Performance of penalized maximum likelihood in estimation of genetic covariances matrices

Karin Meyer

Animal Genetics and Breeding Unit[1], University of New England, Armidale NSW 2351, Australia

1 Introduction

Estimation of genetic parameters, i.e. the partitioning of phenotypic variation between individuals into (co)variances due to genetic effects and other sources, is one of the basic tasks in quantitative genetics. Increasingly, recording schemes in livestock improvement programmes are becoming more sophisticated and detailed, along with a trend for breeding objectives to involve more and more components. This results in a continual growth in the number of traits of interest, and, in turn, necessitates increasingly complex, multivariate analyses considering more than just a few traits simultaneously.

Advances in modelling, improvements of computational algorithms and of the corresponding software for estimation, paired with the capabilities of modern day computer hardware available have brought us to a point where large-scale analyses comprising numerous traits and records on tens of thousands of animals are within the realms of reality. For example, Tyrisevä et al. (2011) recently demonstrated that simultaneous estimation of the complete genetic covariance matrix required by Interbull, the international evaluation service for dairy bulls, for its multiple-trait across country evaluation is feasible, presenting multivariate analyses involving 25 traits with more than 100 000 sires and up to 325 parameters to be estimated. However, comparatively little attention has been paid to the problems associated with sampling variation that are inherent in multivariate analyses, and which increase dramatically with the number of traits and the number of parameters to be estimated.

It has long been known that the eigenvalues of estimated covariance matrices are over-dispersed, i.e. that the largest sample eigenvalues are systematically biased upwards and the smallest values are biased downwards while their mean is expected to be unbiased (Lawley, 1956). Moreover, a large proportion of the sampling variances of estimates of individual

[1] a joint venture with Industries & Investment NSW

covariances can be attributed to this excess variation (Ledoit and Wolf, 2004). The effects of this phenomenon are the more pronounced the narrower the ratio of the matrix dimension to the number of observations and the more similar the population eigenvalues are. Hill and Thompson (1978) showed in an early simulation study how this affected estimates of genetic covariance matrices and that it resulted in high probabilities of obtaining non-positive definite estimates.

While modern, maximum likelihood (ML) based methods of estimation make efficient use of all the data and readily allow estimates of covariance matrices to be constrained to the parameter space (Harville, 1977), the problems of sampling variation remain. Even multivariate analyses based on relatively large data sets are thus likely to yield imprecise estimates, the more so the more traits are considered. At the other end of the spectrum, we have numerous scenarios where the numbers of records are invariably limited. This includes data for new traits of interest or traits which are difficult or expensive to measure but which may have substantial impact on selection decisions in livestock improvement programmes. A typical example for such data are carcass characteristics of meat producing animals, which are never recorded directly for parents of the next generation. Similarly, evolutionary biologist concerned with quantitative genetics of natural populations are usually restricted to rather small samples.

Hence, any avenue to 'improve' estimates, i.e. to obtain estimates which are on average closer to the population values, is of considerable interest and should be given serious consideration. To begin with, we have accumulated a substantial body of knowledge about genetic parameters for various traits. However, typically this is completely ignored. While the Bayesian paradigm directly provides the means to incorporate such prior information, analyses concerned with the estimation of covariance components more often than not assume flat or uninformative priors (Thompson et al., 2005). Clearly, there is considerable scope for using this information more advantageously, especially for small samples arising in evolutionary studies of natural or laboratory populations (Kirkpatrick et al., 2011).

Secondly, multivariate covariance matrices can often be modelled parsimoniously by imposing some structure. This decreases sampling variation by reducing the number of parameters to be estimated. Common examples are factor-analytic and reduced rank models or treating covariance matrices as 'separable', i.e. as the direct product of two or more smaller matrices; see Meyer (2009) for a detailed review. Finally, statistical techniques are available – often referred to as regularization methods – which substantially reduce sampling variance, albeit at the expense of introducing some bias, and thus yield ‘better’ estimates. Interest in regularized estimation for multivariate analyses and the trade-off between sampling variance and bias dates back to the Seventies and earlier, stimulated in particular by the work of Stein (e.g. James and Stein, 1961; Stein, 1975). Recently, there has been a resurgence in attention for applications involving estimation in very high-dimensional settings, in particular for genomic data (e.g. Huang et al., 2006; Warton, 2008; Yap et al., 2009; Witten and Tibshirani, 2009).

In spite of well established literature on regularized estimation of covariance matrices, there has been comparatively little interest in this approach in the context of estimating genetic parameters in quantitative genetics. An early proposal, due to Hayes and Hill (1981), has been to shrink the canonical eigenvalues in a one-way analysis of variance towards their mean and thus to reduce sampling variation. This yielded an estimate of the genetic covariance matrix which was a weighted combination of the standard (i.e. not regularized) estimate and the phenotypic covariance matrix multiplied by the mean eigenvalue. The authors thus described their method as ‘bending’ the genetic towards the phenotypic covariance matrix. Hayes and Hill (1981) presented a simulation study demonstrating that ‘bending’ could substantially increase the achieved response to selection based on an index derived using the modified estimates. However, other than in forcing covariance matrices obtained by pooling estimates from multiple sources to be positive definite, their method has found little application, as there were no clear guidelines on how to choose the amount of shrinkage to be applied.

Recently, Meyer and Kirkpatrick (2010) proposed to employ penalized restricted maximum likelihood (REML) to obtain ‘better’ estimates of genetic covariance matrices, and showed that imposing a penalty proportional to the variance among the canonical eigenvalues acted analogously to ‘bending’. They demonstrated by simulation that this resulted in estimates of genetic parameters from multivariate analyses which had greatly reduced sampling and mean square errors, and, moreover, that this held not only for the paternal half-sib design considered by Hayes and Hill (1981), but equally for animal model analyses with a complicated pedigree structure and many different types of covariances between relatives.

This paper extends the approach of Meyer and Kirkpatrick (2010) to different types of penalties and, in an extensive simulation study, examines the performance of various strategies to determine the amount of penalization to be applied. To begin with, we briefly review the underlying statistical principles and outline a penalized maximum likelihood estimation scheme, presenting a number of suitable choices of penalties. This is followed by a simulation study to compare the efficacy of different types of penalty and schemes to estimate the tuning factor required, considering different numbers of traits and sample sizes. The paper concludes with a discussion and recommendations for practical applications.

2 Penalized maximum likelihood estimation

2.1 Improved estimation

The quality of a statistical estimator is generally quantified by some measure of the difference between the estimator and the true value, or *loss*. A widely used quantity is the mean square error. This is a quadratic loss, comprised of the sampling variance and the square of the bias in the estimator. We talk about improving an estimator when we are able to modify it in some way so that, on average, it is closer to the true value, i.e. has reduced loss. Usually this involves a trade-off between a reduction in sampling variance and additional bias.

For covariance matrices, commonly employed measures of divergence are the entropy (L_1) and quadratic (L_2) loss (James and Stein, 1961):

$$L_1\left(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}\right) = \text{tr}\left(\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}}\right) - \log\left|\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}}\right| - q \qquad \text{and} \qquad L_2\left(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}\right) = \text{tr}\left(\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}} - \mathbf{I}\right)^2 \tag{1}$$

where $\boldsymbol{\Sigma}$ and $\hat{\boldsymbol{\Sigma}}$ denote a covariance matrix of size $q \times q$ and its estimator, respectively, and q represents the number of traits.

A reduction in loss can often be achieved by regularizing estimators. In broad terms, *regularization* describes a scenario where estimation for somewhat ill-posed or overparameterized problems is improved through use of some form of additional information. Frequently the latter involves a penalty for the deviation from a desired outcome. For example, in modelling curves using splines a 'roughness penalty' is employed to place preference on simple, smooth functions (Green, 1998). Well known forms of regularization are ridge regression (Hoerl and Kennard, 1970) and the LASSO (Least absolute shrinkage and selection operator; Tibshirani, 1996, 2011). Whilst these methods were originally developed to encourage shrinkage of regression coefficients, corresponding applications for the estimation of high-dimensional covariance matrices have been described; see Meyer and Kirkpatrick (2010) for a review and references.

2.2 Penalizing the likelihood

Consider a simple 'animal model' for q traits, $\mathbf{y} = \mathbf{Xb} + \mathbf{Zg} + \mathbf{e}$ with $\mathbf{y}$, $\mathbf{b}$, $\mathbf{g}$ and $\mathbf{e}$ the vectors of observations, fixed effects, additive genetic and residual effects, respectively, and $\mathbf{X}$ and $\mathbf{Z}$ the corresponding incidence matrices. Let $\mathbf{\Sigma}_G$ and $\mathbf{\Sigma}_E$ denote the matrices of additive genetic and residual covariances among the q traits. This gives a vector of parameters to be estimated, $\boldsymbol{\theta}$, of length $q(q+1)$ comprising the distinct elements of $\mathbf{\Sigma}_G$ and $\mathbf{\Sigma}_E$. Further, let $\text{Var}\,(\mathbf{g}) = \mathbf{\Sigma}_G \otimes \mathbf{A} = \mathbf{G}$, where $\mathbf{A}$ is the numerator relationship matrix between individuals. Let $\mathbf{R}_k$ denote the sub-matrix of $\mathbf{\Sigma}_E$ corresponding to the traits recorded for the $k-$th individual. This gives $\text{Var}\,(\mathbf{e}) = \sum_k^+ \mathbf{R}_k = \mathbf{R}$, where '$\sum^+$' is the direct matrix sum. The phenotypic covariance matrix of the vector of observations is then $\text{Var}\,(\mathbf{y}) = \mathbf{ZGZ}' + \mathbf{R} = \mathbf{V}$, and the pertaining REML log likelihood is, apart from a constant,

$$\log \mathcal{L}(\boldsymbol{\theta}) = -\tfrac{1}{2}\left(\log|\mathbf{V}| + \log\left|\mathbf{X}_0'\mathbf{V}^{-1}\mathbf{X}_0\right| + (\mathbf{y} - \mathbf{Xb})'\,\mathbf{V}^{-1}\,(\mathbf{y} - \mathbf{Xb})\right) \tag{2}$$

for $\mathbf{X}_0$ a full-rank submatrix of $\mathbf{X}$ (e.g. Harville, 1977). Regularized estimates can be obtained by maximizing the *penalized* likelihood

$$\log \mathcal{L}_P(\boldsymbol{\theta}) = \log \mathcal{L}(\boldsymbol{\theta}) - \tfrac{1}{2}\,\psi\,\mathcal{P}(\boldsymbol{\theta}) \tag{3}$$

where the penalty $\mathcal{P}(\boldsymbol{\theta})$ is a selected function of the parameters, aimed at reducing loss in their estimates, and ψ is a tuning factor which specifies the relative emphasis to be given to the penalty compared to the usual, unpenalized estimator. For $\psi = 0$, this simplifies to the standard, unpenalized likelihood. Here, the factor of ½ in (Eq. 3) is for algebraic consistency and could be omitted.

A general way to select a penalty is to specify a prior distribution for the parameters to be estimated for a suitable choice of parameterisation. The penalty can then be obtained as minus the logarithmic value of the density of the prior. Hence, penalizing the likelihood provides a direct link to Bayesian estimation, with the tuning factor performing an analogous rôle to the degree of belief attached to the prior. Meng (2008) described penalized estimation as a way of "enjoying the Bayesian fruits without paying the B-club fee".

2.2.1 Penalties on eigenvalues

Recognition of the systematic upwards bias in the largest and downwards bias in the smallest eigenvalues of estimated covariance matrices early on has led to the development of various improved estimators which modify the eigenvalues in some fashion whilst retaining the corresponding eigenvectors. As the mean eigenvalue is expected to be unbiased, a specific proposal has been to regress all eigenvalues towards their mean in order to reduce their excessive spread. This is equivalent to assuming eigenvalues have a prior that is a Normal distribution.

As outlined above, Hayes and Hill (1981) proposed to apply this type of shrinkage to the canonical eigenvalues (λ_i), i.e. the eigenvalues of $\mathbf{\Sigma}_P^{-1}\mathbf{\Sigma}_G$, with $\mathbf{\Sigma}_P = \mathbf{\Sigma}_G + \mathbf{\Sigma}_E$ the phenotypic covariance matrix. The equivalent to bending in a (RE)ML framework can be obtained by placing a penalty proportional to the variance among the estimated canonical eigenvalues on the likelihood (Meyer and Kirkpatrick, 2010):

$$\mathcal{P}_\lambda \propto \text{tr}\left(\mathbf{\Lambda} - \bar{\lambda}\mathbf{I}\right)^2 \qquad \text{with} \quad \bar{\lambda} = \text{tr}\left(\mathbf{\Lambda}\right)/q \tag{4}$$

for $\mathbf{\Lambda} = \text{Diag}\left\{\hat{\lambda}_i\right\}$. The canonical decomposition gives $\mathbf{\Sigma}_G = \mathbf{T\Lambda T}'$ and the residual covariance matrix, $\mathbf{\Sigma}_E = \mathbf{T}(\mathbf{I} - \mathbf{\Lambda})\mathbf{T}'$, with $\mathbf{I}$ an identity matrix and $\mathbf{T}$ the matrix of eigenvectors of $\mathbf{\Sigma}_P^{-1}\mathbf{\Sigma}_G$ scaled by a matrix square root of $\mathbf{\Sigma}_P$. Hence, $\mathcal{P}_\lambda$ can be thought of as penalizing both $\mathbf{\Sigma}_G$ and $\mathbf{\Sigma}_E$ at the same time.

