Bayesian Estimation of Within and Across Country Genetic Parameters for MACE

Just Jensen and Per Madsen
Danish Institute of Agricultural Sciences, Dept. of Animal Breeding and Genetics, Research Centre Foulum, P.O. Box 50, DK-8830 Tjele Denmark

Introduction

In international genetic evaluation of dairy bulls using MACE, information from each country or region is regarded as a different trait. This leads to a requirement of knowing variances and covariances among trait expressions in different countries. If only one trait is included per country no residual covariances exists. However, it is often desired to include several traits per country and in these cases also residual covariances between traits recorded within country must be known. In practice genetic and residual variances and covariances are never known without uncertainty. However, current practice has been to estimate these parameters and then proceed performing genetic evaluations as if the estimated parameters were the true values.

When new breeds are initiating international genetic evaluations, or new countries are added to existing evaluations, the genetic ties between different countries or regions may be weak. An example in question is the desire to conduct international comparisons for conformation traits in Ayrshires (Klei & Lawlor, 2001). In such cases information available is inadequate to estimate across country genetic (co) variances using REML methodology.

For conformation there are usually several traits of interest recorded per country so that within country residual covariances must be estimated. Madsen *et al.* (2000) presented methodology to estimate such within country residual covariances. Their methods furthermore estimated the genetic variances and covariances based on all available information and not only on selected well-connected subsets (Sigurdsson *et al.*, 1998; Klei & Weigel, 1998). In addition the method of Madsen *et al.* (2000) allowed the estimation of standard errors of across country genetic correlations. Large standard errors on a genetic correlation between two countries indicates weak genetic ties between the countries involved and, therefore, the standard error provides a summary of both direct and indirect genetic ties between the countries involved.

The problems mentioned above have lead to the desire of developing a method where prior information on variances and covariances can be introduced in the estimation of (co) variance components for MACE and a Bayesian framework is then a natural choice. Prior information on variances and covariances for use in international genetic evaluation can come from either similar traits recorded in other breeds or from the same traits recorded in ”similar” countries. The two approaches can of course be combined.

The purpose of this paper is to present a Bayesian method for the estimation of across country genetic variances and covariances and within country residual variances and covariances for use in MACE. The method presented will also include Bayesian prediction of breeding values for all bulls on all national scales, where the posterior uncertainty about the variances and covariances is taken into account.

Methods

Model

The general model for MACE as presented by Schaeffer & Zhang, (1993) is:

$$\mathbf{y} = \mathbf{Cc} + \mathbf{ZQg} + \mathbf{Zs} + \mathbf{e}\,, \qquad [1]$$

where **y** is a vector of deregressed national proofs or daughter yield deviations (DYDs), **c** is a vector of country effects, **g** is a vector of phantom parent group effects, **s** is a vector of random bull effects and **e** is a vector of random residuals.

Using the same formulation as Madsen *et al.* (2000), where phantom parent groups are assumed to be defined within country of origin, the model can be written as:

$$\mathbf{y} = \mathbf{ZQf} + \mathbf{Zs} + \mathbf{e}, \qquad [2]$$

where **f** is a vector of phantom parent group + country effects.

The conditional distribution of **y** given the location parameters **f** and **s** is:

$$\mathbf{y} \mid \mathbf{f}, \mathbf{s}, \mathbf{R}_o \sim \mathrm{N}\left[\mathbf{ZQf} + \mathbf{Zs}, \mathbf{R}\right], \qquad [3]$$

where $\mathbf{R}_0$ is a matrix of residual variances and covariances. The covariances are only defined when both traits involved are recorded in the same country; otherwise the corresponding element is set to zero.

The matrix **R** is block diagonal with each block corresponding to the traits a bull have recorded in each country where he have data. In general the residual variance of records for use in MACE is heterogeneous. This is because the number of daughters per bulls varies considerably. Therefore the degree of heterogeneity is known up to proportionality. Often this can be solved using a weighted analysis:

$$\mathbf{R}_{kc} = \mathbf{W}^{-1}\mathbf{R}_0\mathbf{W}^{-T} \qquad [4]$$

where $\mathbf{R}_{kc}$ is the diagonal block in **R** for bull k in country c, where all traits is recorded and **W** is a diagonal weighting matrix with $\sqrt{n_{ci}}$ on the i´th diagonal where n_{ci} is the number of records on trait i in country c. If some traits are missing in country c the corresponding row in $\mathbf{R}_k$ must be deleted.

