Proceedings, 10th World Congress of Genetics Applied to Livestock Production

Improving REML estimates of genetic parameters through penalties on correlation matrices

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

ABSTRACT: Penalized REML estimation can substantially reduce sampling variation in estimates of covariance matrices, and yield estimates of genetic parameters closer to population values than standard analyses. A number of suitable penalties based on prior distributions of correlation matrices from the Bayesian literature are described, and a simulation study is presented demonstrating their efficacy. Results show that reductions of 'loss' in estimates of the genetic covariance matrix, a conglomerate of sampling variance and bias, well over 50% are readily obtained for multivariate analyses of small samples. Default settings for a mild degree of penalization are proposed, which make such analyses suitable for routine use without increasing computational requirements.
Keywords:
Estimation of genetic parameters
Penalized REML
Improved estimator

Introduction

Estimates of genetic parameters are afflicted by sampling variation, increasingly so with the number of traits considered. A sobering but realistic view is that, "Few datasets, whether from livestock, laboratory or natural populations, are of sufficient size to obtain useful estimates of many genetic parameters" (Hill, 2010). Fortunately, estimates can be 'improved' – i.e. modified so that, on average, they are closer to the population values – by utilising additional knowledge. This is inherent in Bayesian analyses through the assumed prior distributions. For REML estimation, imposing a penalty proportional to the logarithmic value of the prior densities on the likelihood can yield analogous benefits.

Some Bayesian approaches to estimating covariance matrices decompose these into variances and correlations with separate priors, generally shrinking correlations towards zero. Estimates of genetic correlations have been found to be often close to their phenotypic counterparts (Cheverud, 1988). In addition, the latter are generally estimated much more accurately. Hence, an alternative may be to 'borrow strength' by shrinking the genetic towards the phenotypic correlation matrix. This paper investigates the scope for REML estimation penalizing correlation matrices.

Priors and penalties

Let $\mathbf{R}$ denote the correlation matrix corresponding to covariance matrix $\mathbf{\Sigma}$ for q traits. Few families of density functions for $\mathbf{R}$ have been considered, with most priors encouraging shrinkage of $\mathbf{R}$ towards an identity matrix, $\mathbf{I}$. Barnard et al. (2000) suggest uniform distributions, either for individual correlations, r_{ij}, or jointly within the permissible space, i.e. $p(\mathbf{R}) \propto 1$. These can be formulated as parameter extended Inverse Wishart and Wishart priors with degrees of freedom $\nu = q + 1$ and scale matrix $\mathbf{I}$, while $\nu = 0$ for the latter yields Jeffreys' rule prior (Zhang et al., 2006). Chung et al. (2013) propose a weakly informative prior aimed specifically at penalized REML estimation assuming a Wishart prior for $\mathbf{\Sigma}$ with scale matrix $(2\omega)^{-1}\mathbf{I}$, incorporating information on individual correlations r_{ij}, believed to be close to a value of ρ_{ij}, by multiplying with the density from a Normal distribution, $N(\rho_{ij}, \tau^2)$, with suggested defaults of $\nu = q + 2$, $\omega \to 0$ and $\tau = 0.25$.

More generally, unit diagonals and the requirement for $\mathbf{R}$ to be positive definite make handling r_{ij} challenging. This can be alleviated by parameterizing to partial auto-correlations (PAC), i.e. the correlations between traits i and j given the 'intervening' traits $i+1$ to $j-1$

$$\pi_{ij} = \left[r_{ij} - \mathbf{r}'_{1(i,j)}\mathbf{R}^{-1}_{3(i,j)}\mathbf{r}_{2(i,j)}\right] / \sqrt{R_{1(i,j)}R_{2(i,j)}} \qquad (1)$$

with $\mathbf{R}_{3(i,j)} = \{r_{kl}\}$, $\mathbf{r}_{1(i,j)} = \{r_{ik}\}$, $\mathbf{r}_{2(i,j)} = \{r_{jk}\}$ for $k, l = i+1$ to $j-1$, and $R_m = \left[1 - \mathbf{r}'_{m(i,j)}\mathbf{R}^{-1}_{3(i,j)}\mathbf{r}_{m(i,j)}\right]$. PAC fall in the interval $[-1, 1]$ and are unconstrained otherwise. Daniels and Pourahmadi (2009) assume independent shifted Beta priors for the PAC, $\pi_{ij} \propto \mathrm{B}(\alpha_{ij}, \beta_{ij})$. The authors show that $\alpha_{ij} = \beta_{ij} = 1 + (q - 1 - j + i)/2$ recovers the joint uniform prior for $\mathbf{R}$ of Barnard et al. (2000).