A related penalty, $\mathcal{P}_\lambda^\ell$, is obtained by penalizing the eigenvalues on the logarithmic scale, i.e. defining $\mathbf{\Lambda} = \text{Diag}\left\{\log(\hat{\lambda}_i)\right\}$. This is analogous to the log eigenvalue posterior mean shrinkage estimator considered by Daniels and Kass (2001) for a single matrix. Placing a quadratic penalty on $(1 - \lambda_i)$ is equivalent to penalizing λ_i, but this does not hold on the log scale. Hence a third penalty is

$$\mathcal{P}_\lambda^{\ell 2} \propto \text{tr}\left(\mathbf{\Lambda}_1 - \bar{\lambda}_1\mathbf{I}\right)^2 + \text{tr}\left(\mathbf{\Lambda}_2 - \bar{\lambda}_2\mathbf{I}\right)^2 \tag{5}$$

for $\mathbf{\Lambda}_1 = \text{Diag}\left\{\log(\hat{\lambda}_i)\right\}$ and $\mathbf{\Lambda}_2 = \text{Diag}\left\{\log(1 - \hat{\lambda}_i)\right\}$, with $\bar{\lambda}_i = \text{tr}\left(\mathbf{\Lambda}_i\right)/q$.

For $\mathbf{\Sigma}_G$ positive semi-definite, the canonical eigenvalues lie in the interval $[0, 1]$. Hence a natural alternative to a Normal prior is the Beta distribution, which is defined on this domain and is thus frequently used as prior for binomial proportions in a Bayesian setting. It has two shape parameters, $\alpha > 0$ and $\beta > 0$, and probability density function

$$p\left(x\right) = \frac{\Gamma\left(\alpha + \beta\right)}{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)}\, x^{\alpha-1}\left(1 - x\right)^{\beta-1} \tag{6}$$

with $\Gamma(\cdot)$ denoting the Gamma function, and mean $\alpha/(\alpha + \beta)$. Hence, for $\alpha = \beta$ the function $p(x)$ is symmetric with mean at 0.5. For $\alpha > 1$ and $\beta > 1$ it is uni-modal with probability mass increasingly concentrated at the mean as α and β increase. Figure 1 (a) illustrates this for $\alpha = \beta = 2, \ldots, 5$. A restricted domain $[x_1, x_2]$ (with x_1 and x_2 the lower and upper limits for x) can be taken into account by expanding $p(x)$ to a four parameter function, replacing $x^{\alpha-1}$ and $(1-x)^{\beta-1}$ in (Eq. 6) with $(x-x_1)^{\alpha-1}$ and $(x_2-x)^{\beta-1}$, respectively, and scaling by $(x_2-x_1)^{-(\alpha+\beta-1)}$ (Evans et al., 2000). Alternatively, this can be achieved by replacing x in (Eq. 6) with $x^\star = (x-x_1)/(x_2-x_1)$.

The distribution of estimates of the canonical eigenvalues clearly depends on the population parameters and may well not cover the whole interval $[0,1]$. As we expect standard estimates of eigenvalues to be over-dispersed, a suitable, if somewhat inflated, estimate of the range may be given by the estimates of the extreme values from an unpenalized analysis, i.e. for $\psi = 0$, denoted henceforth by a superscript of 0. Assuming eigenvalues are numbered in descending order of magnitude, this gives $\hat\lambda_1^0$ and $\hat\lambda_q^0$ for the upper and lower bound, respectively. To utilise the standard form of the Beta distribution, as given in (Eq. 6), we then base the penalty on scaled values $\lambda_i^\star = (\hat\lambda_i - \hat\lambda_q^0)/(\hat\lambda_1^0 - \hat\lambda_q^0)$. For chosen values α and β, this gives penalty

$$\mathcal{P}^a_\beta \propto (\alpha-1)\log(\lambda_i^\star) + (\beta-1)\log(1-\lambda_i^\star) \tag{7}$$

A suitable choice might be $\alpha = \beta = 2, 3, \ldots$ which implies a symmetric distribution for λ_i^* with probability mass somewhat more spread out than a Normal distribution (*c.f.* Figure 1, (a))

Alternatively, we may try to obtain estimates of the scale parameters from the unpenalized estimates of the canonical eigenvalues. Using that the mean and variance of the standard Beta distribution are $\alpha/(\alpha+\beta)$ and $\alpha\beta(\alpha+\beta)^{-2}(\alpha+\beta+1)^{-1}$, respectively, gives method of moment estimators $\tilde\alpha = \bar\lambda\nu$ and $\tilde\beta = (1-\bar\lambda)\nu$, with $\nu = q\bar\lambda(1-\bar\lambda)/\sum_{i=1}^{q}(\hat\lambda_i^0 - \bar\lambda)^2) - 1$ (Evans et al., 2000) and $\bar\lambda$ the mean of the $\hat\lambda_i^0$. This may result in estimates of α and β with are less than unity, implying probability distributions that are U- or J-shaped with a high mass at the extremes. To counteract effects of over-dispersion of the $\hat\lambda_i^0$ and ensure a uni-modal Beta distribution, we thus choose to augment these values by a constant z, $\hat\alpha = \tilde\alpha + z$ and $\hat\beta = \tilde\beta + z$. Figure 1 (b) demonstrates the effect that a scale parameter less than unity has on the probability distribution and how adding a constant of z=1 yields a prior with more appropriate shape. This gives penalty

$$\mathcal{P}^b_\beta \propto (\hat\alpha-1)\log(\lambda_i) + (\hat\beta-1)\log(1-\lambda_i) \tag{8}$$

As above, we can combine estimates of the scale parameter with scaling to account for a range smaller than $[0,1]$ by replacing λ_i in (Eq. 8) with $\lambda_i^\star$, yielding penalty $\mathcal{P}^c_\beta$.

Penalties considered so far implied that estimated eigenvalues were samples from a distribution with common mean $\bar{\lambda}$. However, while quadratic penalties on eigenvalues or eigenvalues transformed to logarithmic scale have been found to be highly effective when the corresponding population values were similar, they have been reported to result in substantial over-shrinkage when the latter were spread apart (Daniels and Kass, 2001; Ledoit and Wolf, 2004; Meyer and Kirkpatrick, 2010). Hence, if population eigenvalues are markedly different, it may be advantageous to shrink towards individual targets. Ordering values sampled from a statistical distribution are according to size introduces a specific distribution. The $i-$th order statistic of a $q-$variate sample is the $i-$th smallest value. Assuming a uniform distribution, the order statistics on the unit interval have marginal Beta distributions with scale parameters $z+i$ and $z+q-i+1$ for $z=0$. Treating the scaled estimates of canonical eigenvalues as independent order statistics results in a penalty

$$\mathcal{P}_{\beta}^{d} \propto \sum_{i=1}^{q} (z+i-1)\log(\lambda_i^{\star}) + (z+q-i)\log(1-\lambda_i^{\star}) \qquad \text{for} \quad c=0 \tag{9}$$

Again we have allowed for a modifying constant z in (Eq. 9). For the distribution of order statistics this is z=0 . Figure 1 (c) shows the corresponding probability density functions for $q=5$ variables. As illustrated this results in rather different distributions for different variables. A value of $z>0$ causes individual distributions to be 'squashed' together, i.e. allows for a compromise between the assumption of a common mean for the $\lambda_i^{\star}$ and that of an even distribution over the unit interval. Figure 1 (d) demonstrates the effect of using z=1.

2.2.2 Penalties on matrix divergence

Motivated by the historical emphasis on the rôle of sample eigenvalues of covariance matrices, we have concentrated on penalties on these characteristics so far. A simple alternative is to consider a covariance matrix as a whole and its prior distribution, or to penalize the deviation from a specific target.

A standard assumption in Bayesian estimation of covariance matrices is that of an Inverse Wishart prior distribution, as, for observations with a multivariate Normal distribution, this is a conjugate prior. It has probability density function $p\left(\boldsymbol{\Sigma}|\boldsymbol{\Omega},\nu\right) \propto |\boldsymbol{\Sigma}|^{\frac{1}{2}(\nu+q+1)} \exp\left[-\frac{1}{2}\,\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Omega}\right)\right]$ (e.g Sorensen and Gianola, 2002), with $\boldsymbol{\Omega}$ denoting the scale parameter and ν the degree of belief we assign to the prior. Omitting terms not depending on $\boldsymbol{\Sigma}$ or $\boldsymbol{\Omega}$ and taking logarithms gives $(\nu+q+1)\log|\boldsymbol{\Sigma}| + \nu\,\mathrm{tr}\left(\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Omega}\right)$.

Corresponding to the penalties 'borrowing strength' from the phenotypic covariance matrix considered above, a penalty which regularizes $\hat{\mathbf{\Sigma}}_G$ by shrinking it towards $\mathbf{\Sigma}_P$ can be obtained by substituting the latter for the scale matrix $\mathbf{\Omega}$. Adopting an empirical Bayes approach, as suggested by Meyer et al. (2011), we replace $\mathbf{\Sigma}_P$ with its estimate from a standard, unpenalized (RE)ML analysis, $\hat{\mathbf{\Sigma}}_P^0$. Further, replacing ν with the tuning factor ψ, gives a penalty

$$\mathcal{P}_\Sigma \propto C \log |\hat{\mathbf{\Sigma}}_G| + \text{tr}\left(\hat{\mathbf{\Sigma}}_G^{-1}\hat{\mathbf{\Sigma}}_P^0\right) \tag{10}$$

with $C = \left(\psi + q + 1\right)/\psi$. If C is approximated with unity, $\mathcal{P}_\Sigma$ is proportional to the Kullback-Leibler divergence between $\hat{\mathbf{\Sigma}}_G$ and $\hat{\mathbf{\Sigma}}_P^0$, which is the entropy loss $L_1(\cdot)$ (Eq. 1) with $\mathbf{\Sigma}$ and $\hat{\mathbf{\Sigma}}$ exchanged (Levina et al., 2008). The relationship between $\mathcal{P}_\Sigma$ and $\mathcal{P}_\lambda$ can be seen by rewriting (Eq. 11) in terms of the canonical decomposition which gives $\mathcal{P}_\Sigma \propto C\Big(\log |\hat{\mathbf{\Lambda}}| + \log |\hat{\mathbf{T}}\hat{\mathbf{T}}'|\Big) + \text{tr}\left(\hat{\mathbf{\Lambda}}^{-1}\hat{\mathbf{T}}^{-1}\hat{\mathbf{\Sigma}}_P^0\hat{\mathbf{T}}^{-T}\right)$. Assuming that $\hat{\mathbf{\Sigma}}_P^0 \approx \hat{\mathbf{T}}\hat{\mathbf{T}}'$, i.e. that the estimate of transformation and phenotypic covariance matrix are largely unaffected by penalised estimation, gives $\mathcal{P}_\Sigma \propto C \log |\hat{\mathbf{\Lambda}}| + \text{tr}\left(\hat{\mathbf{\Lambda}}^{-1}\right) \propto \sum_i^q C \log\left(\hat{\lambda}_i\right) + \hat{\lambda}_i^{-1}$. This shows that $\mathcal{P}_\Sigma$ implies a substantial penalty on the smallest canonical eigenvalues. Analogous to penalty $\mathcal{P}_\lambda^{\ell 2}$, we may also consider to penalize both $\mathbf{\Sigma}_G$ and $\mathbf{\Sigma}_E$ using

$$\mathcal{P}_\Sigma^2 \propto C \log |\hat{\mathbf{\Sigma}}_G| + \text{tr}\left(\hat{\mathbf{\Sigma}}_G^{-1}\hat{\mathbf{\Sigma}}_P^0\right) + C \log |\hat{\mathbf{\Sigma}}_E| + \text{tr}\left(\hat{\mathbf{\Sigma}}_E^{-1}\hat{\mathbf{\Sigma}}_P^0\right) \tag{11}$$

Based on empirical evidence that estimates of genetic (r_G) and phenotypic (r_P) correlations are often similar, Cheverud (1988) proposed to substitute r_P for r_G if the data did not support accurate estimation of r_G. Adopting this suggestion, Meyer and Kirkpatrick (2009) demonstrated that estimating $\mathbf{\Sigma}_G$ and $\mathbf{\Sigma}_E$ or $\mathbf{\Sigma}_P$ by assuming a joint correlation structure resulted in highly parsimonious models and a dramatic reduction in mean square errors when the underlying assumptions were approximately true. Conversely, estimates could be substantially biased if they were not. A more flexible alternative is to penalize the divergence between estimates of the genetic ($\mathbf{R}_G$) and phenotypic correlation ($\mathbf{R}_P$) matrix, i.e. to shrink $\hat{\mathbf{R}}_G$ towards $\hat{\mathbf{R}}_P^0$. Analogous to (Eq. 11), this can be achieved using a penalty

$$\mathcal{P}_\rho \propto C \log |\hat{\mathbf{R}}_G| + \text{tr}\left(\hat{\mathbf{R}}_G^{-1}\hat{\mathbf{R}}_P^0\right) \tag{12}$$

or

$$\mathcal{P}_\rho^2 \propto C \log |\hat{\mathbf{R}}_G| + \text{tr}\left(\hat{\mathbf{R}}_G^{-1}\hat{\mathbf{R}}_P^0\right) + C \log |\hat{\mathbf{R}}_E| + \text{tr}\left(\hat{\mathbf{R}}_E^{-1}\hat{\mathbf{R}}_P^0\right) \tag{13}$$

More generally, this type of penalty can be used to shrink an estimated covariance matrix towards any chosen structure. This allows for a data-driven compromise between the assumed structure and an unstructured matrix. For instance, Chen (1979) presented an empirical Bayesian approach to estimate a covariance matrix shrinking towards a prior assumed to have a factor-analytic or compound symmetric structure. More recently, Schäfer and Strimmer (2005) considered shrinkage towards a number of target matrices with diagonal structure or constant correlations. Within our penalized (RE)ML framework this can be achieved by substituting the structured matrix for the scale matrix $\mathbf{\Omega}$ in (Eq. 11). This may be a suitable matrix chosen *a priori* or, in an empirical vein, an unpenalized estimate obtained from the data, imposing the structure selected.

