Proc. Assoc. Advmt. Anim. Breed. Genet. ***19***:*83–86*

PERFORMANCE OF CROSS-VALIDATION AND LIKELIHOOD BASED STRATEGIES TO SELECT TUNING FACTORS FOR PENALIZED ESTIMATION

Karin Meyer

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

SUMMARY

Using simulation, the efficacy of penalized maximum likelihood estimation of genetic covariances when employing different strategies to determine the necessary tuning parameter is investigated. It is shown that errors in estimating the tuning factor from the data using cross-validation can reduce the percentage reduction in average loss at modest sample sizes from 70% or more to 60% or less. Mild penalization by limiting the change in likelihood is shown to perform well and to yield choices which are highly correlated with those based on the population parameters. Likelihood based selection of the tuning parameter is recommended as a simple and effective alternative to cross-validation.

INTRODUCTION

Penalized estimation of genetic parameters has been shown to be capable of yielding 'better' estimates, i.e. estimates that are on average closer to the population values than standard, non-penalized estimates (Meyer and Kirkpatrick 2010). An exposé of the underlying principles and salient features is given in a companion paper in this volume (Meyer 2011). Penalized estimation requires the choice of a so-called tuning factor, denoted as ψ, which determines the relative emphasis to be given to the penalty. Simulation studies examining the benefits of penalization so far have relied on the knowledge of the population parameters to select the optimal value of ψ (Meyer and Kirkpatrick 2010; Meyer *et al.* 2011), and results should therefore be regarded as optimistic. In practical applications we need to estimate ψ and are bound to do so with error, which affects the gains achievable.

This paper presents a simulation study investigating the performance of penalized estimation of genetic covariances matrices ($\mathbf{\Sigma}_G$) when tuning factors are estimated using cross-validation techniques or are determined by limiting the change in the likelihood due to penalization to a given value.

MATERIAL AND METHODS

Data were simulated for a paternal half-sib design, considering q=5 traits recorded on each of 10 progeny of s unrelated sires. Sample sizes considered were s=50, 100, 150, 200, 300, 400 and 1000. Population parameters were obtained combining 12 sets of heritabilities (A to L; see Table 1) with 5 scenarios for genetic (r_G) and residual (r_E) correlations (S1 to S5; see Table 2, $i \neq j$). This resulted in 60 different cases. Phenotypic variances were set to $\sigma^2_{P\,i}$=1 for S1 and $\sigma^2_{P\,i}$=1.5^{i-1} for S2 (i=1, q), and $\sigma^2_{P\,1}$=$\sigma^2_{P\,5}$=3, $\sigma^2_{P\,2}$=$\sigma^2_{P\,4}$=2 and $\sigma^2_{P\,3}$=1 for S3 to S5. Data were generated by sampling from appropriate multivariate Normal distributions, with 1000 replicates per case.

Penalty. Let $\log \mathcal{L}(\boldsymbol{\theta})$ denote the log likelihood for a given model of analysis with parameters $\boldsymbol{\theta}$. Penalized estimates were obtained by maximizing $\log \mathcal{L}_P(\boldsymbol{\theta}) = \log \mathcal{L}(\boldsymbol{\theta}) - \frac{1}{2}\,\psi\,\mathcal{P}(\boldsymbol{\theta})$, with ψ the tuning factor, and a quadratic penalty $\mathcal{P}(\boldsymbol{\theta})$ on the canonical eigenvalues λ_i, i.e. the eigenvalues of $\mathbf{\Sigma}_P^{-1}\mathbf{\Sigma}_G$ ($\mathbf{\Sigma}_P$: phenotypic covariance matrix). For $\mathbf{\Lambda}_1 = \text{Diag}\{\log(\hat{\lambda}_i)\}$ and $\mathbf{\Lambda}_2 = \text{Diag}\{\log(1-\hat{\lambda}_i)\}$, the penalty was

$$\mathcal{P}(\boldsymbol{\theta}) \propto \text{tr}\left(\mathbf{\Lambda}_1 - \overline{\lambda}_1\mathbf{I}\right)^2 + \text{tr}\left(\mathbf{\Lambda}_2 - \overline{\lambda}_2\mathbf{I}\right)^2 \qquad \text{with} \quad \overline{\lambda}_i = \text{tr}\left(\mathbf{\Lambda}_i\right)/q$$

Analyses. Restricted ML (REML) estimates of $\mathbf{\Sigma}_G$, $\hat{\mathbf{\Sigma}}^{\psi}_G$, and the residual covariance, $\hat{\mathbf{\Sigma}}^{\psi}_E$, were obtained as described by Meyer and Kirkpatrick (2010) for a range of values of ψ: 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