Gaskins et al. (2013) model the variance of individual PAC as a function of lag between traits, $\mathrm{Var}(\pi_{ij}) = \varepsilon\,|j-i|^{-\gamma}$ with $\varepsilon \in [0, 1]$ and $\gamma \geq 0$, and use this to define $\alpha_{ij} = \beta_{ij} = (1/\mathrm{Var}(\pi_{ij}) - 1)/2$. Setting $\alpha_{ij} = \beta_{ij}$ implies $\mathrm{E}[\pi_{ij}] = 0$. A value $\tau_{ij} \neq 0$ is accommodated by setting $\beta_{ij} = T\alpha_{ij}$ with $T = (1 - \tau_{ij})/(1 + \tau_{ij})$, and α_{ij} as chosen or calculated as $\left[4T/\left(\mathrm{Var}(\pi_{ij})(T+1)^2\right) - 1\right]/(T+1)$.

Penalties. The density functions suggested give penalties based on Jeffreys' rule prior, $\mathcal{P}_J$,

$$\mathcal{P}_J = \tfrac{q+1}{2} \log|\mathbf{R}| \qquad (2)$$

and assuming shifted Beta priors on PAC

$$\begin{aligned}\mathcal{P}_\pi = &\textstyle\sum_{i=1}^{q}\sum_{j=i+1}^{q} \log\Gamma(\alpha_{ij}) + \log\Gamma(\beta_{ij}) \\ &- \log\Gamma(\alpha_{ij} + \beta_{ij}) + (\alpha_{ij} + \beta_{ij} - 1)\log(2) \\ &- (\alpha_{ij} - 1)\log(1 + \pi_{ij}) - (\beta_{ij} - 1)\log(1 - \pi_{ij})\end{aligned} \qquad (3)$$

with $\Gamma(\cdot)$ denoting the Gamma function.

Simulation study

Data. Records for $q = 9$ traits were sampled from multivariate normal distributions, assuming a balanced paternal half-sib design comprised of $s = 100$ or 1000 sires families of size 10. Population values for 60 cases, selected to represent an extensive range of possible (including many 'difficult')

scenarios with coefficients of variation for canonical eigenvalues from 0 to 170%, were obtained by combining 12 sets of heritabilities with 5 correlation structures, referred to as C1 to C5. Let $r_{G\,ij}$ and $r_{E\,ij}$ ($i \neq j$) denote the genetic and residual correlations between traits i and j. Values were $r_{G\,ij} = r_{E\,ij} = 0$ for C1, $r_{G\,ij} = 0.5$ and $r_{E\,ij} = 0.3$ for C2, $r_{G\,ij} = 0.7^{|j-i|}$ and $r_{E\,ij} = -1^{j}0.05i + 0.2$ for C3, $r_{G\,ij} = -0.8^{|j-i|} + 0.02i$ and $r_{E\,ij} = -0.2^{|j-i|} + 0.5$ for C4, and $r_{G\,ij} = r_{E\,ij} = 0.7$ if $i, j \in [3, 7]$ and $r_{G\,ij} = r_{E\,ij} = 0.3$ otherwise for C5.

Analyses. REML estimates of genetic ($\boldsymbol{\Sigma}_G$) and residual ($\boldsymbol{\Sigma}_E$) covariance matrices for each sample were obtained fitting a simple animal model with means as the only fixed effects, for different types of penalties:

1. $\mathcal{P}_J$ as given in (2),
2. $\mathcal{P}_B$ invoking the joint uniform of Barnard et al. (2000), using the unconstrained parameterization through PAC,
3. $\mathcal{P}_D$ assuming the shifted Beta prior for PAC, shown in (3), with fixed values for $\alpha = 2, \ldots, 16$, and
4. $\mathcal{P}_G$ as 3., but modelling scale parameters depending on lag between traits, considering values of $\varepsilon = 0.05, 0.1, 0.2$ and $\gamma = 0, 0.5, 0.8, 1, 1.2$.

All were examined imposing a penalty on genetic correlations only and on both genetic and residual values, denoted by superscript '+'. Both $\mathcal{P}_D$ and $\mathcal{P}_G$ were considered shrinking π_{ij} towards zero and towards values τ_{ij} equal to phenotypic PAC. A total of 500 replicates were carried out for each case.