3 Simulation study

3.1 Simulation set-up

Data for a simple paternal half-sib design comprising s unrelated sires with n=10 progeny each were simulated by sampling from appropriate multivariate normal distributions for q=5 and q=9 traits. Sample sizes considered were s=50, 100, 150, 200, 300, 400, 600 and 1000. A total of 90 sets of population parameters, 60 for q=5 and 30 for q=9 traits were considered.

Population parameters for q=5 were obtained by combining 12 sets of heritabilities (A to L) with 5 scenarios for genetic (r_G) and residual (r_E) correlations and phenotypic variances, labelled $\mathcal{I}$ to $\mathcal{V}$. This resulted in 60 combinations, labelled A-$\mathcal{I}$ to L-$\mathcal{V}$ in the following. Similarly, 10 sets of heritabilities (M to V) for q=9 traits were combined with correlation scenarios $\mathcal{I}$, $\mathcal{VI}$ and $\mathcal{VII}$ to yield combinations M-$\mathcal{I}$ to V-$\mathcal{VII}$. Details for heritabilities and correlation scenarios are summarized in Table 1 and Table 2, respectively. Heritabilities were chosen to decline with trait number and represent a range of cases, from equal values for all traits to sets of values which not only spanned almost the entire interval from zero to unity but also were very unevenly distributed. Combined with correlation scenarios ranging from zero throughout to genetic correlations of 0.8, this yielded coefficients of variation among the corresponding canonical eigenvalues ranging from 0 to 175% (see Table 1). A total of 1000 samples per case and sample size were obtained.

3.2 Analyses

REML estimates of $\mathbf{\Sigma}_G$ and $\mathbf{\Sigma}_E$ for each sample were obtained for different penalties and tuning factors using a Method of Scoring algorithm to locate the maximum of $\log \mathcal{L}(\boldsymbol{\theta})$ or $\log \mathcal{L}_P(\boldsymbol{\theta})$, followed by simple derivative-free search steps to ensure that convergence had been reached. This was done using a parameterisation to the elements of the canonical decomposition, λ_i and $t_{ij} \in \mathbf{T}$, as described by Meyer and Kirkpatrick (2010), restraining estimates of λ_i to the interval of $[0.0001, 0.9999]$.

A total of 12 penalties were examined. These comprised 8 penalties on the canonical eigenvalues, $\mathcal{P}_\lambda$, $\mathcal{P}_\lambda^\ell$, $\mathcal{P}_\lambda^{\ell 2}$, $\mathcal{P}_\beta^a$ for α=β=2, $\mathcal{P}_\beta^b$, $\mathcal{P}_\beta^c$, $\mathcal{P}_\beta^d$ for z=0 and $\mathcal{P}_\beta^e$ which is $\mathcal{P}_\beta^d$ for $z = 1$, and 4 penalties on matrices $\mathcal{P}_\Sigma$, $\mathcal{P}_\Sigma^2$, $\mathcal{P}_\rho$ and $\mathcal{P}_\rho^2$, as described above (see Section 2.2). All these employed a single tuning factor. In addition, the effect of applying a different tuning factor to the parts of penalties $\mathcal{P}_\lambda^{\ell 2}$, $\mathcal{P}_\Sigma^2$ and $\mathcal{P}_\rho^2$ corresponding to genetic and residual components were investigated.

3.3 Estimating the tuning factor

To determine the tuning factor ($\hat{\psi}$) for each analysis, estimates of $\mathbf{\Sigma}_G$ and $\mathbf{\Sigma}_E$, denoted as $\hat{\mathbf{\Sigma}}_G^\psi$ and $\hat{\mathbf{\Sigma}}_E^\psi$, were obtained for a range of possible values for ψ. A total of 311 values were used, comprising 0 to 2 in steps of 0.1, 2.2 to 5 in steps of 0.2, 5.5 to 10 in steps of 0.5, 11 to 100 in steps of 1, 102 to 250 in steps of 2, 255 to 500 in steps of 5 and 510 to 1000 in steps of 10. The 'best' value was then chosen using three different approaches.

First, for comparison with previous work, knowledge of the population parameters was utilised. Strategy $L_1(\mathbf{\Sigma}_G)$ simply involved calculating the entropy loss in the estimate of $\mathbf{\Sigma}_G$ for each tuning factor, selecting the value of ψ for which the loss in $\hat{\mathbf{\Sigma}}_G^\psi$ was minimized as best. In contrast, strategies V∞ and V1 considered the effect of penalization on both covariance matrices: For each ψ and estimates $\hat{\mathbf{\Sigma}}_G^\psi$ and $\hat{\mathbf{\Sigma}}_E^\psi$ the corresponding *unpenalized* log likelihood was calculated as

$$\log \mathcal{L}(\boldsymbol{\theta})^\psi = -\frac{1}{2}\Big[\big(s-1\big)\Big(\log|\mathbf{\Sigma}_B| + \mathrm{tr}\big(\mathbf{\Sigma}_B^{-1}\mathbf{M}_B\big)\Big) + s\big(n-1\big)\Big(\log|\mathbf{\Sigma}_W| + \mathrm{tr}\big(\mathbf{\Sigma}_W^{-1}\mathbf{M}_W\big)\Big)\Big] \quad (14)$$

with $\mathbf{\Sigma}_W = \hat{\mathbf{\Sigma}}_E^\psi + \frac{3}{4}\hat{\mathbf{\Sigma}}_G^\psi$ and $\mathbf{\Sigma}_B = \mathbf{\Sigma}_W + \frac{1}{4}n\hat{\mathbf{\Sigma}}_G^\psi$. This requires validation 'data', i.e. matrices of mean squares and cross-products between ($\mathbf{M}_B$) and within ($\mathbf{M}_W$) sires. For strategy V1 these were obtained by sampling one additional data set from the same distribution as the data for the analysis were sampled from. For strategy V∞, $\mathbf{M}_B$ and $\mathbf{M}_W$ were constructed from the population parameters. This can be thought of as equivalent to sampling an infinite number of additional data sets for the same data structure, hence the notation V∞. For both strategies, the value of ψ which maximised $\log \mathcal{L}(\boldsymbol{\theta})^{\psi}$ was then chosen as $\hat{\psi}$.

Secondly, $K-$fold cross-validation was used to estimate ψ using only the data available. This is a widely used strategy applicable to a range of problems; see, for instance, Hastie et al. (2001, Chapter 7). In brief, cross-validation involves splitting the data into so-called 'training' and 'validation' sets. Analyses are then carried out for a range of values for the quantity to be determined (e.g. ψ) using the training data and a corresponding criterion to assess the quality of the estimates (e.g. residual sums of squares) is obtained using the validation data. For K-fold cross-validation the data is split into K subsets of approximately equal size. K analyses are then carried out for each value of ψ, with the $i-$th subset treated in turn as the validation set and the remaining $K-1$ subsets forming the training set, and the tuning parameter is chosen based on the criterion averaged across the K validation sets.

Here, data were split into K folds of approximately equal size by sequentially assigning complete sire families to subsets. For i=1, K, the $i-$th subset was set aside for validation. The remaining $K-1$ subsets together where used to obtain estimates $\hat{\boldsymbol{\Sigma}}_G^{\psi}$ and $\hat{\boldsymbol{\Sigma}}_E^{\psi}$ for all values of ψ considered. Corresponding values for the unpenalized likelihood, $\log \mathcal{L}(\boldsymbol{\theta})_i^{\psi}$ (Eq. 14), in the validation data were then obtained and accumulated across folds. Finally, $\hat{\psi}$ was chosen as the value for which the average likelihood, $\sum_{i=1}^{K} \log \mathcal{L}(\boldsymbol{\theta})_i^{\psi}/K$, was maximized. Values of K=2, 3, 5 and 10 were considered, with the corresponding strategies denoted as CV2, CV3, CV5 and CV10 in the following.

The third approach used simply involved choosing $\hat{\psi}$ as the largest value of ψ for which the reduction in the unpenalized likelihood due to penalization from the maximum at ψ=0, $|\log \mathcal{L}(\boldsymbol{\theta})^{\psi} - \log \mathcal{L}(\boldsymbol{\theta})^{0}|$, did not exceed a selected value. Limits were chosen as the χ^2_{γ} values ($\times\frac{1}{2}$) which would be employed in a likelihood ratio test of a single parameter with error probability γ, 0.82 for γ=0.2, 1.36 for γ=0.1, 1.92 for γ=0.05 and 2.51 for γ=0.025, referred to as strategies L20%, L10%, L5% and L2.5% subsequently.

3.4 Summary statistics

As suggested by Lin and Perlman (1985), the effect of penalized estimation was evaluated as the percentage reduction in average loss (PRIAL) due to penalization,

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

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

In addition, the absolute and relative bias (in %) for parameter θ_i were calculated as $|\hat{\theta}_i - \theta_i|$ and $100\,(\hat{\theta}_i - \theta_i)/\theta_i$, respectively.

4 Results

4.1 Comparing penalties

Mean PRIAL values across all cases for individual covariance matrices and all penalties considered are summarized in Table 3 for a sample size of s=100. Using known population values (strategy V∞), reductions in average loss in estimates of $\mathbf{\Sigma}_G$ achieved were substantial, ranging form about 60% to more than 72%. Somewhat lower levels overall for q=9 than q=5 traits were, in part at least, due to the fact that the cases chosen for 9 traits involved a higher proportion of unfavourable scenarios, i.e. population values with substantially and unevenly spread canonical eigenvalues. The main exception was $\mathcal{P}_\lambda$ which penalized the untransformed canonical eigenvalues rather than their logarithmic values. For this penalty, PRIALs for estimates of $\mathbf{\Sigma}_E$ were substantially higher than for $\mathbf{\Sigma}_G$, suggesting that for strategy V∞ tuning parameter selection was more appropriate for the former.

As found earlier by Meyer and Kirkpatrick (2010), taking logarithms of the canonical eigenvalues ($\mathcal{P}_\lambda^\ell$) greatly improved the efficacy of a penalty proportional to the variance among them. Because canonical eigenvalues are a function of both $\mathbf{\Sigma}_G$ and $\mathbf{\Sigma}_E$, all penalties on the λ_i yielded marked improvements in estimates of $\mathbf{\Sigma}_E$ simultaneous to that for $\mathbf{\Sigma}_G$. Considering $\log(1 - \lambda_i)$ in addition to $\log(\lambda_i)$ ($\mathcal{P}_\lambda^{\ell 2}$ and all $\mathcal{P}_\beta^{\cdot}$) increased PRIALs for $\mathbf{\Sigma}_E$ further without affecting estimates of $\mathbf{\Sigma}_G$ detrimentally. Among the penalties invoking a Beta distribution for the canonical eigenvalues, those estimating the scale parameters tended to perform best. For q=5 traits, applying this to unscaled eigenvalues ($\mathcal{P}_\beta^b$; see (Eq. 8)) yielded higher PRIALs than scaling them in addition ($\mathcal{P}^c_\beta$), but corresponding differences for q=9 were reversed and much smaller. A possible explanation is that for the smaller number of traits attempting to estimate both range and scale parameters exacerbated errors. Considering the quite different underlying assumptions, the similarity of results for $\mathcal{P}^d_\beta$ and $\mathcal{P}^e_\beta$, i.e. the penalties based on the distribution of order statistics on the unit interval, and the other penalties assuming a common distribution of all λ_i was somewhat surprising.

Whilst achieving comparable PRIALs, penalizing the difference between genetic and phenotypic covariance or correlation matrices acted differently to penalties on canonical eigenvalues. As to be expected, considering $\mathbf{\Sigma}_G$ or $\mathbf{R}_G$ only ($\mathcal{P}_\Sigma$ and $\mathcal{P}_\rho$) yielded relatively small improvements in estimates of $\mathbf{\Sigma}_E$. Adding a corresponding penalty for the residual matrices ($\mathcal{P}^2_\Sigma$ and $\mathcal{P}^2_\rho$) increased PRIALs for estimates of $\mathbf{\Sigma}_E$ to levels comparable to those obtained penalizing canonical eigenvalues, again without reducing mean PRIALs for estimates of $\mathbf{\Sigma}_G$ notably. For q=9 traits, there was an unexpected, substantial difference between penalties on covariance and correlation matrix and shrinking both genetic and residual correlations towards their phenotypic counterparts increased the PRIAL for $\hat{\mathbf{\Sigma}}_G$ by 2% ($\mathcal{P}^2_\rho$ *vs.* $\mathcal{P}_\rho$). In contrast, corresponding differences for q=5 were considerably smaller. It is not clear how much this was an effect of the dimension or due to differences in population values.

