in this procedure, data were split  
398 by assigning all animals in a CG (for trait 4) to a subset, processing CGs in order of size.  
399 Initially values of  $\psi = 0, 1, 2, \dots, 20$  and  $\psi = 25, 30, \dots, 100$  were considered, and, in a second  
400 pass, all values between 20 and 35 in steps of 1 were evaluated.

## 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  $0.03 \pm 0.05$ , with a mean of 0.37. Conducting a simulation study, corresponding to simulation II above with 125 families, for 6 traits (measured on all individuals) with canonical heritabilities of 0.8, 0.5, 0.4, 0.3, 0.2 and 0.1 suggested that this was a scenario in which a penalty on the eigenvalues would be preferable to a penalty on values transformed to logarithmic scale. For the simulation, the average estimate of the tuning parameter was 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  $0.05 \pm 0.05$ , 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  $\lambda_1$  was reduced by more than 20%.

Resulting estimates of heritabilities (on the original scale) and correlations from the two analyses are contrasted in Figure 3. 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 unusually high 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, both

430 due to advancement in computing facilities and improvements in software available. However,  
431 problems associated with substantial sampling variation, inherent in multivariate estimation  
432 especially for relatively small data sets, remain. In particular, the fact that large eigenvalues  
433 tend to be biased upwards while small eigenvalues tend to be biased downwards is generally  
434 given little consideration. Emphasis on unbiased estimation of breeding values has fostered  
435 a corresponding preference for unbiased methods of estimation, often ignoring the fact that  
436 standard methods such as REML are, by definition, biased as they require estimates to be  
437 within the parameter space, i.e. constrain estimates of covariance matrices to be positive (semi-  
438 ) definite.

439 Our review has shown that trading additional bias against a lower statistical risk in the esti-  
440 mation of covariance matrices is a well established practice. The literature available ranges  
441 from theoretical studies, which predominantly are interested in establishing that certain classes  
442 of ‘improved’ estimators dominate over others, to applications which demonstrate that using  
443 ‘regularized’ estimates of covariance matrices in regression problems, discriminant analyses or  
444 portfolio estimation results in more reliable estimates or statistical test. In a quantitative genetic  
445 context, an early form of regularization – though not labelled as such – has been suggested in  
446 the form of ‘bending’ and has been shown to improve the achieved response to selection based  
447 on indexes derived using regularized estimates of genetic covariance matrices (HAYES and HILL,  
448 1981).

449 We propose to implement the equivalent to ‘bending’ in REML analyses fitting an animal model  
450 by penalizing the corresponding log likelihood, with the penalty term proportional to the sum  
451 of squared deviations of the canonical heritabilities from their mean. Our simulation results  
452 demonstrate the statistical risks associated with standard REML estimates of covariance matri-  
453 ces and show that these can be dramatically reduced using penalized estimation. On a relative  
454 scale, penalization is most effective when the population eigenvalues are close together, which  
455 is the scenario when sampling variances in estimates of eigenvalues are largest. However, mean  
456 risks increase considerably as the true roots are spread further apart, so that a proportionally  
457 much smaller reduction for these cases can still represent a substantial decrease in absolute  
458 values.

459 Analyses examining the eigenvalues of estimated genetic covariance matrices usually show  
460 that a substantial proportion of the total genetic variance is explained by the leading princi-

461 pal components, with genetic eigenvalues declining in an approximately exponential fashion  
462 (KIRKPATRICK, 2009). Corresponding canonical heritabilities have not been examined in a me-  
463 thodical fashion. While the pattern in genetic eigenvalues does not imply that the eigenvalues  
464 of  $\Sigma_p^{-1}\Sigma_G$  follow suit, our applied example suggests that practical cases with a relatively large  
465 spread may not be unusual.

466 Clearly, an alternative to penalized (RE)ML is Bayesian estimation where regularization is  
467 implicit through the prior distributions specified. While such analyses have become a standard  
468 in quantitative genetics, for estimation of variance components uninformative priors appear to  
469 be used more often than not, i.e. “only lip service is paid to the Bayesian paradigm” (THOMPSON  
470 *et al.*, 2005). This demonstrates that specification of suitable prior distributions or of the  
471 associated hyper-parameters is often not all that straightforward. Hence penalized REML may  
472 provide an easier option in practice.