Summary statistics. For each sample, the loss in estimates was determined as (for $X = G, E$ and P)

$$L_1\left(\boldsymbol{\Sigma}_X, \hat{\boldsymbol{\Sigma}}_X\right) = \text{tr}\left(\boldsymbol{\Sigma}_X^{-1}\hat{\boldsymbol{\Sigma}}_X\right) - \log\left|\boldsymbol{\Sigma}_X^{-1}\hat{\boldsymbol{\Sigma}}_X\right| - q \qquad (4)$$

with $\boldsymbol{\Sigma}_X$ the matrix of population values, $\hat{\boldsymbol{\Sigma}}_X$ the corresponding estimate, and $\boldsymbol{\Sigma}_P = \boldsymbol{\Sigma}_G + \boldsymbol{\Sigma}_E$. The **P**ercentage **R**eduction **I**n **A**verage **L**oss due to penalization was then evaluated as

$$\text{PRIAL} = 100\left[1 - \bar{L}_1\left(\boldsymbol{\Sigma}_X, \hat{\boldsymbol{\Sigma}}_X^{\nu}\right)/\bar{L}_1\left(\boldsymbol{\Sigma}_X, \hat{\boldsymbol{\Sigma}}_X^{0}\right)\right] \qquad (5)$$

with $\hat{\boldsymbol{\Sigma}}_X^{\nu}$ and $\hat{\boldsymbol{\Sigma}}_X^{0}$ the penalized and unpenalized estimates of $\boldsymbol{\Sigma}_X$, and $\bar{L}_1(\cdot)$ the average loss over replicates. In addition, the mean reduction in unpenalized likelihood due to penalization (from its maximum for unpenalized estimates), $\Delta\mathcal{L}$, was calculated.

Results

Mean PRIAL values and $\Delta\mathcal{L}$ across the 60 cases for selected analyses are summarized in Table 1. Overall, penalization yielded dramatically 'better' estimates than standard REML, especially for the smaller sample size, accompanied in general by modest reductions in likelihood. For $q = 9$ and 90 covariance components, for $\Delta\mathcal{L}$ (absolute value) to be significant at 5% error probability, it would have needed to exceed a much larger value of 56.57.

Means hide marked differences for individual cases – clearly penalties performed the better the closer the prior matched the population parameters. Milder penalties tended to achieve most of the benefits feasible whilst reducing the risk of distorting estimates of phenotypic covariances due to over-shrinkage. Figure 1 shows the distribution of PRIAL across cases for increasing values of $\alpha = \beta$ for penalty $\mathcal{P}_D^+$. While PRIAL values for $\boldsymbol{\Sigma}_G$ increased with α up to about 10, the higher values

Table 1. Mean PRIAL for selected penalties.