Allowing for different tuning factors for parts of the penalty corresponding to genetic and residual effects increased the PRIAL for $\hat{\mathbf{\Sigma}}_G$ for q=5 from 72.9 to 73.7% for $\mathcal{P}^{\ell 2}_\lambda$, from 70.0 to 72.7% for $\mathcal{P}^2_\Sigma$ and from 72.2 to 74.3% for $\mathcal{P}^2_\rho$, i.e. by less than 3%. Corresponding PRIALs for $\hat{\mathbf{\Sigma}}_E$ were 65.6% ($\mathcal{P}^{\ell 2}_\lambda$), 64.9% ($\mathcal{P}^2_\Sigma$) and 62.7%, i.e. increased by more than 10% for $\mathcal{P}^2_\Sigma$. While non-negligible, the gains for estimates of $\mathbf{\Sigma}_G$ were deemed too small to off-set the dramatically increased computational requirements arising from the two-dimensional search for the optimal tuning factors needed, and not given any further consideration.

Mean PRIAL values discussed so far conceal a considerable range and variation in the ranking of penalties for individual cases. This is illustrated in Figure 2, which shows in PRIAL for $\hat{\mathbf{\Sigma}}_G$ for q=9 traits with individual cases in declining order of that achieved using penalty $\mathcal{P}^{\ell 2}_\lambda$. For strategy V∞, penalties on canonical eigenvalues assuming a common mean performed best when populations values for the λ_i were fairly similar, e.g. for R-I and M-I all population values were equal. For q=9 there was little difference in PRIALs for $\hat{\mathbf{\Sigma}}_G$ obtained between penalties assuming a Normal distribution on the logarithmic scale ($\mathcal{P}^\ell_\lambda$ and $\mathcal{P}^{\ell 2}_\lambda$) and a Beta distribution with estimated scale parameters ($\mathcal{P}^b_\beta$), though a tendency for $\mathcal{P}^b_\beta$ to yield slightly higher values for cases where penalized estimation worked least well was evident. Conversely, penalties derived assuming an Inverse Wishart matrix prior mostly yielded larger PRIALs for the other cases, in particular when penalizing the difference between genetic and phenotypic correlations. For q=5 trait, penalties $\mathcal{P}_\rho$ and $\mathcal{P}_\rho^2$ performed best for 35% of the individual cases considered, mainly those for which PRIALs for $\hat{\mathbf{\Sigma}}_G$ were less than average, while $\mathcal{P}_\lambda^\ell$ and $\mathcal{P}_\lambda^\ell$ yielded the highest values for 37% of cases. For q=9 where population values were predominantly chosen to represent scenarios for which penalties on the λ_i worked least well, penalty $\mathcal{P}_\rho^2$ thus yielded the highest PRIAL for 80% of cases.

4.2 Estimating tuning factors

A crucial part of penalized estimation is the estimation of the appropriate tuning factor to be used. Mean PRIAL values for $\hat{\mathbf{\Sigma}}_G$ for different strategies to determine $\hat{\psi}$ are summarized in Table 4 for selected penalties, q=5 traits and s=100 sires, together with the average proportion of replicates for which penalization increased rather than decreased the entropy loss in $\hat{\mathbf{\Sigma}}_G$. Corresponding PRIAL values for all penalties for strategies V∞, CV3 and L5% are given in Table 3. Clearly, mean values well above 70% when utilizing the population values (V∞ or $L_1(\mathbf{\Sigma}_G)$) present an overly optimistic view of the efficacy of penalized estimation. Considering only one additional sample for validation (strategy V1) introduced considerable sampling error and thus reduced PRIALs achieved by about 10%.

Examining regularized estimation of covariance matrix, Rothman et al. (2009) reported that using strategy V1 yielded similar results to cross-validation. However, in our case, mean PRIAL values obtained using cross-validation to determine $\hat{\psi}$ were consistently lower, i.e. suffered from additional noise introduced. Somewhat surprisingly, PRIALs achieved tended to decrease with the number of folds considered, K. This was accompanied by increasing variability of results for individual cases. Clearly, there was a trade-off between the sizes of the training and validation sets. One might expect that using a small training set (low K) would result in a $\hat{\psi}$ which was somewhat too large as it pertained to the sample size of the subset. On the other hand, a larger validation set might favour more accurate estimation of ψ. Similarly, a larger number of replications or folds might off-set potential inabilities to ascertain optimal values for ψ due to the limited size of the validation set. However, results for CV5 and CV10 were consistently worse than for lower values of K.

Inspection of the mean tuning factors did reveal a trend for $\hat{\psi}$ to decline with increasing number of folds. For penalties $\mathcal{P}_\beta^b$, $\mathcal{P}_\Sigma$ and $\mathcal{P}_\rho$ means where substantially higher than those obtained for strategy V∞, suggesting that lower PRIALs obtained using cross-validation were indeed due to over-penalization. For $\mathcal{P}_\lambda^\ell$ and $\mathcal{P}_\lambda^{\ell\,2}$ results were less consistent: for these penalties, estimates of ψ for cases with low coefficients of variation in the population canonical eigenvalues from strategy V∞ were very high. Using cross-validation, corresponding values tended to be substantially lower, so that overall means from strategies V∞ and CVK were similar. Using cross-validation also tended to reduce differences between penalties somewhat. Interestingly, as shown in Table 3, penalized estimation using penalties derived from the Beta distribution of order statistics appeared least affected by the noise introduced when estimating ψ. For strategy CV3 penalties $\mathcal{P}_\beta^d$ and $\mathcal{P}_\beta^e$ yielded the highest PRIAL in $\hat{\boldsymbol{\Sigma}}_G$ for 35% of the individual cases (q=5 and s=100), compared to 2% for strategy V∞.

Difficulties in deriving the optimal 'bending' factor theoretically led Hayes and Hill (1981) to suggest a choice on the basis of the sample size. An alternative in a likelihood framework of estimation is to select the tuning factor so that the corresponding reduction in the unpenalized likelihood does not exceed a given limit. When carrying out a likelihood ratio test for the difference between estimates from different models, minus twice the difference in log likelihood is contrasted to a value of the χ^2 distribution corresponding to the number of parameters tested and an error probability of γ. The smallest number of parameters which can be tested is p=1. Hence, choosing ψ as the largest value for which the resulting change in $\log \mathcal{L}(\boldsymbol{\theta})$ (sign ignored) does not exceed $\frac{1}{2}\chi^2_\gamma$ for one degree of freedom will result in a change in estimates which is not statistically significant. While it may not result in the optimal amount of regularization, it is appealing as a strategy to select a mild degree of penalization to exploit at least some of the advantages of penalized estimation without having to justify significant changes in parameter estimates. In addition, computational requirements to determine such ψ are considerably less than for cross-validation.

As shown in Table 3 and Table 4 employing such strategy yielded substantially improved estimates of $\boldsymbol{\Sigma}_G$, with PRIALs achieved consistently higher than for cross-validation. For a sample size of s = 100, an error probability of 5% or 10% appeared most appropriate. Mean estimates of ψ were markedly and consistently lower than for strategy V∞, indicating that this approach indeed resulted in under-penalization. This held especially for cases with similar population canonical eigenvalues (E-I, H-I, I-I, M-I and R-I). As illustrated in Figure 2, choosing ψ in this way also blurred differences between penalties. In a number of cases, in particular for q=9 traits, PRIALs for $\hat{\boldsymbol{\Sigma}}_G$ from strategy L5% were higher than those from V∞, but lower than from $L_1(\boldsymbol{\Sigma}_G)$.

4.3 Effects of sample size

The effect of sample size on the efficacy of regularized estimation is illustrated in Figure 3 for q=5. Clearly, penalization was most advantageous for small samples, with mean PRIALs for $\hat{\mathbf{\Sigma}}_G$ decreasing substantially as the number of sire families increased. There were marked differences between penalties and strategies to determine ψ, especially in the rate of decline of PRIALs with increasing s. This was least for penalty $\mathcal{P}^2_\rho$ and, moreover, choosing tuning factors on the basis of the change in log $\mathcal{L}(\boldsymbol{\theta})$ performed almost as well if knowledge of the population values could be exploited. In addition, $\mathcal{P}^2_\rho$ resulted in the highest PRIAL for both $\hat{\mathbf{\Sigma}}_G$ and $\hat{\mathbf{\Sigma}}_E$ for all sample sizes when using the change in likelihood to decide on the degree of penalization to be applied (strategy Lk%).

As noted above, improvements in $\hat{\mathbf{\Sigma}}_G$ when using cross-validation to determine the tuning factor were substantially less than for the other strategies. This difference tended to increase with sample size. Whilst consistently performing worst for strategy V∞, penalties derived assuming the distribution of canonical eigenvalues resembled that of order statistics on the unit interval yielded the highest PRIAL in $\hat{\mathbf{\Sigma}}_G$ for strategy CV3, with values for $\mathcal{P}^e_\beta$ almost 2% higher than for $\mathcal{P}^d_\beta$ for s=1000. It is not clear what this comparatively larger robustness against noise in estimates of ψ can be attributed to.

The decline in PRIAL with sample size was clearly a function of the number of traits considered, with reductions for q=9 markedly smaller. For instance, for $\mathcal{P}^2_\rho$ and strategy L5% the average PRIAL in $\hat{\mathbf{\Sigma}}_G$ declined from 69.4% for s=100 to 64.1% for s=400 and 60.2% for s=1000. Similarly, respective values for $\mathcal{P}^{\ell 2}_\lambda$ were 67.7%, 64.2% and 54.2%. This suggests that mild penalization is advantageous even for larger samples as the dimensions of the covariance matrices to be estimated increases.

4.4 Bias

As emphasized above, regularized estimation entails a trade-off between sampling variance and bias. Table 5 gives the mean relative bias in estimates of canonical eigenvalues for a sample size of s=100 sires and strategy V∞. Figure 4 further illustrates the relationship between estimates of λ_i and their true values for selected penalties and strategy V∞, with the solid line showing a one-to-one correspondence (unbiased estimates) and the dashed line representing the linear regression of estimates on population values. Patterns obtained when selecting the tuning factor based on the likelihood or using cross-validation were very similar. As indicated by theory, unpenalized estimates of the largest values were biased upwards and those of the smallest values biased downwards. Whilst the mean was expected to be estimated unbiasedly, a small upwards bias in $\bar{\lambda}$ – corresponding to a clustering of the smallest $\hat{\lambda}_i$ at zero – was evident, reflecting the effects of constraints on the parameter space.

Estimation placing a penalty on canonical eigenvalues tended to result in over-shrinkage, resulting in a downward bias of the largest and upward bias of the smallest values. This was the more pronounced the further the corresponding population values were spread apart. Similar results for shrinkage of the eigenvalues of a single matrix have been reported by Daniels and Kass (2001). While the relative bias in the smallest $\hat{\lambda}_i$ was substantial, absolute changes tended to be small and penalization clustered estimates closer to the one-to-one line.

Though PRIALs achieved were, by and large, comparable, penalties on matrix divergence clearly acted in a different manner to those on canonical eigenvalues. For penalty $\mathcal{P}_\Sigma$ upwards bias in $\hat{\lambda}_1$ was of similar magnitude and individual estimates showed the same pattern of distribution (Figure 4) than for unpenalized estimation, with penalization predominantly affecting the smallest values. This could be attributed to the fact that this penalty involved a component approximately proportional to the reciprocal of the $\hat{\lambda}_i$ (see Section 2.2.2). Shrinking genetic correlations towards their phenotypic counterparts ($\mathcal{P}_\rho$) yielded the least relative bias in estimates of the leading canonical eigenvalues. Penalizing both genetic and environmental components tended to shrink the largest $\hat{\lambda}_i$ more and the smallest $\hat{\lambda}_i$ less ($\mathcal{P}^\ell_\lambda$ *vs.* $\mathcal{P}^{\ell 2}_\lambda$ and $\mathcal{P}_\rho$ *vs.* $\mathcal{P}^2_\rho$). Allowing for separate tuning factors for the two parts of the respective penalties increased the downwards relative bias in $\hat{\lambda}_1$ somewhat (to -10.9% for $\mathcal{P}^{\ell 2}_\lambda$ and -5.3 for $\mathcal{P}^2_\rho$) whilst increasing the corresponding PRIALs, again illustrating that more improvement in estimates can come at the price of more bias.

It has to be stressed tough that bias in estimates of eigenvalues does not directly translate into bias in the corresponding covariance components or genetic parameters derived from them. As illustrated by various authors (e.g Ledoit and Wolf, 2004), eigenvalues of sample covariance matrices are systematically over-dispersed and biased, but the sample covariance matrix is an unbiased estimator. Standard, unpenalized REML estimates are biased, however, because estimates are constrained to the parameter space. This implies that for scenarios where no constraints are needed, no bias is notable. Mean estimates of heritabilities for individual scenarios for q=9 traits are shown in Figure 5. Not imposing a penalty, a slight bias for those with the highest and lowest population values is evident, arising from constrained estimation. The corresponding plot for a larger sample with s=1000 (not shown) exhibited virtually no bias.