473 While penalized estimation is appealing, it can increase the computational requirements com-  
474 pared to standard REML analyses by orders of magnitude. In particular, cross-validation to  
475 estimate the tuning parameter can be laborious. An alternative may be to choose a mild penalty  
476 on the basis of sample size, pedigree structure and spread in the unpenalized estimates of the  
477 canonical heritabilities. Further work is required to see whether suitable rules of thumb can  
478 be established. In addition, using the parameterisation to canonical heritabilities and the cor-  
479 responding transformation matrix, computational requirements per iterate and the number of  
480 iterates required to locate the maximum of the penalized likelihood function tend to be higher  
481 than for standard REML. This is due to the fact that the covariance matrices,  $\Sigma_G$  and  $\Sigma_E$ , are  
482 ‘less’ linear in the parameters to be estimated than, say, a parameterisation to the elements of  
483 their Cholesky factors. Furthermore, derivatives with respect to all parameters to be estimated  
484 are non-zero increasing the amount of effort needed to compute derivatives of  $\log \mathcal{L}$  or  $\log \mathcal{L}_P$ .  
485 However, additional computing may be a small price to pay to make the best possible use of  
486 limited and precious data.

## 487 **Acknowledgments**

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

## APPENDIX

490  
 491 Let  $\Delta_{ij}$  represent a  $q \times q$  matrix with  $ij$ -th element of unity and zero otherwise. The non-zero  
 492 derivatives needed to adapt standard REML algorithm to the canonical parameterisation and  
 493 penalized estimation for  $\mathcal{P} = \sum_{i=1}^q (\lambda_i - \bar{\lambda})^2$  are given in the following.

### 494 First derivatives

$$\begin{aligned} \frac{\partial \Sigma_A}{\partial \lambda_i} &= \mathbf{T} \Delta_{ii} \mathbf{T}' & \frac{\partial \Sigma_A}{\partial t_{ij}} &= \Delta_{ij} \Lambda \mathbf{T}' + \mathbf{T} \Lambda \Delta'_{ij} \\ \frac{\partial \Sigma_E}{\partial \lambda_i} &= -\mathbf{T} \Delta_{ii} \mathbf{T}' & \frac{\partial \Sigma_E}{\partial t_{ij}} &= \Delta_{ij} (\mathbf{I} - \Lambda) \mathbf{T}' + \mathbf{T} (\mathbf{I} - \Lambda) \Delta'_{ij} \\ \frac{\partial \mathcal{P}}{\partial \lambda_i} &= 2(\lambda_i - \bar{\lambda}) \end{aligned}$$

### 495 Second derivatives

$$\begin{aligned} \frac{\partial^2 \Sigma_A}{\partial t_{ij} \partial \lambda_k} &= \Delta_{ij} \Delta_{kk} \mathbf{T}' + \mathbf{T} \Delta_{kk} \Delta'_{ij} & \frac{\partial^2 \Sigma_A}{\partial t_{ij} \partial t_{kl}} &= \Delta_{ij} \Lambda \Delta'_{kl} + \Delta_{kl} \Lambda \Delta'_{ij} \\ \frac{\partial^2 \Sigma_E}{\partial t_{ij} \partial \lambda_k} &= -(\Delta_{ij} \Delta_{kk} \mathbf{T}' + \mathbf{T} \Delta_{kk} \Delta'_{ij}) & \frac{\partial^2 \Sigma_E}{\partial t_{ij} \partial t_{kl}} &= \Delta_{ij} (\mathbf{I} - \Lambda) \Delta'_{kl} + \Delta_{kl} (\mathbf{I} - \Lambda) \Delta'_{ij} \\ \frac{\partial^2 \mathcal{P}}{\partial \lambda_i \partial \lambda_j} &= 2 \left( \delta_{ij} - \frac{1}{q} \right) & \text{with } \delta_{ij} &= \begin{cases} 1 & \text{for } i = j \\ 0 & \text{for } i \neq j \end{cases} \end{aligned}$$