*AGBU is a joint venture of NSW Department of Industry and Investment and the University of New England

to 250 in steps of 2, 255 to 500 in steps of 5 and 510 to 1000 in steps of 10, 311 in total. The 'optimal' tuning factor, $\hat{\psi}$, was then determined using 10 different strategies:

Table 1. Population heritability values (×100)

A	B	C	D	E	F	G	H	I	J	K	L
40	50	60	70	90	70	80	90	20	30	50	60
40	45	50	55	50	70	30	30	20	25	20	10
40	40	40	40	30	40	30	10	20	20	15	10
40	35	30	25	20	10	30	10	20	15	10	10
40	30	20	10	10	10	30	10	20	10	5	10

Using population values. 1) For known $\boldsymbol{\Sigma}_G$, $\hat{\psi}$ was chosen as the value which maximized the unpenalized likelihood $\log \mathcal{L}(\boldsymbol{\theta})^{\psi}$, for data represented by mean squares between and within sires constructed from the population values; see Meyer *et al.* (2011) for details. This was like sampling an infinite number of additional data sets for the same data structure (V∞). 2) Sampling one additional data set for validation and maximizing $\log \mathcal{L}(\boldsymbol{\theta})^{\psi}$ in these data (V1).

Using K−fold cross-validation. For each replicate, 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 $\hat{\boldsymbol{\Sigma}}_E^{\psi}$. Corresponding values for $\log \mathcal{L}(\boldsymbol{\theta})_i^{\psi}$ in the validation data were then obtained for all ψ, and $\hat{\psi}$ was chosen as the value for which the average, $\sum_{i=1}^{K} \log \mathcal{L}(\boldsymbol{\theta})_i^{\psi}/K$, was highest. Values of 3) K=2 (strategy CV2), 4) K=3 (CV3), 5) K=5 (CV5) and 6) K=10 (CV10) were considered.

Using the likelihood. Finally, $\hat{\psi}$ was chosen as the largest value for which $|\log \mathcal{L}(\boldsymbol{\theta})^{\psi} - \log \mathcal{L}(\boldsymbol{\theta})^{0}|$, i.e. the reduction in the unpenalized likelihood due to penalization from the maximum (at ψ=0) (sign ignored) did not exceed a selected value. Limits were chosen as the χ^2_{α} values ($\times\frac{1}{2}$) which would be employed in a likelihood ratio test of a single parameter with error probability α, i.e. 7) 0.82 for α=0.2 (strategy L20%), 8) 1.36 for α=0.1 (L10%), 9) 1.92 for α=0.05 (L5%) and 10) 2.51 for α=0.025 (L2.5%).

PRIAL. The effect of penalization on $\hat{\boldsymbol{\Sigma}}_G$ was summarized as percentage reduction in average loss

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

with $\hat{\boldsymbol{\Sigma}}_G^0$ and $\hat{\boldsymbol{\Sigma}}_G^{\hat{\psi}}$ the unpenalized and penalized estimates, respectively, $L_1\left(\boldsymbol{\Sigma}_G, \hat{\boldsymbol{\Sigma}}_G^{\psi}\right) = \text{tr}\left(\boldsymbol{\Sigma}_G^{-1}\hat{\boldsymbol{\Sigma}}_G^{\psi}\right) - \log|\boldsymbol{\Sigma}_G^{-1}\hat{\boldsymbol{\Sigma}}_G^{\psi}| - q$ the entropy loss in $\hat{\boldsymbol{\Sigma}}_G$, and $\bar{L}_1(\cdot)$ the average of $L_1(\cdot)$ over replicates.

Table 2. Correlations values

	$r_{G\,ij}$	$r_{E\,ij}$
S1	0	0
S2	0.8	0
S3	$0.6^{\lvert i-j\rvert}$	$-0.4^{\lvert i-j\rvert}$+0.5
S4	$-0.8^{\lvert i-j\rvert}$+0.02	$-0.4^{\lvert i-j\rvert}$+0.5
S5	$-1^{i}\,0.05\,j$+0.5	$-1^{j}\,0.1\,i$+0.2

RESULTS

Mean PRIAL values across the 60 cases for the different strategies are summarized in Table 3. Values declined with sample size, and were highest for strategy V∞. For the balanced case considered here, V∞ yielded the same results as minimizing the sum of the entropy losses in $\hat{\boldsymbol{\Sigma}}_G$ and $\hat{\boldsymbol{\Sigma}}_E$. Simulating a single validation set only in strategy V1 introduced considerable sampling error which reduced mean PRIAL values by 8 to 10% compared to V∞.