Penalized estimation, however, yielded biased estimates of heritabilities, with a pattern of biases and differences between penalties analogous to those observed for the canonical eigenvalues. For instance, for $\mathcal{P}_\Sigma$ the smallest heritabilities were substantially biased upwards while estimates for the largest values were similar to those from unpenalized analyses. Penalties on the canonical eigenvalues resulted in marked underestimates of the highest heritabilities, with mean differences between estimates and population values for trait 1 of -0.130 for $\mathcal{P}^c_\beta$ and -0.113 for $\mathcal{P}^{\ell\,2}_\lambda$, whilst corresponding values for $\mathcal{P}_\Sigma$ and $\mathcal{P}^2_\rho$ were 0.009 and -0.054, respectively. Taking the average of absolute deviations across traits yielded values of 0.019 for $\mathcal{P}_\rho$ and 0.025 for $\mathcal{P}^2_\rho$, compared to 0.013 for unpenalized estimates, whilst mean absolute differences for the other penalties were about twice as high, ranging from 0.048 to 0.054. Using a likelihood based strategy (L5%) to determine the tuning factor approximately halved the bias in the heritability for trait 1 and reduced the mean absolute bias to 0.018 for $\mathcal{P}^2_\rho$ and 0.023 to 0.027 for the other penalties, except $\mathcal{P}_\rho$ for which this value remained unchanged. Analogous differences between penalties were found for q=5 traits, but using strategy L5% rather than V∞ had little effect on the mean absolute bias due to penalization.

The effects of penalized estimation on estimates of genetic correlations are illustrated in Figure 6 for case T-$\mathcal{VI}$ and a sample with s=100 sire families. Shown is a box-and-whisker plot of individual estimates across replicates, with correlations in ascending order of their population values, depicted by horizontal bars. Not surprisingly for such small sample, unpenalized estimates were subject to substantial sampling variation, and spread furthest for pairs of traits with the lowest heritabilities. Again, unpenalized estimates were clearly biased due to the effects of constraints on the parameter space, with mean deviations from the population values ranging from -0.504 $(8, 9)$ to 0.035 $(3, 8)$ and a mean, absolute bias across replicates of 0.064. Penalization dramatically reduced the spread of estimates, but increased bias to a range of -0.734 $(8, 9)$ to 0.103 $(4, 8)$, with a mean absolute value of 0.142. In all cases, genetic correlations were shrunk towards the corresponding phenotypic correlations (population values shown as dashed horizontal lines). In spite of the increase in bias, penalized estimation reduced the PRIAL in the estimate of the genetic correlation matrix by 77.3%. The corresponding value for $\hat{\mathbf{\Sigma}}_G$ was less, 58.1% for strategy V∞ and 60.5% for L5%, i.e. this was a scenario for which penalization worked somewhat less well (*c.f.* Figure 2).

Across all cases simulated, the mean absolute bias in estimates of genetic correlations for unpenalized estimates for s=100 was 0.046 for q=9 and 0.033 for q=5. Excluding $\mathcal{P}_\lambda$, penalized estimation using strategy V∞ to determine the tuning factor increased this to 0.082 ($\mathcal{P}_\Sigma$) to 0.105 ($\mathcal{P}_\rho^2$) for q=9 and 0.085 ($\mathcal{P}_\Sigma$) to 0.101 ($\mathcal{P}_\lambda^\ell$) for q=5. For strategy L5%, corresponding values ranged from 0.058 ($\mathcal{P}_\Sigma$) to 0.068 ($\mathcal{P}_\beta^a$) and 0.099 ($\mathcal{P}_\Sigma$) to 0.109 ($\mathcal{P}_\beta^a$). Thus penalized estimation increased the average bias in estimates of genetic correlation by a factor of two to three. Again, there was a tendency for the bias to be most pronounced for penalties imposed directly on the canonical eigenvalues.

5 Discussion

An extension of current, standard methodology to estimate genetic parameters in a mixed model framework has been outlined that has the scope to yield 'better' estimates, especially for multivariate analyses comprising more than just a few traits. This is achieved by penalizing the likelihood, with the penalty a function of the parameters aimed at reducing sampling variation. A number of suitable penalties have been investigated with emphasis on those 'borrowing strength' from estimates of the corresponding phenotypic covariance or correlation matrices, which are typically estimated much more accurately than their genetic counterparts. All penalties presented have a Bayesian motivation, i.e. can be derived assuming certain prior distributions for covariance matrices or their eigenvalues. In contrast to 'full' Bayesian analyses, location or scale parameters for the priors are estimated from the data at hand, i.e. our penalized maximum likelihood procedure can be considered as analogous to an empirical Bayes approach.

Simulation results have been presented demonstrating that substantial reductions in loss, i.e. the (average) difference between true and estimated covariance matrices, can be achieved. As expected, this comes at the price of increasing bias, over and above that introduced by constraining estimates to the parameter space in standard analyses. The magnitude and direction of the additional bias depend on the population parameters and penalty applied, but in general penalization caused estimates of the highest heritabilities to be reduced and those of the smallest heritabilities to be increased while estimates of genetic correlations were reduced in absolute value. With comparable (or better) reductions in loss to other penalties, $\mathcal{P}_\rho$ and $\mathcal{P}_\rho^2$ which shrink the genetic towards the phenotypic correlation matrix appeared to result in least bias.

As described by Meyer and Kirkpatrick (2010), penalized REML estimation for penalties on canonical eigenvalues is best implemented by reparameterising to the elements of $\mathbf{\Lambda}$ and $\mathbf{T}$

(*c.f.* Section 2.2.1), i.e. the canonical decomposition. In contrast to implementations for standard REML algorithms, which usually parameterize to the elements of the Cholesky factors of the covariance matrices to be estimated, this yields a parameterization in which derivatives of all covariance matrices with respect to all parameters are non-zero. Further, initial experience with this parameterization has been that it resulted in slower convergence rates than estimation of covariance matrices or of the corresponding Cholesky factors. Similar results for the parameterization of a single matrix to the elements of its eigen-decomposition have been reported by Pinheiro and Bates (1996). An additional disadvantage is that extension to models with additional random effects and penalties on their covariance matrices is not straightforward. Estimation using the penalties on matrix divergence proposed, however, is readily carried out using standard parameterizations, with calculation of derivatives of the penalty the only modification to existing REML algorithms required. Furthermore, penalties on additional covariance matrices can easily be imposed, provided appropriate tuning factors are available.

Cross-validation is a widely used technique to estimate the tuning or shrinkage factor in regularization problems from the data at hand. For our application, however, it was found to be only moderately successful, with errors in estimating ψ limiting PRIALs achieved and increasing the proportion of replicates for which penalization was detrimental. These errors appeared especially important for larger samples, i.e. in small samples any degree of penalization is likely to have a substantial effect while over-penalization becomes more harmful as sample size increases. An added problem with cross-validation for data with a genetic family structure is that of representative sampling of data subsets. In our simulation setting, assigning whole sire families to individual folds was a natural choice and yielded higher PRIAL values than a random assignment. In practical data sets with arbitrary relationships and fixed effects, choices are less obvious and while procedures to optimize sampling exist (e.g Tillé, 2006), guidelines to good sampling strategies in a mixed model setting are scarce.

Moreover, cross-validation is laborious, increasing the number of analyses required by orders of magnitude. A sequential search for the optimal tuning factor was used in our simulation study. A more efficient strategy would have been to use one of the many structured, one-dimensional optimization methods available, e.g. a quadratic approximation of the average likelihood from the validation sets. However, this relies on the ‘validation’ curves to be smooth, increasing monotonically to a maximum and then decreasing again. This was not always the case in the simulations presented – some jagged curves were encountered, in particular for the smallest sample sizes. Presumably this was due likelihood surfaces which were very flat around the area of the maxima, resulting in inaccurate location of these points. Use of such techniques was thus disregarded here.

Fortunately, choice of $\hat{\psi}$ based on the decrease in the unpenalized likelihood from its maximum at $\psi = 0$ can result in penalized estimates closely related to those which would be obtained if population values were known. As demonstrated, such strategies yielded average reductions in loss for estimates of the genetic covariance matrix substantially higher than those estimating ψ by cross-validation, and values comparable to those achieved using knowledge of the population parameters for some penalties. Choosing the limit so that the change in likelihood was just not statistically significant appeared to be a sensible choice to select a mild degree of penalization. While it did not perform quite as well for individual cases where all population canonical eigenvalue were very similar, this is a constellation which is unlikely to be of practical relevance in quantitative genetic applications.

Work so far has considered a balanced scenario, with all traits in a multivariate analysis measured for all individuals. Often, however, we have a substantial discrepancy between the number of observations available for different traits. For instance, we may have a number of traits recorded on a substantial number of individuals whilst records for other, hard to measure traits are available for a small subset only. In that case, it is necessary to penalize parts of the genetic covariance matrix corresponding to such grouping of traits differently. To achieve this, a possible extension of the penalties on the divergence between genetic and phenotypic matrices might involve assuming a Generalized Inverse Wishart prior distribution (e.g. Brown, 2006), similar to the approach taken, for instance, by Cantet (2010) to allow for different degrees of belief. Future work should consider the scope for such differential regularization.

Even with today's computational resources, there may be problems where an analysis considering all traits of interest is not feasible, so that elements of the complete covariance matrix have to be obtained through a series of analyses of selected subsets of traits. This yields multiple estimates of variance and some covariance components which need to be pooled whilst ensuring the resulting matrix is positive definite. Typically, this is done by considering one matrix at a time, e.g. genetic or residual, using some method as the iterative summation of expanded part matrices (Mäntysaari, 1999) or treating estimates from individual analyses as 'pseudo-data' (Thompson et al., 2005). Alternatively, a strategy comprising simple averaging combined with a regression of the eigenvalues of the resulting matrix towards their mean to ensure the smallest value is greater than zero is frequently employed. The latter is commonly referred to as bending, though it differs from the original suggestion by Hayes and Hill (1981) as it involves a single matrix only. Results from this paper suggest that considering all matrices of interest simultaneously when combining estimates from analyses of subsets, together with some shrinkage towards the phenotypic covariance matrix may be advantageous.

6 Conclusions

Penalized maximum likelihood estimation provides the means to 'make the most' of limited and precious data and facilitates more stable estimation for multi-dimensional analyses even when samples are somewhat larger. We anticipate that it will become part of our everyday toolkit as truly multivariate estimation for quantitative genetic problems becomes routine. At the present state of knowledge, a mild penalty on the divergence of the genetic from the phenotypic correlation matrix, chosen on the basis of the change in likelihood from an unpenalized analysis, appears the most suitable option for practical applications.

Acknowledgments

This work was supported by Meat and Livestock Australia under grant B.BFG.0050. I am indebted to Mark Kirkpatrick, Daniel Gianola and a reviewer of Meyer and Kirkpatrick (2010) for numerous helpful suggestions and discussions.

References

Brown P.J. Inverted Wishart Distribution, Generalized. In: Encyclopedia of Environmetrics. John Wiley and Sons, Ltd (2006). doi: 10.1002/9780470057339.vag008.

Cantet R.J.C. Bayesian estimation of a genetic covariance matrix with different degrees of belief via a Generalized Inverted Wishart distribution. J. Anim. Sci. Vol. 88, E-Suppl. 2/J. Dairy Sci. Vol. 93, E-Suppl. 1/Poult. Sci. Vol. 89, E-Suppl. 1 88 (2010) 186 (Abstr).

Chen C.F. Bayesian inference for a normal dispersion matrix and its application to stochastic multiple regression analysis. J. Roy. Stat. Soc. B 41 (1979) 235–248.

Cheverud J.M. A comparison of genetic and phenotypic correlations. Evolution 42 (1988) 958–968.

Daniels M.J., Kass R.E. Shrinkage estimators for covariance matrices. Biometrics 57 (2001) 1173–1184.

Evans M., Hastings N., Peacock B. Beta distribution. In: Statistical distributions, Series in Probability and Statistics, chap. 5. Wiley, New York, 3rd edn. (2000), pp. 34–42.

Green P.J. Penalized likelihood. In: Encyclopedia of Statistical Sciences, vol. 2. John Wiley & Sons (1998), pp. 578–586.

Harville D.A. Maximum likelihood approaches to variance component estimation and related problems. J. Amer. Stat. Ass. 72 (1977) 320–338.

Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. Springer Series in Statistics. Springer Verlag, New York, NY, USA (2001).

Hayes J.F., Hill W.G. Modifications of estimates of parameters in the construction of genetic selection indices ('bending'). Biometrics 37 (1981) 483–493.

Hill W.G., Thompson R. Probabilities of non-positive definite between-group or genetic covariance matrices. Biometrics 34 (1978) 429–439.

Hoerl A.E., Kennard R.W. Ridge regression: applications to nonorthogonal problems. Technometrics 12 (1970) 69–82.

Huang J.Z., Liu N., Pourahmadi M., Liu L. Covariance matrix selection and estimation via penalised normal likelihood. Biometrika 93 (2006) 85–98. doi: 10.1093/biomet/93.1.85.

James W., Stein C. Estimation with quadratic loss. In: Fourth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1 (1961), pp. 361–379.

Kirkpatrick M., Helper A., Coolaborator B., Other C. The trials and tribulations of Mr. Bayes. Evolution 000 (2011) 000–000.