## References

- 496
- 497 AMEMIYA, Y., 1985 What should be done when an estimated between-group covariance matrix  
498 is not nonnegative definite ? Amer. Stat. **39**: 112–117.
- 499 ANDERSON, B. M., T. W. ANDERSON, and I. OLKIN, 1986 Maximum likelihood estimators and  
500 likelihood ratio criteria in multivariate components of variance. Ann. Stat. **14**: 405–417.
- 501 ANDERSON, T. W., 1984 *An Introduction to Multivariate Statistical Analysis*. Wiley, New York,  
502 2nd edition.
- 503 BHARGAVA, A. K., and D. DISCH, 1982 Exact probabilities of obtaining estimated non-positive  
504 definite between-group covariance matrices. J. Stat. Comp. Simul. **15**: 27–32.
- 505 BICKEL, P. J., and E. LEVINA, 2008 Regularized estimation of large covariance matrices. Ann.  
506 Stat. **36**: 199–227.
- 507 BICKEL, P. J., and B. LI, 2006 Regularization in statistics. Test **15**: 271–303.
- 508 BÖHM, H., 2008 *Shrinkage methods for multivariate spectral analysis*. Ph.D. thesis, Catholic  
509 University Louvain, Belgium.
- 510 BONDARI, K., R. L. WILLHAM, and A. E. FREEMAN, 1978 Estimates of direct and maternal genetic  
511 correlations for pupa weight and family size of *Tribolium*. J. Anim. Sci. **47**: 358–365.
- 512 DANIELS, M. J., and R. E. KASS, 2001 Shrinkage estimators for covariance matrices. Biometrics  
513 **57**: 1173–1184.
- 514 DEMPSTER, A. P., 1972 Covariance selection. Biometrics **28**: 157–175.
- 515 DEY, D., and C. SRINIVASAN, 1985 Estimation of a covariance matrix under Stein’s loss. Ann.  
516 Stat. **13**: 1581–1591.
- 517 FOSTER, S. D., A. P. VERBYLA, and W. S. PITCHFORD, 2009 Estimation, prediction and inference  
518 for the LASSO random effects model. Austr. New Zeal. J. Stat. **51**: 43–61.
- 519 FRANK, I. E., and J. H. FRIEDMAN, 1993 A statistical view of some chemometrics regression  
520 tools. Technometrics **35**: 109–135.
- 521 FRIEDMAN, J., T. HASTIE, and R. TIBSHIRANI, 2008 Sparse inverse covariance estimation with the  
522 graphical lasso. Biostat **9**: 432–441.
- 523 GREEN, P. J., 1998 Penalized likelihood. In *Encyclopedia of Statistical Sciences*, volume 2. John  
524 Wiley & Sons, 578–586.
- 525 HAFF, L. R., 1980 Empirical Bayes estimation of the multivariate normal covariance matrix.  
526 Ann. Stat. **8**: 586–597.
- 527 HARVILLE, D. A., 1977 Maximum likelihood approaches to variance component estimation and  
528 related problems. J. Amer. Stat. Ass. **72**: 320–338.
- 529 HASTIE, T., R. TIBSHIRANI, and J. FRIEDMAN, 2001 *The Elements of Statistical Learning*. Springer  
530 Series in Statistics. Springer Verlag, New York, NY, USA.

- 531 HAYES, J. F., and W. G. HILL, 1981 Modifications of estimates of parameters in the construction  
532 of genetic selection indices ('bending'). *Biometrics* **37**: 483–493.
- 533 HILL, W. G., and R. THOMPSON, 1978 Probabilities of non-positive definite between-group or  
534 genetic covariance matrices. *Biometrics* **34**: 429–439.
- 535 HOERL, A. E., and R. W. KENNARD, 1970 Ridge regression: applications to nonorthogonal prob-  
536 lems. *Technometrics* **12**: 69–82.
- 537 HOFFMANN, K., 2000 Stein estimation – a review. *Statistical Papers* **41**: 127–158.
- 538 HUANG, J. Z., N. LIU, M. POURAHMADI, and L. LIU, 2006 Covariance matrix selection and esti-  
539 mation via penalised normal likelihood. *Biometrika* **93**: 85–98.
- 540 JAMES, W., and C. STEIN, 1961 Estimation with quadratic loss. In *Proc. Fourth Berkeley Symp.*  
541 *Math. Stat. Prob.*, volume 1. 361–379.
- 542 KIRKPATRICK, M., 2009 Patterns of quantitative genetic variation in multiple dimensions. *Ge-*  
543 *netica* **136**: 271–284.
- 544 KIRKPATRICK, M., D. LOFSVOLD, and M. BULMER, 1990 Analysis of the inheritance, selection and  
545 evolution of growth trajectories. *Genetics* **124**: 979–993.
- 546 KIRKPATRICK, M., and K. MEYER, 2004 Direct estimation of genetic principal components: Sim-  
547 plified analysis of complex phenotypes. *Genetics* **168**: 2295–2306.
- 548 KLOTZ, J., and J. PUTTER, 1969 Maximum likelihood estimation of multivariate covariance com-  
549 ponents for the balanced one-way layout. *Ann. Math. Stat.* **40**: 1100–1105.
- 550 KRUUK, L. E. B., J. SLATE, and A. J. WILSON, 2008 New answers for old questions: The evolu-  
551 tionary quantitative genetics of wild animal populations. *Ann. Rev. Ecol. Evol. System.* **39**:  
552 525–548.
- 553 KUBOKAWA, T., 1999 Shrinkage and modification techniques in estimation of variance and the  
554 related problems: A review. *Comm. Stat. - Theo. Meth.* **28**: 613–650.
- 555 KUBOKAWA, T., and M. T. TSAI, 2006 Estimation of covariance matrices in fixed and mixed  
556 effects linear models. *J. Multiv. Anal.* **97**: 2242–2261.
- 557 LAWLEY, D. N., 1956 Tests of significance for the latent roots of covariance and correlation  
558 matrices. *Biometrika* **43**: 128–136.
- 559 LEDOIT, O., and M. WOLF, 2004 A well-conditioned estimator for large-dimensional covariance  
560 matrices. *J. Multiv. Anal.* **88**: 365–411.
- 561 LEVINA, E., A. J. ROTHMAN, and J. ZHU, 2008 Sparse estimation of large covariance matrices via  
562 a nested Lasso penalty. *Ann. Appl. Stat.* **2**: 245–263.
- 563 LIN, S. P., and M. D. PERLMAN, 1985 A Monte Carlo comparison of four estimators of a covari-  
564 ance matrix. In P. R. Krishnaish, editor, *Multivariate Analysis*, volume 6. North-Holland,  
565 Amsterdam, 411–428.
- 566 LOH, W. L., 1991 Estimating covariance matrices. *Ann. Stat.* **19**: 283–296.