Examining regularization of covariance matrices via thresholding, Rothman *et al.* (2009) commented that cross-validation yielded similar results than strategy V1. However, in our case, mean PRIAL values obtained using cross-validation to determine $\hat{\psi}$ were but consistently lower, only slightly so for small samples but increasingly as sample size increased. Somewhat surprisingly, the PRIAL achieved using cross-validation decreased with the number of folds considered, K. As illustrated in Figure 1, 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. Our expectation was that a small training set (low K) would result in a $\hat{\psi}$ which was somewhat too large as it pertained to the sample

Table 3. Mean PRIAL for estimates of Σ_G

s=	50	100	150	200	300	400	1000
V∞	72.1	72.9	72.1	71.6	68.2	65.4	55.4
V1	63.7	63.7	63.2	62.9	59.3	55.2	47.0
CV2	62.3	61.8	60.5	58.0	52.6	47.5	30.5
CV3	61.3	60.7	58.2	54.4	48.9	43.6	27.2
CV5	59.7	58.1	55.5	51.5	44.7	39.9	23.6
CV10	57.7	55.3	52.1	47.4	40.6	34.9	21.7
L20%	69.5	69.3	67.8	66.4	62.2	59.0	46.5
L10%	71.4	70.7	68.8	67.4	62.8	59.2	45.5
L5%	71.3	70.2	68.1	66.6	61.6	57.6	42.7
L2.5%	70.3	69.0	66.6	65.0	59.7	55.2	39.1

Table 4. Mean tuning factors (S2 to S5)

s=	50	100	150	200	300	400	1000
V∞	2.6	1.9	1.8	1.7	1.7	1.7	1.8
V1	7.7	3.2	2.5	2.4	2.2	2.2	2.4
CV2	17.8	7.4	3.8	2.6	2.1	1.8	1.6
CV3	15.5	4.8	2.8	2.2	1.8	1.7	1.5
CV5	13.9	4.3	2.4	1.9	1.7	1.6	1.5
CV10	12.4	3.5	2.2	1.8	1.6	1.5	1.4
L20%	0.5	0.7	0.8	0.9	1.2	1.3	2.1
L10%	0.9	1.2	1.4	1.6	1.9	2.1	3.0
L5%	1.5	1.8	2.1	2.2	2.6	2.9	4.0
L2.5%	2.4	2.4	2.7	2.9	3.3	3.6	4.9

size of the subset, and that the number of replications for larger K would off-set potential inabilities to ascertain optimal values for ψ due to the limited size of the validation set. Mean tuning factors for scenarios S2 to S5 are shown in Table 4. As expected, at small sample sizes, cross-validation resulted in substantially larger estimates $\hat{\psi}$ than the strategies exploiting knowledge of the population parameters, i.e. the reduction in PRIAL was due to excessive penalization. S1 was excluded from these averages as it included several cases (A, B, I and J) for which the optimal tuning factor was very large. While the pattern of PRIAL values across strategies for S1 was comparable to that for the other population correlation values, cross-validation for these cases resulted in underestimates of $\hat{\psi}$. If S1 had been included in the averages shown in Table 4, results would have been distorted due the magnitude of $\hat{\psi}$ for these special cases.

In part, large values of $\hat{\psi}$ for small sample sizes could be attributed to a few cases where the cross-validation procedure failed and selected overly large values. For instance, disregarding any replicates with a $\hat{\psi}$ more than 5 standard deviations above the mean (within case), reduced values for CV2 to 12.8, 4.9, 3.1 and 2.4 for s=50 to s=200, but had virtually no effect on the average $\hat{\psi}$ for larger sample sizes. This may partially explain the relative small difference in PRIAL obtained from CV2 or CV3 and V1 for the smaller samples. Other reasons may be that the variation in $\hat{\psi}$ in individual replicates has relatively little effect on the average loss in penalized estimates of Σ_G and that, for relatively large entropy losses of unpenalized estimates at small s, these translate to small changes in PRIAL only. While inflation in estimates $\hat{\psi}$ from cross-validation decreased with the number of folds considered, mean PRIAL values decreased as K increased. Reasons for this are not clear. Results suggest that repetition of $K-$fold cross-validation for small K is advantageous over larger K at similar computational expense.