Lawley D.N. Tests of significance for the latent roots of covariance and correlation matrices. Biometrika 43 (1956) 128–136. doi: 10.1093/biomet/43.1-2.128.

Ledoit O., Wolf M. A well-conditioned estimator for large-dimensional covariance matrices. J. Multiv. Anal. 88 (2004) 365–411.

Levina E., Rothman A.J., Zhu J. Sparse estimation of large covariance matrices via a nested Lasso penalty. Ann. Appl. Stat. 2 (2008) 245–263. doi: 10.1214/07-AOAS139.

Lin S.P., Perlman M.D. A Monte Carlo comparison of four estimators of a covariance matrix. In: Krishnaish P.R., ed., Multivariate Analysis, vol. 6. North-Holland, Amsterdam (1985), pp. 411–428.

Mäntysaari E.A. Derivation of multiple trait reduced random regression (RR) model for the first lactation test day records of milk, protein and fat. In: 50th Annual Meeting. Europ. Ass. Anim. Prod. (1999). Mimeo., 8pp.

Meng X.L. 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 (2008) 1542–1552.

Meyer K. Factor-analytic models for genotype x environment type problems and structured covariance matrices. Genet. Select. Evol. 41 (2009) 21. doi: 10.1186/1297-9686-41-21.

Meyer K., Kirkpatrick M. Cheverud revisited: Scope for joint modelling of genetic and environmental covariance matrices. Proc. Ass. Advan. Anim. Breed. Genet. 18 (2009) 438–441. URL `http://www.aaabg.org/proceedings18/files/meyer438.pdf`.

Meyer K., Kirkpatrick M. Better estimates of genetic covariance matrices by 'bending' using penalized maximum likelihood. Genetics 185 (2010) 1097–1110. doi: 10.1534/genetics.109.113381.

Meyer K., Kirkpatrick M., Gianola D. Penalized maximum likelihood estimates of genetic covariance matrices with shrinkage towards phenotypic dispersion. Proc. Ass. Advan. Anim. Breed. Genet. 19 (2011) 000–000.

Pinheiro J.C., Bates D.M. Unconstrained parameterizations for variance-covariance matrices. Stat. Comp. 6 (1996) 289–296.

Rothman A.J., Levina E., Zhu J. Generalized thresholding of large covariance matrices. J. Amer. Stat. Ass. 104 (2009) 177–186. doi: 10.1198/jasa.2009.0101.

Schäfer J., Strimmer K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Stat. Appl. Genet. Mol. Biol. 4 (2005) 32. doi: 10.2202/1544-6115.1175.

Sorensen D., Gianola D. Likelihood, Bayesian and MCMC Methods in Quantitative Genetics. Springer Verlag (2002).

Stein C. Estimation of a covariance matrix. In: Reitz lecture. 39th Annual Meeting of the Institute of Mathematical Statistics. Atlanta (1975).

Thompson R., Brotherstone S., White I.M.S. Estimation of quantitative genetic parameters. Phil. Trans. R. Soc. B 360 (2005) 1469–1477. doi: 10.1098/rbstb.2005.1676.

Tibshirani R. Regression shrinkage and selection via the lasso. J. Roy. Stat. Soc. B 58 (1996) 267–288.

Tibshirani R. Regression shrinkage and selection via the lasso: a retrospective. J. Roy. Stat. Soc. B 73 (2011) 273–282. doi: 10.1111/j.1467-9868.2011.00771.x.

Tillé Y. Sampling algorithms. Springer Series in Statistics. Springer Verlag (2006).

Tyrisevä A.M., Meyer K., Fikse F., Ducrocq V., Jakobsen J., Lidauer M.H., Mäntysaari E.A. Principal component approach in variance component estimation for international sire evaluation. Genet. Select. Evol. 43 (2011) 21. doi: 10.1186/1297-9686-43-21.

Warton D.I. Penalized normal likelihood and ridge regularization of correlation and covariance matrices. J. Amer. Stat. Ass. 103 (2008) 340–349. doi: 10.1198/016214508000000021.

Witten D.M., Tibshirani R. Covariance-regularized regression and classification for high dimensional problems. J. Roy. Stat. Soc. B 71 (2009) 615–636. doi: 10.1111/j.1467-9868.2009.00699.x.

Yap J.S., Fan J., Wu R. Nonparametric modeling of longitudinal covariance structure in functional mapping of quantitative trait loci. Biometrics 65 (2009) 1068–1077. doi: 10.1111/j.1541-0420.2009.01222.x.

Table 1: Population values for heritabilities (×100) for individual cases together with the coefficient of variation (in %) amongst canonical eigenvalues for different correlation scenarios

(a) 5 traits

	A	B	C	D	E	F	G	H	I	J	K	L
						Heritability						
1	40	50	60	70	90	70	80	90	20	30	50	60
2	40	45	50	55	50	70	30	30	20	25	20	10
3	40	40	40	40	30	40	30	10	20	20	15	10
4	40	35	30	25	20	10	30	10	20	15	10	10
5	40	30	20	10	10	10	30	10	20	10	5	10
						Coefficient of variation						
I	0	20	40	59	79	75	56	115	0	40	88	112
II	115	116	118	122	134	124	127	168	148	151	164	175
III	64	67	73	83	95	92	81	129	87	96	123	135
IV	76	79	86	95	112	101	101	145	98	108	137	150
$\mathcal{V}$	70	70	74	82	96	93	83	124	81	81	103	120

(b) 9 traits

	M	N	O	P	Q	R	S	T	U	V
					Heritability					
1	40	60	90	75	70	20	35	50	60	80
2	40	55	60	70	70	20	30	50	50	40
3	40	50	50	60	70	20	25	20	10	10
4	40	45	50	50	40	20	20	15	10	10
5	40	40	30	40	40	20	20	15	10	10
6	40	35	30	30	40	20	20	10	10	10
7	40	30	20	20	10	20	15	10	10	10
8	40	25	20	10	10	20	10	5	10	5
9	40	20	10	5	10	20	5	5	10	5
					Coefficient of variation					
I	0	34	63	64	65	0	47	88	100	124
$\mathcal{V}I$	73	74	85	85	83	97	102	113	113	131
$\mathcal{V}II$	77	81	93	90	89	102	111	127	132	150

Table 2: Population values for genetic ($r_{G\,ij}$) and environmental ($r_{E\,ij}$) correlations between traits i and j together with values for phenotypic variances (σ_i^2) for different scenarios

Scenario	$r_{G\,ij}$	$r_{E\,ij}$	σ_i^2
I	0.0	0.0	1.0
II	0.8	0.0	1.5^{i-1}
III	$0.6^{\lvert i-j\rvert}$	$0.5 + (-0.4)^{\lvert i-j\rvert}$	3.0, 2.0, 1.0, 2.0, 3.0
IV	$0.02\,i + (-0.8)^{\lvert i-j\rvert}$	$0.5 + (-0.4)^{\lvert i-j\rvert}$	as III
$\mathcal{V}$	$0.5 + (-1)^i\,0.05\,j$	$0.2 + (-1)^j\,0.1\,i$	as III
$\mathcal{V}I$	$0.7^{\lvert i-j\rvert}$	$0.2 + (-1)^j\,0.05\,i$	2.0, 1.0, 3.0, 2.0, 1.0, 2.0, 3.0, 1.0, 2.0
$\mathcal{V}II$	$0.02\,i + (-0.8)^{\lvert i-j\rvert}$	$0.5 + (-0.2)^{\lvert i-j\rvert}$	as $\mathcal{V}I$

Table 3: Mean percentage reduction in average loss (PRIAL) in estimates of covariance matrices ($\mathbf{\Sigma}_G$ genetic, $\mathbf{\Sigma}_E$ residual and $\mathbf{\Sigma}_P$ phenotypic) for different penalties (see text) and three strategies to determine the tuning factor (Data for 100 sires).

		$\mathcal{P}_\lambda$	$\mathcal{P}_\lambda^\ell$	$\mathcal{P}_\lambda^{\ell\,2}$	$\mathcal{P}_\beta^a$	$\mathcal{P}_\beta^b$	$\mathcal{P}_\beta^c$	$\mathcal{P}_\beta^d$	$\mathcal{P}_\beta^e$	$\mathcal{P}_\Sigma$	$\mathcal{P}_\Sigma^2$	$\mathcal{P}_\rho$	$\mathcal{P}_\rho^2$
5 traits													
$\mathbf{\Sigma}_G$	V∞	35.8	71.3	72.9	66.7	71.4	66.1	68.1	67.9	70.6	70.0	72.0	72.2
	CV3	23.1	55.9	60.7	59.2	58.1	58.3	61.2	61.1	54.9	52.9	54.4	56.9
	L5%	41.3	68.3	70.2	67.6	69.5	70.0	69.8	69.3	64.1	66.7	70.5	71.5
$\mathbf{\Sigma}_E$	V∞	57.9	43.4	61.6	59.3	60.9	59.8	59.7	59.7	13.3	54.2	37.3	60.0
	CV3	14.1	26.7	44.3	38.7	36.0	32.5	38.0	39.6	10.7	43.0	22.8	40.9
	L5%	43.6	35.0	55.9	54.2	54.1	51.6	53.9	54.0	7.2	51.4	33.2	55.7
$\mathbf{\Sigma}_P$	V∞	1.1	1.2	1.3	1.3	1.2	1.1	1.2	1.2	1.2	1.7	2.2	2.4
	CV3	-0.4	0.4	0.5	0.3	0.1	0.0	0.2	0.3	0.2	0.1	0.4	0.8
	L5%	-0.7	0.7	0.8	0.5	0.5	0.2	0.4	0.5	0.3	1.0	1.0	1.2
9 traits													
$\mathbf{\Sigma}_G$	V∞	48.4	64.8	68.4	65.3	68.9	69.2	66.9	66.7	64.0	62.8	71.3	73.3
	L5%	24.1	67.5	67.7	65.4	66.5	66.0	66.3	66.4	68.0	67.7	69.5	69.4
$\mathbf{\Sigma}_E$	V∞	62.9	60.5	68.8	67.8	67.3	66.1	68.0	68.3	10.4	61.1	57.9	70.2
	L5%	63.0	16.4	59.3	60.9	62.6	63.3	61.6	61.7	9.9	47.4	17.2	56.3
$\mathbf{\Sigma}_P$	V∞	1.3	1.9	1.9	2.0	1.8	1.7	2.0	2.0	1.2	1.7	2.5	3.0
	L5%	1.2	0.5	1.1	1.2	1.3	1.3	1.2	1.2	0.6	0.7	1.1	1.2

Table 4: Mean percentage reduction in average loss (PRIAL) in estimates of the genetic covariance matrix together with average proportion (in %) of replicates for which penalisation increased the loss in estimates, for different penalties (see text) and strategies to determine the tuning factor (Data for 5 traits and 100 sires).

	Population values			Crossvalidation				Likelihood			
	$L_1(\mathbf{\Sigma}_G)$	V∞	V1	CV2	CV3	CV5	CV10	L20%	L10%	L5%	L2.5%
PRIAL											
$\mathcal{P}^{\ell}_{\lambda}$	75.6	71.3	60.6	55.8	55.9	50.4	44.4	68.8	69.6	68.3	66.3
$\mathcal{P}^{\ell 2}_{\lambda}$	76.1	72.9	63.7	61.8	60.7	58.1	55.3	69.3	70.7	70.2	69.0
$\mathcal{P}^{b}_{\beta}$	74.9	71.4	62.9	59.8	58.1	53.9	48.2	68.2	69.6	69.5	68.6
$\mathcal{P}_{\Sigma}$	75.2	70.6	60.6	56.7	54.9	52.7	50.0	68.7	68.0	64.1	61.0
$\mathcal{P}_{\rho}$	75.9	72.0	62.9	58.1	54.4	51.6	46.1	70.2	71.2	70.5	68.9
Increased loss											
$\mathcal{P}^{\ell}_{\lambda}$	0.0	7.3	8.7	15.3	14.6	14.6	14.7	8.7	10.5	12.0	13.6
$\mathcal{P}^{\ell 2}_{\lambda}$	0.0	6.5	7.5	13.4	13.0	13.2	13.2	7.0	8.5	10.0	11.4
$\mathcal{P}^{b}_{\beta}$	0.0	6.4	7.5	14.1	13.6	14.0	14.1	7.0	8.4	9.8	11.1
$\mathcal{P}_{\Sigma}$	0.0	4.6	8.9	15.6	15.4	15.5	15.4	10.3	12.8	15.6	17.9
$\mathcal{P}_{\rho}$	0.0	4.0	7.1	10.5	9.9	10.2	10.4	6.6	8.0	9.2	10.4

Table 5: Mean relative bias (in %) in estimates of the canonical eigenvalues and their mean ($\bar{\lambda}$) for different penalties (strategy V∞; 100 sires)