- 567 MATHEW, T., A. NIYOGI, and B. K. SINHA, 1994 Improved nonnegative estimation of variance  
568 components in balanced multivariate mixed models. *J. Multiv. Anal.* **51**: 83–101.
- 569 MENG, X. L., 2008 Who cares if it is a white cat or a black cat? Discussion: “One-step sparse  
570 estimates in nonconcave penalized likelihood models” [Ann. Statist. **36** (2008), 1509–1533]  
571 by H. Zou and R. Li. *Ann. Stat.* **36**: 1542–1552.
- 572 MEYER, K., 2009 Factor-analytic models for genotype x environment type problems and struc-  
573 tured covariance matrices. *Genet. Select. Evol.* **41**: 21.
- 574 MEYER, K., and W. G. HILL, 1983 A note on the effects of sampling errors on the accuracy of  
575 genetic selection indices. *Z. Tierz. Zücht. Biol.* **100**: 27–32.
- 576 MEYER, K., and M. KIRKPATRICK, 2008 Perils of parsimony: Properties of reduced rank estimates  
577 of genetic covariances. *Genetics* **108**: 1153–1166.
- 578 ODELL, P. L., and A. H. FEIVESON, 1966 A numerical procedure to generate a sample covariance  
579 matrix. *J. Amer. Stat. Ass.* **61**: 199–203.
- 580 PINHEIRO, J. C., and D. M. BATES, 1996 Unconstrained parameterizations for variance-  
581 covariance matrices. *Stat. Comp.* **6**: 289–296.
- 582 POURAHMADI, M., 1999 Joint mean-covariance models with applications to longitudinal data :  
583 Unconstrained parameterisation. *Biometrika* **86**: 677–690.
- 584 REVERTER, A., D. J. JOHNSTON, H.-U. GRASER, M. L. WOLCOTT, and W. H. UPTON, 2000 Ge-  
585 netic analyses of live-animal ultrasound and abattoir carcass traits in Australian Angus and  
586 Hereford cattle. *J. Anim. Sci.* **78**: 1786–1795.
- 587 ROTHMAN, A. J., P. J. BICKEL, E. LEVINA, and J. ZHU, 2008 Sparse permutation invariant covari-  
588 ance estimation. *Electron. J. Statist.* **2**: 494–515.
- 589 RUPPERT, D., M. P. WAND, and R. J. CARROLL, 2003 *Semiparametric Regression*. Cambridge  
590 University Press, New York.
- 591 SANCETTA, A., 2008 Sample covariance shrinkage for high dimensional dependent data. *J. Mul-  
592 tiv. Anal.* **99**: 949 – 967.
- 593 SCHÄFER, J., and K. STRIMMER, 2005 A shrinkage approach to large-scale covariance matrix  
594 estimation and implications for functional genomics. *Stat. Appl. Genet. Mol. Biol.* **4**: 32.
- 595 SRIVASTAVA, M. S., and T. KUBOKAWA, 1999 Improved non-negative estimation of multivariate  
596 components of variance. *Ann. Stat.* **27**: 2008–2032.
- 597 STEIN, C., 1975 Estimation of a covariance matrix. In *Reitz lecture. 39th Annual Meeting of the  
598 Institute of Mathematical Statistics*. Atlanta.
- 599 THOMPSON, R., 1976 The estimation of maternal genetic variance. *Biometrics* **32**: 903–917.
- 600 THOMPSON, R., S. BROTHERSTONE, and I. M. S. WHITE, 2005 Estimation of quantitative genetic  
601 parameters. *Phil. Trans. R. Soc. B* **360**: 1469–1477.
- 602 TIBSHIRANI, R., 1996 Regression shrinkage and selection via the lasso. *J. Roy. Stat. Soc. B* **58**:  
603 267–288.