$\mathcal{P}^*$	$\nu^\dagger$	100 sires				1000 sires			
		$\boldsymbol{\Sigma}_G$	$\boldsymbol{\Sigma}_E$	$\boldsymbol{\Sigma}_P$	$\Delta\mathcal{L}^\ddagger$	$\boldsymbol{\Sigma}_G$	$\boldsymbol{\Sigma}_E$	$\boldsymbol{\Sigma}_P$	$\Delta\mathcal{L}$
Shrinking towards zero									
$\mathcal{P}_B$	–	55	28	1	-4.94	34	7	0	-1.23
$\mathcal{P}_B^+$	–	57	50	1	-4.66	34	12	0	-1.21
$\mathcal{P}_J$	–	64	35	3	-9.05	39	9	0	-2.34
$\mathcal{P}_J^+$	–	68	57	2	-8.83	40	16	0	-2.31
$\mathcal{P}_D$	2	58	20	1	-1.69	32	4	0	-0.32
$\mathcal{P}_D^+$	2	59	45	2	-1.76	32	11	0	-0.33
$\mathcal{P}_D^+$	4	66	54	2	-5.41	38	15	0	-1.23
$\mathcal{P}_D^+$	6	68	57	2	-8.82	39	16	0	-2.30
$\mathcal{P}_G^{\,1}$	5	64	36	4	-19.79	37	3	0	-11.58
$\mathcal{P}_G^{\,1}$	10	71	39	4	-8.03	46	15	1	-4.32
$\mathcal{P}_G^{\,1}$	20	72	36	4	-6.93	45	10	1	-1.67
$\mathcal{P}_G^{+1}$	5	67	51	-1	-23.73	43	10	-2	-11.11
$\mathcal{P}_G^{+1}$	10	72	61	5	-13.77	48	22	1	-4.45
$\mathcal{P}_G^{+1}$	20	74	61	5	-7.57	46	21	1	-1.79
Shrinking towards phenotypic correlations									
$\mathcal{P}_B^+$	–	48	45	0	-2.93	29	8	0	-1.37
$\mathcal{P}_D$	2	41	5	-1	-1.72	26	1	0	-0.41
$\mathcal{P}_D^+$	2	43	39	0	-1.81	26	9	0	-0.46
$\mathcal{P}_D^+$	4	60	51	1	-3.83	34	11	0	-1.40
$\mathcal{P}_G^{+1}$	5	67	61	2	-13.88	41	13	0	-8.54
$\mathcal{P}_G^{+0}$	10	62	54	1	-4.53	35	13	0	-1.66
$\mathcal{P}_G^{+0.5}$	10	67	60	2	-7.16	41	18	0	-2.49
$\mathcal{P}_G^{+0.8}$	10	68	62	2	-8.83	43	20	1	-3.35
$\mathcal{P}_G^{+1}$	10	68	62	2	-9.88	45	21	1	-4.11
$\mathcal{P}_G^{+1.2}$	10	69	63	2	-10.89	46	21	1	-5.12
$\mathcal{P}_G^{+1}$	20	66	59	2	-6.41	44	21	1	-1.82

* Penalty; superscripts give value of γ for $\mathcal{P}_G$ † Scale parameter α for $\mathcal{P}_D$, $\varepsilon \times 100$ for $\mathcal{P}_G$ ‡ Mean change in unpenalized log likelihood

Figure 1. PRIAL due to penalty $\mathcal{P}_D^+$ for $\alpha = \beta$

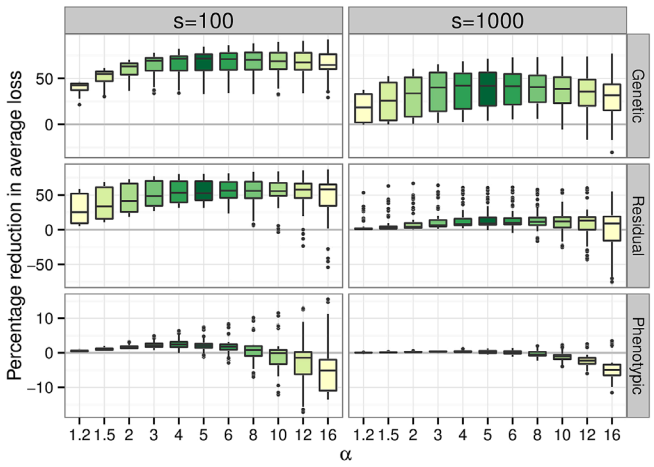
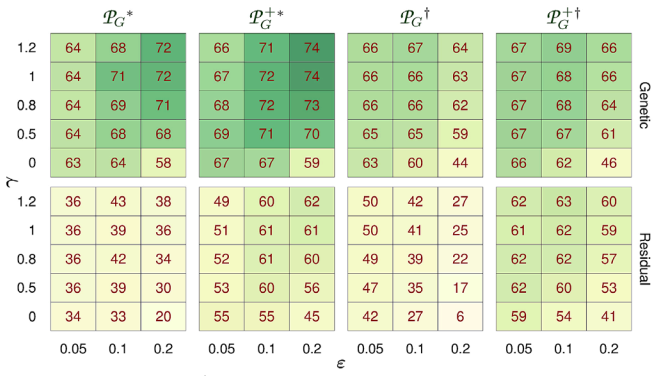


Figure 2. Mean PRIAL from penalties $\mathcal{P}_G$ and $\mathcal{P}_G^+$



* Shrinkage towards 0 † Shrinkage towards phenotypic PAC

resulted in larger spreads in PRIAL for $\mathbf{\Sigma}_E$ and thus $\mathbf{\Sigma}_P$, and more cases for which estimates for the latter were worse than their unpenalized counterparts. Results suggest that a default value of $\alpha = 2$ to 4 may be a sensible choice.