λ_i	None	$\mathcal{P}_\lambda$	$\mathcal{P}_\lambda^\ell$	$\mathcal{P}_\lambda^{\ell\,2}$	$\mathcal{P}_\beta^a$	$\mathcal{P}_\beta^b$	$\mathcal{P}_\beta^c$	$\mathcal{P}_\beta^d$	$\mathcal{P}_\Sigma$	$\mathcal{P}_\Sigma^2$	$\mathcal{P}_\rho$	$\mathcal{P}_\rho^2$
5 traits												
$\bar{\lambda}$	2.3	-5.4	6.6	2.1	3.4	0.9	-1.2	1.0	11.2	10.9	4.7	2.3
1	9.5	-12.9	-3.7	-9.6	-8.9	-11.2	-12.9	-11.5	8.1	3.2	1.3	-3.0
2	26.5	16.1	16.3	16.1	24.7	19.5	19.5	19.5	24.9	26.3	16.2	15.5
4	-19.4	9.1	57.7	48.3	38.8	41.3	31.0	39.4	39.1	47.0	37.3	37.1
5	-78.8	-38.1	101.3	81.6	36.1	44.7	26.6	52.2	75.3	88.6	57.2	56.7
av.[a]	30.2	19.6	41.6	36.4	28.3	29.4	23.4	30.3	34.4	38.8	26.6	26.5
9 traits												
$\bar{\lambda}$	4.4	-9.9	9.5	3.2	11.8	2.1	0.8	7.2	19.7	18.2	6.3	2.5
1	22.4	-22.4	-3.8	-13.7	-6.9	-16.8	-18.5	-12.7	21.6	8.8	2.9	-4.2
2	16.6	-17.5	-6.8	-10.0	0.5	-10.9	-11.4	-6.2	16.1	11.0	-0.7	-3.1
5	15.3	23.3	33.6	29.4	47.4	36.4	35.3	39.7	33.2	39.2	23.7	23.6
8	-85.6	-16.4	139.4	111.7	80.8	86.2	77.8	104.4	87.5	110.1	86.5	82.2
9	-97.9	-35.0	270.1	217.5	133.2	147.7	134.0	190.5	184.1	217.0	133.4	131.7
av.	39.9	16.6	68.4	57.3	48.8	48.4	45.1	56.9	54.0	61.9	40.0	39.1

[a] Average of all q absolute values

Figure 1: Probability density function for various Beta distributions: (a) α=β: — · — α=2, ——— α=3, — — — α=4 and — - — α=5 (b) α=0.6 + z, β=1.2 + z: - - - - z=0 and — · — z=1, (c) order statistics for 5 variables (z=0): - - - - first, — · — second, ——— third, — — — fourth and — - — fifth (d) as (c) for z=1

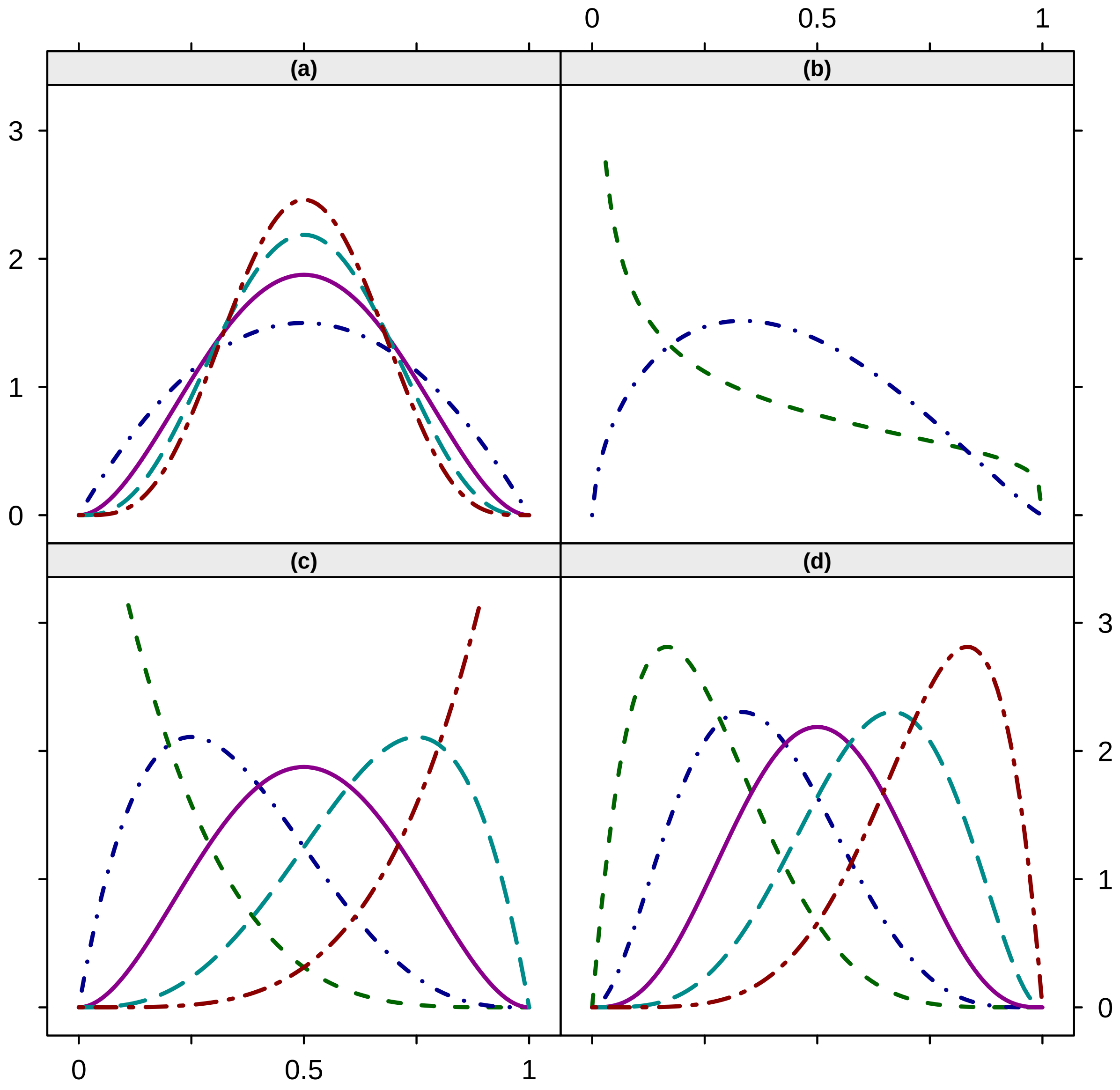


Figure 2: Percentage reduction in average loss (PRIAL) in estimates of the genetic covariance matrix for individual cases comprising 9 traits and different penalties (▼ $\mathcal{P}_\Sigma$, ■ $\mathcal{P}^2_\rho$, ▲ $\mathcal{P}^b_\beta$ and ● $\mathcal{P}^{\ell 2}_\lambda$; see text) , determining tuning factors on the basis of population values (V∞) and by limiting the change in likelihood (L5%)

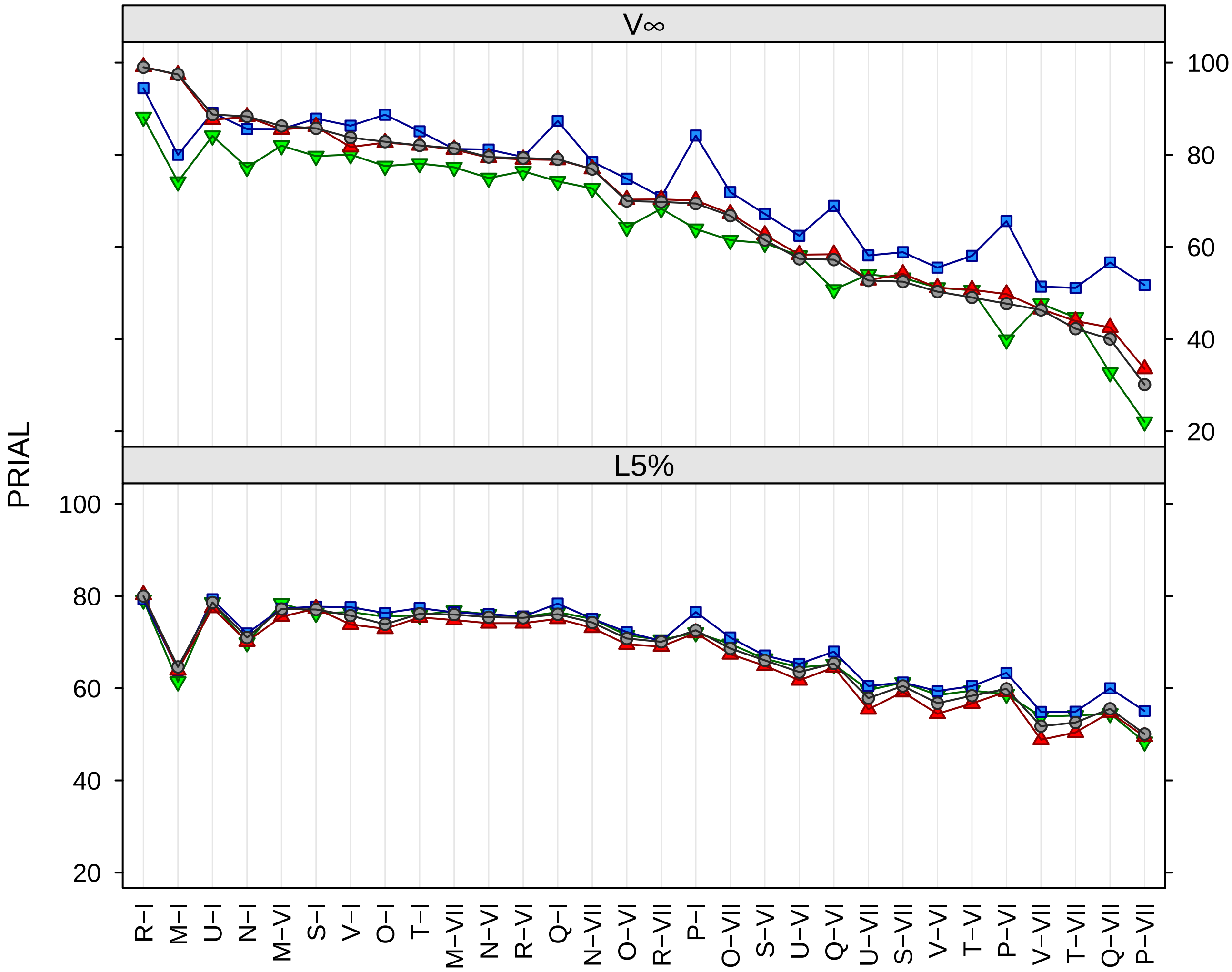


Figure 3: Mean percentage reduction in average loss (PRIAL) in estimates of the genetic covariance matrix (5 traits) for different sample sizes, penalties (see text) and strategies to determine the tuning factor (● using population values (V∞), ■ limiting the change in likelihood (L5%) and ▼ using cross-validation (CV3))

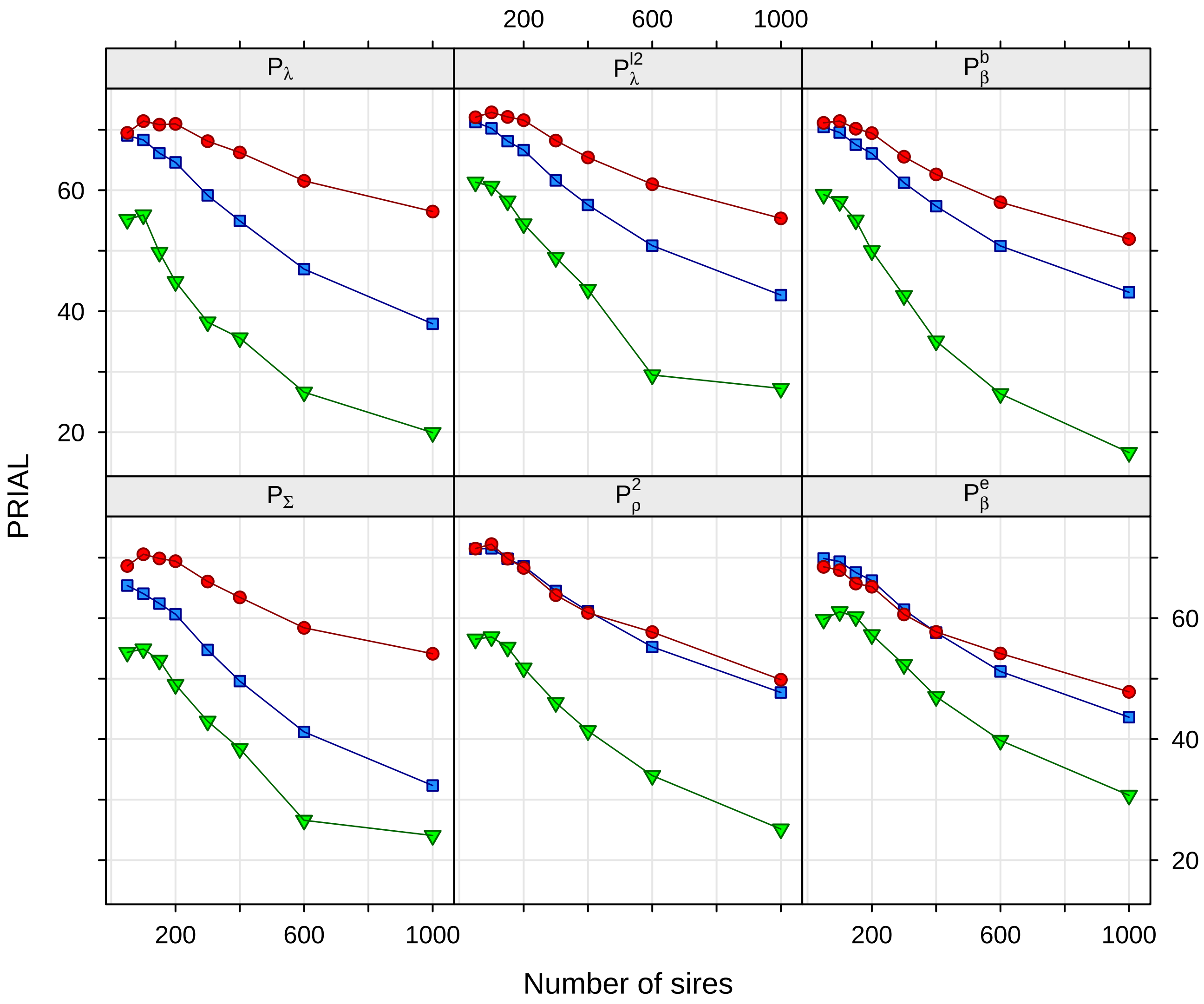


Figure 4: Mean estimates of canonical eigenvalues for individual cases (5 traits, 100 sires) for different penalties (see text) using population values (strategy V∞) to determine the tuning factor (● first, ■ second, ▼ third, ♦ fourth and ▲ fifth eigenvalue)