- 604 TUTZ, G., and J. ULBRICHT, 2009 Penalized regression with correlation-based penalty. *Stat.*  
605 *Comput.* **19**: 239–253.
- 606 WARTON, D. I., 2008 Penalized normal likelihood and ridge regularization of correlation and  
607 covariance matrices. *J. Amer. Stat. Ass.* **103**: 340–349.
- 608 YAP, J. S., J. FAN, and R. WU, 2009 Nonparametric modeling of longitudinal covariance structure  
609 in functional mapping of quantitative trait loci. *Biometrics* Published on-line 3/3/2009.
- 610 YE, R. D., and S. G. WANG, 2009 Improved estimation of the covariance matrix under Stein's  
611 loss. *Stat. Prob. Lett.* **79**: 715 – 721.
- 612 ZOU, H., and T. HASTIE, 2005 Regularization and variable selection via the elastic net. *J. Roy.*  
613 *Stat. Soc. B* **67**: 301–320.

TABLE 1. – Reduction in average loss (PRIAL, in %), for estimates of the genetic ( $\Sigma_G$ ), error ( $\Sigma_E$ ) and phenotypic ( $\Sigma_P$ ) covariance matrix together with mean entropy loss ( $\times 100$ ) in unpenalized REML estimates of  $\Sigma_G$  ( $\bar{L}_1(\hat{\Sigma}_G^0, \Sigma_G)$ ) and the proportion of replicates (W, in %) for which penalized estimation increased the loss in  $\Sigma_G$ , for different constellations (A, ..., K) of population values ( $\lambda_i$ : canonical heritabilities) and penalties on the original (ORG) or logarithmic (LOG) scale; balanced paternal half-sib design with 500, 200 or 100 sires and 10 progeny per sire.

| Scale     |                                         | A    | B    | C    | D    | E    | F    | G    | H    | I    | J    | K    |
|-----------|-----------------------------------------|------|------|------|------|------|------|------|------|------|------|------|
|           | $\lambda_1$                             | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | 0.20 | 0.30 | 0.60 | 0.50 | 0.90 |
|           | $\lambda_2$                             | 0.40 | 0.45 | 0.50 | 0.55 | 0.30 | 0.50 | 0.20 | 0.25 | 0.10 | 0.20 | 0.30 |
|           | $\lambda_3$                             | 0.40 | 0.40 | 0.40 | 0.40 | 0.30 | 0.30 | 0.20 | 0.20 | 0.10 | 0.15 | 0.10 |
|           | $\lambda_4$                             | 0.40 | 0.35 | 0.30 | 0.25 | 0.30 | 0.20 | 0.20 | 0.15 | 0.10 | 0.10 | 0.10 |
|           | $\lambda_5$                             | 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   | 10   | 10   | 14   | 12   | 15   | 29   | 28   | 89   | 53   | 49   |
| ORG       | $\Sigma_G$                              | 79   | 47   | 20   | 21   | 16   | 13   | 86   | 40   | 21   | 23   | 11   |
|           | $\Sigma_E$                              | 83   | 51   | 30   | 28   | 40   | 35   | 72   | 36   | 19   | 16   | 40   |
|           | $\Sigma_P$                              | 8    | 5    | 2    | 2    | 0    | 0    | 2    | 1    | 1    | 0    | 0    |
|           | W                                       | 0    | 1    | 30   | 42   | 3    | 30   | 0    | 19   | 5    | 26   | 13   |
| LOG       | $\Sigma_G$                              | 91   | 45   | 15   | 16   | 40   | 25   | 93   | 33   | 72   | 36   | 60   |
|           | $\Sigma_E$                              | 86   | 52   | 26   | 12   | 38   | 7    | 77   | 35   | 19   | 17   | 18   |
|           | $\Sigma_P$                              | 8    | 5    | 2    | 1    | 1    | 0    | 2    | 1    | 0    | 0    | 0    |
|           | W                                       | 0    | 2    | 42   | 47   | 1    | 43   | 0    | 35   | 1    | 46   | 7    |
| 200 sires |                                         |      |      |      |      |      |      |      |      |      |      |      |
|           | $\bar{L}_1(\hat{\Sigma}_G^0, \Sigma_G)$ | 31   | 27   | 27   | 39   | 32   | 41   | 114  | 79   | 291  | 102  | 144  |
| ORG       | $\Sigma_G$                              | 91   | 68   | 36   | 31   | 38   | 20   | 94   | 62   | 26   | 24   | 9    |
|           | $\Sigma_E$                              | 86   | 70   | 50   | 50   | 50   | -35  | 76   | 53   | 31   | 25   | -18  |
|           | $\Sigma_P$                              | 8    | 6    | 4    | 2    | 0    | -1   | 2    | 2    | 0    | 0    | -1   |
|           | W                                       | 0    | 0    | 16   | 48   | 1    | 30   | 0    | 4    | 13   | 24   | 22   |
| LOG       | $\Sigma_G$                              | 91   | 67   | 30   | 12   | 56   | 32   | 95   | 57   | 83   | 30   | 72   |
|           | $\Sigma_E$                              | 87   | 70   | 49   | 34   | 56   | -17  | 77   | 54   | 31   | 30   | -14  |
|           | $\Sigma_P$                              | 8    | 6    | 4    | 2    | 1    | 0    | 2    | 2    | 1    | 1    | 0    |
|           | W                                       | 0    | 0    | 30   | 61   | 1    | 46   | 0    | 13   | 1    | 38   | 7    |
| 100 sires |                                         |      |      |      |      |      |      |      |      |      |      |      |
|           | $\bar{L}_1(\hat{\Sigma}_G^0, \Sigma_G)$ | 72   | 60   | 60   | 68   | 83   | 75   | 466  | 146  | 473  | 156  | 218  |
| ORG       | $\Sigma_G$                              | 93   | 80   | 54   | 25   | 65   | 22   | 96   | 69   | 29   | 26   | 6    |
|           | $\Sigma_E$                              | 88   | 80   | 69   | 64   | 42   | -198 | 76   | 63   | 45   | 41   | -104 |
|           | $\Sigma_P$                              | 8    | 7    | 5    | 3    | 1    | -1   | 2    | 2    | 1    | 1    | -1   |
|           | W                                       | 0    | 0    | 5    | 47   | 0    | 32   | 0    | 1    | 14   | 19   | 34   |
| LOG       | $\Sigma_G$                              | 93   | 79   | 49   | 2    | 72   | 31   | 98   | 68   | 83   | 25   | 68   |
|           | $\Sigma_E$                              | 88   | 81   | 70   | 57   | 53   | -90  | 78   | 65   | 46   | 44   | -70  |
|           | $\Sigma_P$                              | 8    | 7    | 5    | 3    | 2    | 0    | 3    | 2    | 1    | 1    | -1   |
|           | W                                       | 0    | 0    | 14   | 64   | 0    | 42   | 0    | 4    | 1    | 35   | 6    |