Choosing $\hat{\psi}$ on the basis of the reduction in the (unpenalized) likelihood due to penalizing estimates proved highly successful. Except for the largest sample sizes, this resulted in lower values of $\hat{\psi}$ and thus a milder degree of penalization. Nevertheless it outperformed cross-validation in all cases. For instance, strategy L5% corresponds to a change in a single parameter estimate which would not be considered significant at a 5% error level. This yielded mean PRIAL values higher than for strategy V1 for samples with 300 or less sires. Results suggest that a limit based on a χ^2_α value for $\alpha = 0.05$ is appropriate for the smaller sample sizes, while an increase in α

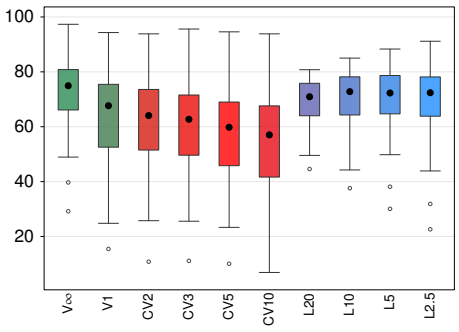

Figure 1. PRIAL for $\hat{\Sigma}_G$ for s=100.

(and thus decrease in the cut-off value) to 0.1 or 0.2 appeared advantageous for larger data sets.

Table 5 summarizes correlations between entropy losses in estimates of $\mathbf{\Sigma}_G$ (i.e. $L_1(\mathbf{\Sigma}_G, \hat{\mathbf{\Sigma}}_G^{\hat{\psi}})$) from V$\infty$ and the other strategies. Values given were calculated across replicates within each of the 60 cases and pooled across cases. Correlations from 0.54 for s=50 to 0.30 for s=1000 between strategies V∞ and V1 again emphasize the effect of sampling variation on estimates of the tuning factor. As to be expected from the means in PRIAL, corresponding values for the cross-validation strategies were low, ranging from 0.50 to 0.06. However, calculating correlations across cases, these rose to 0.78 to 0.50, indicating that these strategies will, on average, determine $\hat{\psi}$ adequately but that there are substantial effects of errors, especially for small validation sets (K large). Conversely, correlations between the likelihood based strategies and V∞ were high throughout, ranging from 0.77 to 0.80. This suggests that a likelihood based choice can determine the optimal tuning factor well.

Table 5. Correlations (×100) between $L_1(\mathbf{\Sigma}_G, \hat{\mathbf{\Sigma}}_G^{\hat{\psi}})$ from V∞ and other strategies

s=	50	100	150	200	300	400	1000
V1	54	46	45	42	38	30	30
CV2	50	36	36	27	25	20	11
CV3	45	31	28	20	19	16	9
CV5	39	26	23	16	16	13	6
CV10	36	23	20	13	13	11	6
L20%	89	87	86	83	84	82	83
L10%	90	88	86	83	83	81	82
L5%	89	86	84	80	80	78	79
L2.5%	87	83	81	77	77	74	76

DISCUSSION

Penalized estimation of genetic parameters is appealing for scenarios where sample sizes are small, regardless of any increased computational demands. Substantial reductions in average loss, i.e. the deviation of estimates from population values can be achieved. However, this relies on the appropriate selection of a tuning factor. Cross-validation is widely advocated as a technique to determine this from the data at hand. Yet, it is laborious and subject to substantial error in determining $\hat{\psi}$. 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 appears to become more detrimental as sample size increases. A particular problem with cross-validation for data with a family structure is that of representative sampling of data subsets. In our simulation setting, assigning whole sire families to individual folds appeared 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.

Fortunately, choice of $\hat{\psi}$ based on the change in likelihood can yield penalized estimates closely related to those which would be obtained if population values were known. As demonstrated, these are at least 'as good' as those obtained using cross-validation. The maximum change in likelihood should be chosen so as to yield a relatively mild penalty and taking account of the sample size and number of traits considered. Further work should evaluate suitable limits for a range of other scenarios.

CONCLUSIONS

Penalized maximum likelihood estimation of genetic parameters can result in estimates with substantially reduced sampling errors. Likelihood based selection of the tuning parameter required is recommended as a simple and effective strategy.

REFERENCES

Meyer K. (2011) *Proc. Ass. Advan. Anim. Breed. Genet.* **19**:71.
Meyer K. and Kirkpatrick M. (2010) *Genetics* **185**:1097.
Meyer K., Kirkpatrick M. and Gianola D. (2011) *Proc. Ass. Advan. Anim. Breed. Genet.* **19**:87.
Rothman A.J., Levina E. and Zhu J. (2009) *J. Amer. Stat. Ass.* **104**:177.