Penalties $\mathcal{P}_G$ yielded the highest PRIAL values but required specification of two parameters, ε and γ. As shown in Figure 2, means depended little on γ, i.e. more aggressive shrinkage of PAC with increasing lag was not as important for our population values as for the longitudinal data for which this prior was originally proposed. A value of $\varepsilon \approx 0.1$, equivalent to assuming intermediate variability of PAC, may be a suitable default value.

Penalties $\mathcal{P}_B$ and $\mathcal{P}_J$ lacked parameters to regulate their strength and thus did not require any choices. Their performance was similar to some of the penalties on PAC, albeit often accompanied by somewhat larger reductions in likelihood. In particular, the simple penalty based on Jeffreys' rule prior – often flagged as an 'improper' prior – resulted in remarkably good reductions in loss.

As observed previously for penalties on covariance matrices (Meyer, 2011), penalizing both $r_{G\,ij}$ and $r_{E\,ij}$ increased PRIAL for $\mathbf{\Sigma}_E$ without detrimental effects on $\mathbf{\Sigma}_G$. This held especially for penalties which did not provide any feedback mechanism on total covariances, as provided, to some extent at least, by shrinkage towards phenotypic correlations. On the whole, no distinct advantage of shrinking correlations towards phenotypic values rather than zero was apparent in terms of mean PRIAL values. In part, this could be attributed to very high PRIAL for cases with matching population values ($r_{G\,ij} = r_{E\,ij} = 0$) for the latter. However, shrinkage towards phenotypic PAC generally resulted in less over-penalization and smaller reductions in the unpenalized likelihood, which presumably equates to less bias.

Discussion

Excessive sampling variation is the bane of multivariate analyses. Penalized REML estimation can substantially reduce this and yield values closer to the population parameters than unpenalized estimates. While reductions in loss are generally highest for smaller samples, very worthwhile improvements can be obtained for larger data sets, especially as the number of traits of interest rises. This can impact dramatically on livestock improvement schemes, for instance, by increasing the achieved response to index selection through more appropriate index weights.

We have demonstrated that penalties on correlation matrices based on assumptions for prior distributions proposed in the Bayesian literature provide suitable functions. An advantage of correlations (including partial auto-correlations) is that they fall in a defined interval and that it is thus feasible to identify default values, e.g. for their variance, which can be used to define relatively mild penalties – conceptually similar to the use of weakly informative priors in Bayesian estimation (Gelman, 2006). This reduces the chance of over-penalizing whilst achieving a large proportion of benefits. Furthermore, use of such default values does not require laborious additional computations, making routine use straightforward. Indeed, mildly penalized likelihood functions often tend to have better defined maxima which may aid convergence in iterative solution schemes. Although direct estimation of the parameters defining strength of penalization is possible in principle, attempts to do so (not shown) at best generally did not proof to be sufficiently advantageous to justify the effort.

As emphasized, efficacy of penalized estimation depends on how well the underlying priors agree with true values. Even when conformity was relatively poor, improvements were obtained in almost all cases for mild penalties, especially for matrices with small eigenvalues. While this may not hold universally, it is a good indication that the method proposed is beneficial for a wide range of scenarios commonly encountered in quantitative genetics.

Conclusions

Penalized REML estimation can dramatically improve estimates of genetic parameters at little 'cost'. Penalties on correlations offer a framework which allows default settings to be specified. We envisage this procedure becoming a future standard for multivariate analyses.

Literature Cited

Barnard, J., McCulloch, R., and Meng, X. (2000). Stat. Sin. 10:1281–1312.

Cheverud, J. M. (1988). Evolution 42:958–968.

Chung, Y., Gelman, A., Rabe-Hesketh, S., et al. (2013). Technical report, Columbia University. `www.stat.columbia.edu/~gelman/research/unpublished/ms.pdf`.

Daniels, M. J., and Pourahmadi, M. (2009). J. Multiv. Anal. 100:2352–2363.

Gaskins, J. T., Daniels, M. J., and Marcus, B. H. (2013). J. Comp. Graph. Stat., published online 30/10/2013.

Gelman, A. (2006). Bayesian Anal. 1:515–533.

Hill, W. G. (2010). Phil. Trans. R. Soc. B 365:73–85.

Meyer, K. (2011). Genet. Sel. Evol. 43:39.

Zhang, X., Boscardin, W. J., and Belin, T. R. (2006). J. Comp. Graph. Stat. 15:880–896.

*AGBU is a joint venture of the University of New England and the NSW Department of Primary Industries