TABLE 2. – Reduction in average loss (PRIAL, in %), for estimates of the genetic ( $\Sigma_G$ ), error ( $\Sigma_E$ ) and phenotypic ( $\Sigma_P$ ) covariance matrix, together with mean entropy loss ( $\times 100$ ) in unpenalized REML estimates of  $\hat{\Sigma}_G$  ( $\bar{L}_1(\hat{\Sigma}_G^0, \Sigma_G)$ ) and the proportion of replicates (W, in %) for which penalized estimation increased the loss in  $\Sigma_G$ , for different constellations (A,...,K) of population values (see Table 1) and penalties on the original (ORG) or logarithmic (LOG) scale; Bondari's design with 125 or 50 families

| Scale        |                                         | A  | B  | C  | D   | E   | F   | G   | H   | I   | J   | K   |
|--------------|-----------------------------------------|----|----|----|-----|-----|-----|-----|-----|-----|-----|-----|
| 125 families |                                         |    |    |    |     |     |     |     |     |     |     |     |
|              | $\bar{L}_1(\hat{\Sigma}_G^0, \Sigma_G)$ | 23 | 22 | 24 | 42  | 30  | 49  | 158 | 92  | 366 | 120 | 183 |
| ORG          | $\Sigma_G$                              | 86 | 58 | 32 | 30  | 41  | 35  | 94  | 62  | 36  | 26  | 35  |
|              | $\Sigma_E$                              | 72 | 53 | 33 | 24  | 29  | 35  | 63  | 45  | 18  | 20  | 31  |
|              | $\Sigma_P$                              | 8  | 6  | 3  | 2   | 1   | 1   | 2   | 2   | 0   | 1   | 1   |
|              | W                                       | 0  | 1  | 31 | 51  | 4   | 36  | 0   | 8   | 13  | 26  | 16  |
| LOG          | $\Sigma_G$                              | 88 | 57 | 24 | 25  | 57  | 37  | 96  | 57  | 84  | 25  | 75  |
|              | $\Sigma_E$                              | 74 | 54 | 30 | 12  | 28  | 9   | 65  | 46  | 18  | 21  | 8   |
|              | $\Sigma_P$                              | 8  | 6  | 3  | 1   | 2   | 1   | 3   | 2   | 1   | 1   | 1   |
|              | W                                       | 0  | 3  | 42 | 54  | 3   | 44  | 0   | 17  | 2   | 41  | 7   |
| 50 families  |                                         |    |    |    |     |     |     |     |     |     |     |     |
|              | $\bar{L}_1(\hat{\Sigma}_G^0, \Sigma_G)$ | 90 | 67 | 70 | 78  | 118 | 95  | 704 | 187 | 618 | 204 | 289 |
| ORG          | $\Sigma_G$                              | 91 | 78 | 54 | 12  | 72  | 28  | 96  | 70  | 43  | 28  | 32  |
|              | $\Sigma_E$                              | 75 | 67 | 53 | 46  | 55  | -25 | 64  | 55  | 32  | 35  | -49 |
|              | $\Sigma_P$                              | 9  | 7  | 5  | 3   | 2   | 1   | 2   | 2   | 1   | 1   | 1   |
|              | W                                       | 0  | 0  | 17 | 54  | 2   | 37  | 0   | 2   | 13  | 23  | 21  |
| LOG          | $\Sigma_G$                              | 92 | 78 | 45 | -13 | 77  | 25  | 98  | 70  | 81  | 11  | 65  |
|              | $\Sigma_E$                              | 76 | 68 | 53 | 36  | 52  | -13 | 65  | 56  | 28  | 34  | -14 |
|              | $\Sigma_P$                              | 8  | 7  | 5  | 3   | 4   | 2   | 3   | 2   | 1   | 1   | 1   |
|              | W                                       | 0  | 1  | 29 | 66  | 3   | 44  | 0   | 4   | 2   | 46  | 9   |