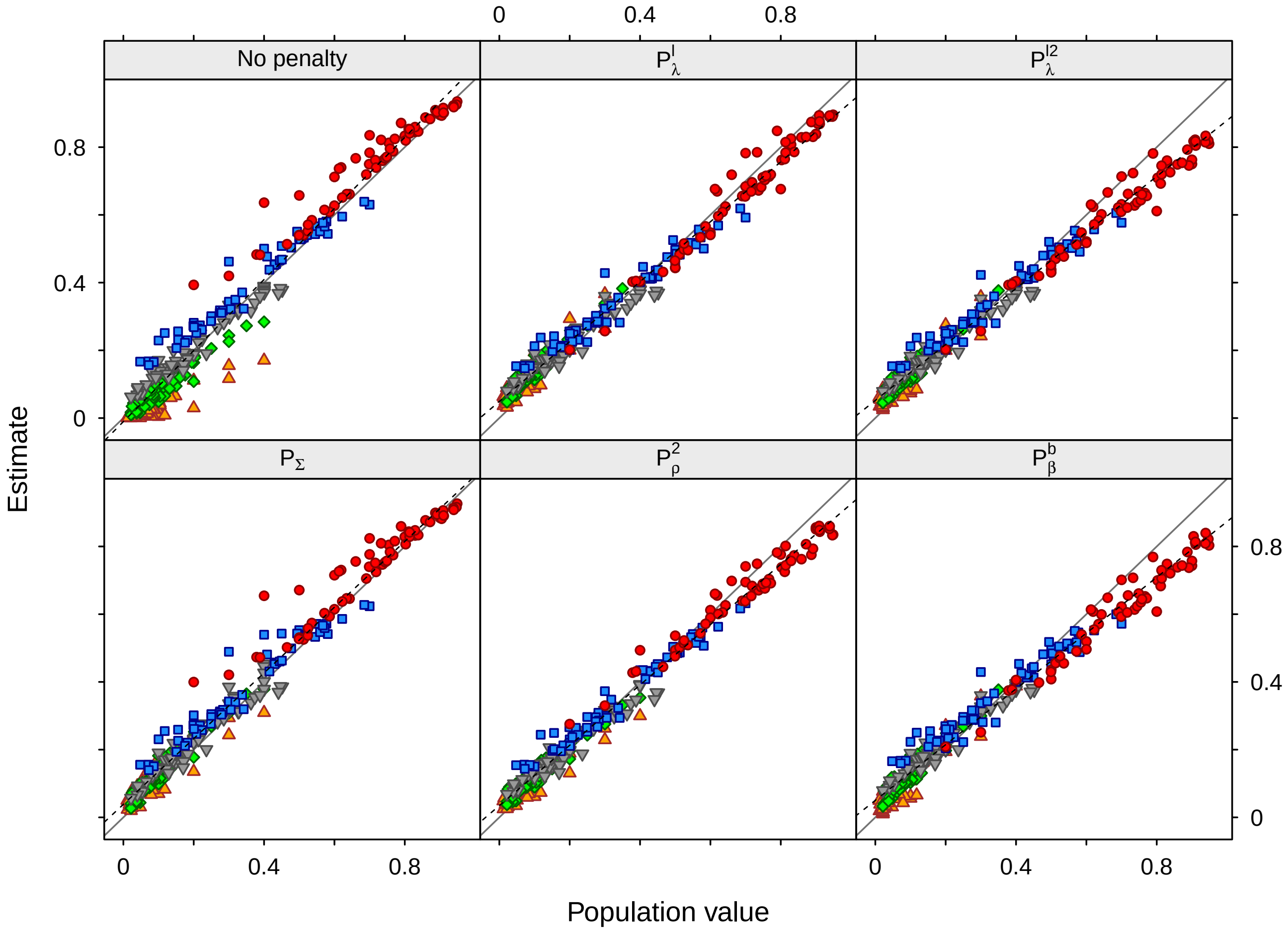


Figure 5: Mean estimates of heritabilities for individual cases (9 traits, 100 sires) for different penalties (see text) using population values (strategy V∞) to determine the tuning factor (● trait 1, ■ trait 2, ▼ trait 3 to 7, ♦ trait 8 and ▲ trait 9)

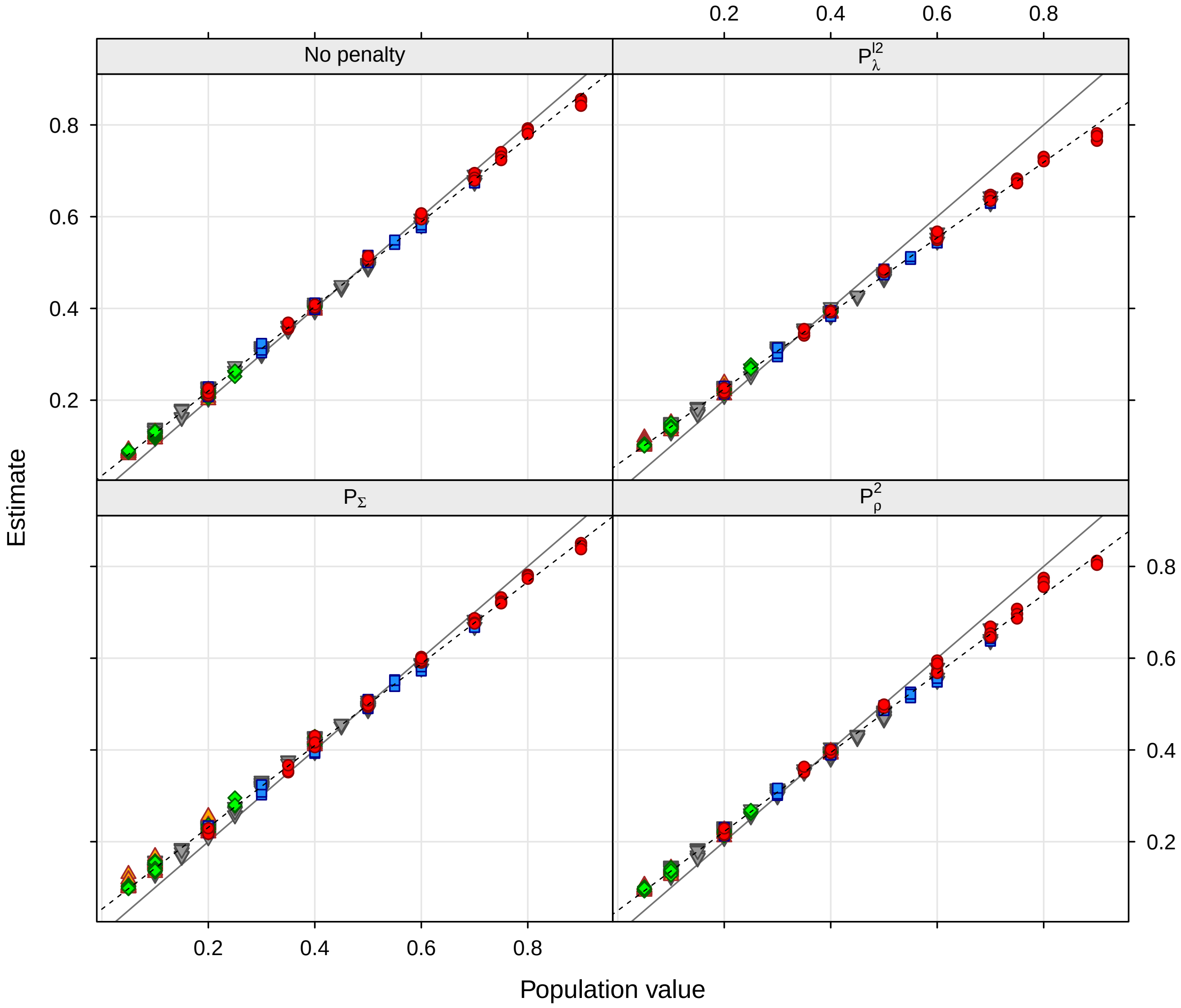
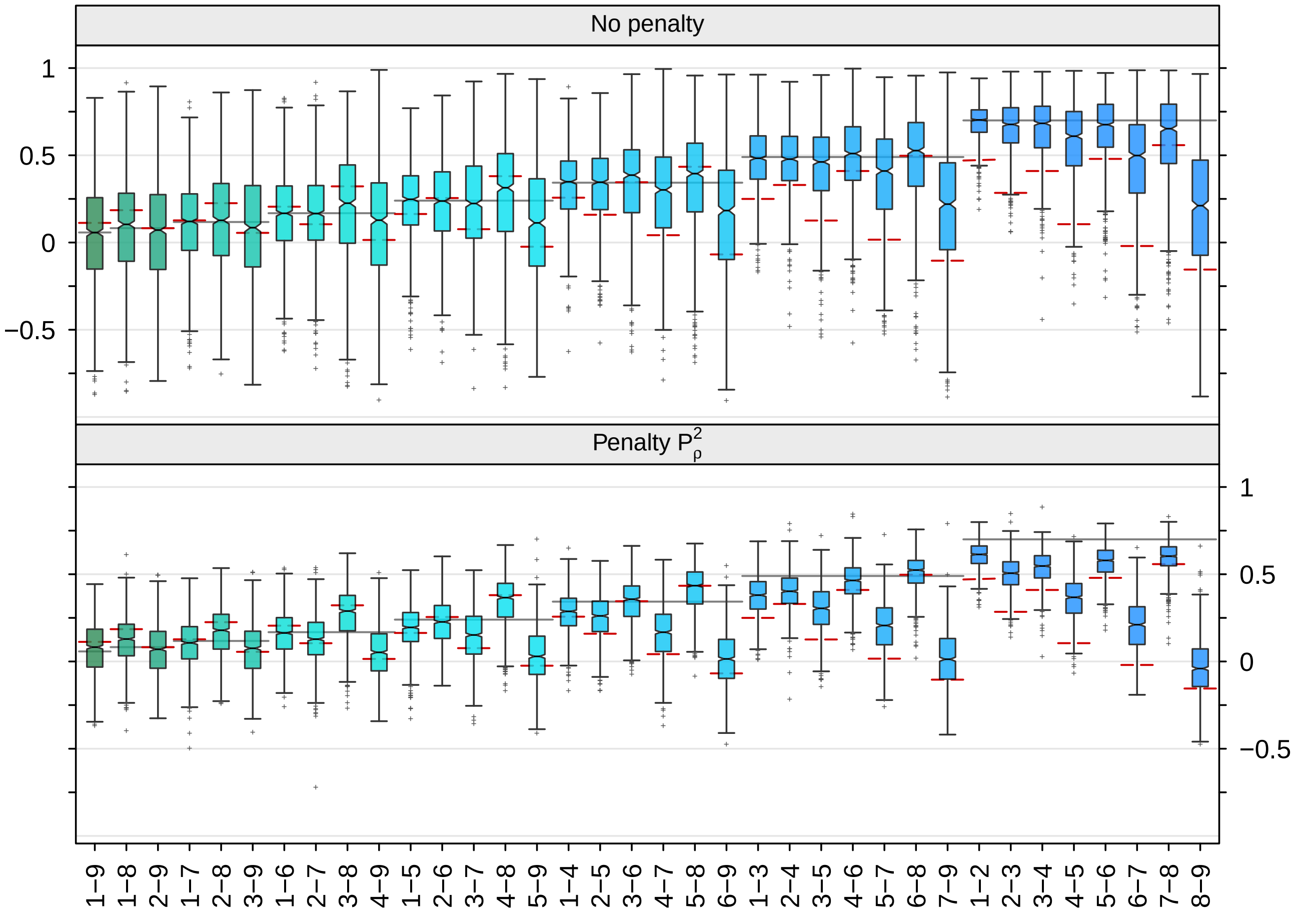


Figure 6: Distribution of estimates of genetic correlations between traits i and j (i–j) across replicates for case T-$\mathcal{VI}$ (s=100 sires, strategy V∞); horizontal bars show population values for genetic (———) and phenotypic (— — —) correlations


No penalty
Penalty P_ρ^2