FIGURE 1. – Mean estimates of canonical heritabilities, as deviation from population values (in %) together with plus/minus one empirical standard deviations (vertical bars) for standard REML analyses (gray), and penalties on eigenvalues on the original (red) and logarithmic (blue) scale, for cases A, B and C and a paternal half-sib design with 500 or 200 sires.

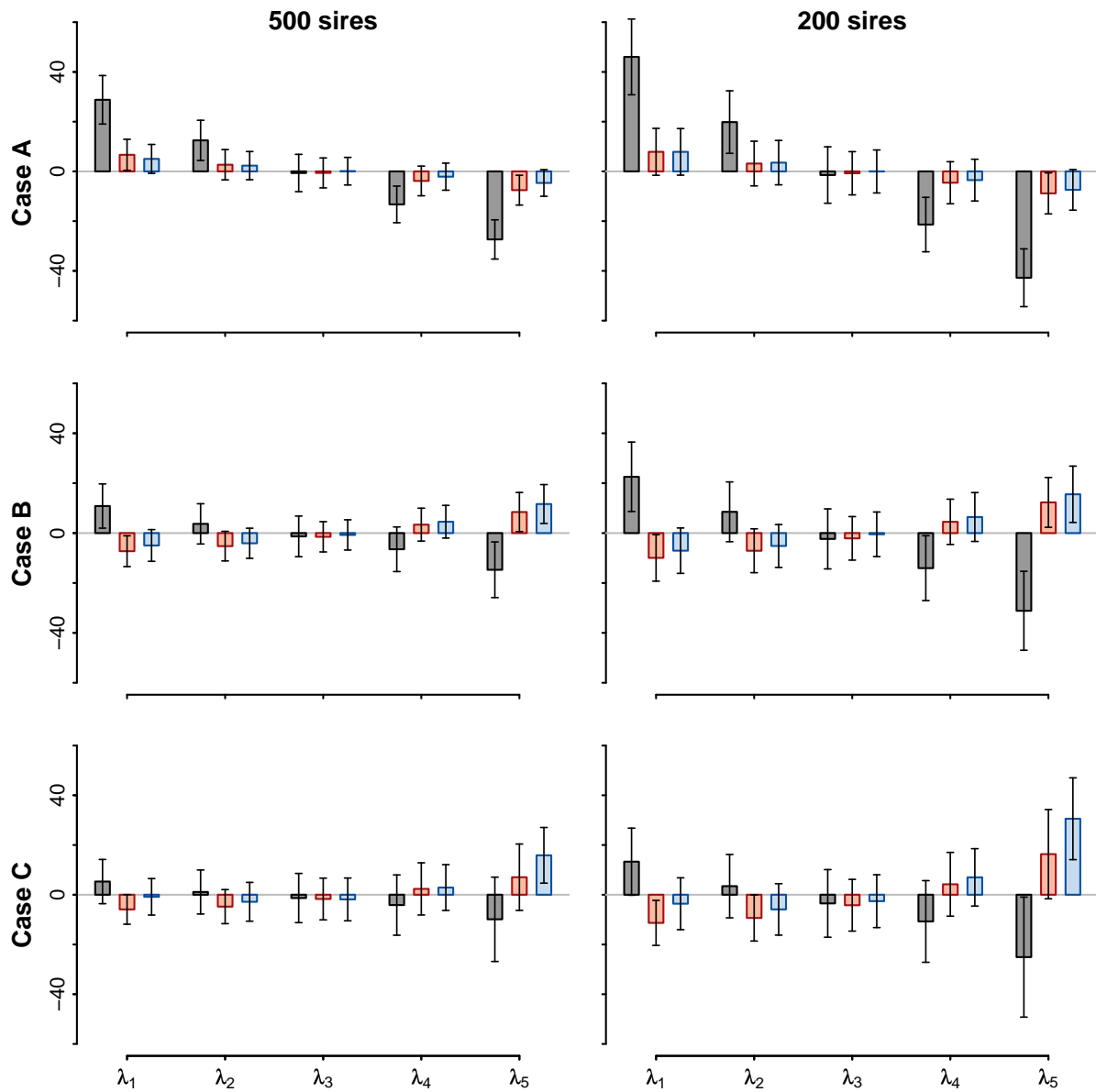


FIGURE 2. – Mean estimates of canonical heritabilities, as deviation from population values (in %) together with plus/minus one empirical standard deviations (vertical bars) for standard REML analyses (gray), and penalties on eigenvalues on the original (red) and logarithmic (blue) scale, for Bondari's design with 125 families.

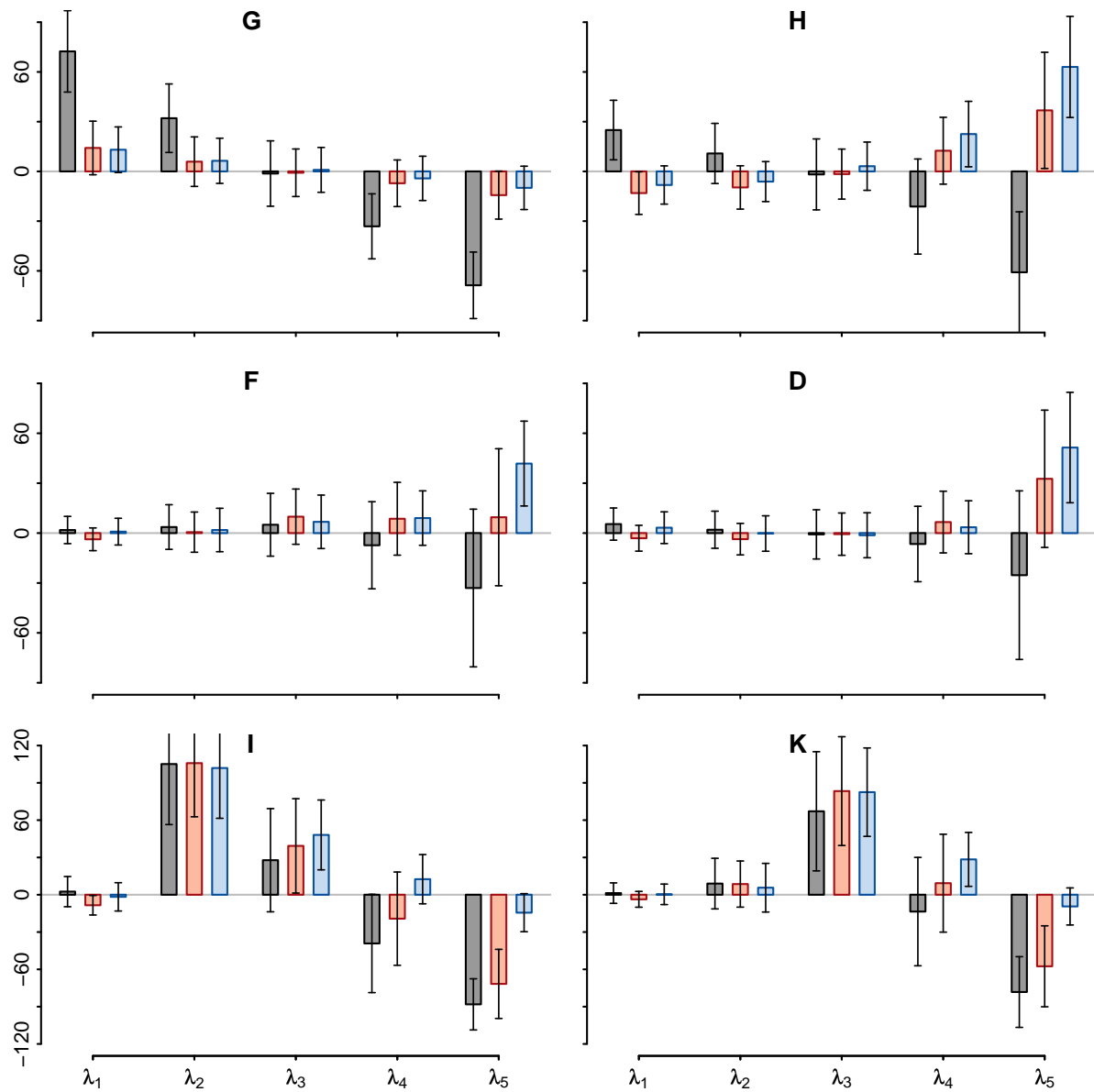


FIGURE 3. – Estimates of genetic parameters ( $h_i^2$ : heritability for trait  $i$ ,  $r_{ij}$ : correlation between traits  $i$  and  $j$ ) for beef cattle example from standard (black symbols) and penalized (light symbols) analyses (• heritability, ♦ genetic correlation, ▼ environmental correlation; vertical bars show range of one standard deviation either side of estimates from standard analyses)