If the number of records per trait varies in a country the elements in $\mathbf{R}_k$ must be computed as:

$$r_{c_{i,j}} = \sigma_{c_{i,j}} / \min(n_{c_i}, n_{c_j}) \qquad [5]$$

where $\sigma_{C_{i,j}}$ is the residual (co) variance between trait i and j in country c and n_{c_i} and n_{c_j} are the number of records for trait i and j in country c. This assumes that the trait with the smallest number of records is measured on a subset of the animals recorded for the other trait.

The distribution of **s** is assumed to be:

$$\mathbf{s} \mid \mathbf{G}_0 \sim \mathrm{N}\left[\mathbf{0}, \mathbf{G}_0 \otimes \mathbf{A}\right], \qquad [6]$$

where $\mathbf{G}_0$ is the matrix of variances and covariances due to bulls and **A** the additive relationship matrix.

Priors and full conditionals

Flat priors were assumed for the elements in **f**. For the covariance matrices we assumed inverse Wishart distributions as:

$$p(\mathbf{G}_0) = \mathrm{IW}_{Ct}\left[(\nu\mathbf{V}_a, \nu_a\right]$$

$$p(\mathbf{R}_0) = \mathrm{IW}_{\sum_{i=1}^{C} t_i}\left[(\nu\mathbf{V}_e, \nu_e\right]$$

where Ct is the product of the number of countries involved (C) times the number of traits included in the analysis (t) and $\sum_{i=0}^{C} t_i$ is a summation over the number of traits recorded per country. This yields the dimension of the bull and residual covariance matrices, respectively. $\mathbf{V}_{x\in\{a,e\}}$ and $\nu_{x\in\{a,e\}}$ denotes the scale parameter and the corresponding degree of belief parameter for the bull and residual covariance matrices. In this parameterisation the scale parameter (**V)** is a prior (co) variance matrix that have the same value as the mean of the prior distribution and, therefore, the elements are in then usual order of magnitude.

The method proposed is implemented using the Gibbs sampler. To do this we need the fully conditional distributions of all parameters in the model given all other parameters and the data. The fully conditional distributions of the location parameters are all normal (Jensen *et al.*, 1995).

The fully conditional distributions of the covariance matrices $\mathbf{G}_0$ and $\mathbf{R}_0$ are inverse Wishart:

$$p(\mathbf{G}_0 \mid \mathbf{y}, \mathbf{f}, \mathbf{s}, \mathbf{R}_0) = \mathrm{IW}\left[q + \nu_a, \nu_a\mathbf{V}_a + \mathbf{S}_a\right]$$

and

$$p(\mathbf{R}_0 \mid \mathbf{y}, \mathbf{f}, \mathbf{s}, \mathbf{e}, \mathbf{G}_0) = \mathrm{IW}\left[n + \nu_e, \nu_e\mathbf{V}_e + \mathbf{S}_e\right]$$

where $\mathbf{S}_a$ is the sum of squares and crossproducts of bull effects and $\mathbf{S}_e$ is the sum of squares and crossproducts of residuals after rescaling to the original scale by applying the weights. In this

formulation it is clearly seen that prior information is weighted with the information in the data.

Implementation of the Gibbs sampler

The method is implemented as an option in the general DMU-software package of Madsen and Jensen (1999), to improve mixing a blocking strategy was applied as described in Jensen *et al.* (1995). Convergence of the algorithm was checked using the method of batching which also estimate the effective posterior sample size. Furthermore we plotted traces of selected variables and computed lag-correlations. Posterior densities were estimated using the non-parametric approach of Scot (1992).

Application

The procedure was tested on a small bivariate example dataset. The data was from an experiment where 428 calves in 56 progeny groups. The progeny groups were split on two different feeding regimes and daily gain under the two regimes were regarded as two different traits. A bivariate model including fixed effects and random effects of sire and residual were used to estimate sire variances and co-variances. Due to the experimental design no residual covariance existed. The data was analysed using AI-REML (Jensen *et al.*, 1996) and using the new procedure where prior information was provided with varying degree of prior belief. In addition analysis, where residuals were assumed to have heterogeneous variance was also run. This was done in order to test the estimation of residual variance in models with heterogeneity known up to proportionality.

Results

Results are summarised in Table 1. It is clearly seen that conducting the analysis using a weak prior yields results similar to the REML estimates. The REML estimates are identical to posterior mode estimates under flat priors and thus there will be differences between Bayesian and REML estimates. The standard errors are large due to the small data set. By increasing the prior degree of belief the posterior means becomes closer to the prior and the posterior standard error of the estimates becomes smaller. Applying different weights for the residual variances rescales the posterior mean of residual variance exactly in proportion to the weights whereas the sire variance remains unchanged (results not shown).

Discussion

A limitation of the procedure developed is that full prior co-variance matrices with equal prior degree of belief on each element must be provided. It may be possible to develop a method with different prior degree of belief for each individual element. In practice, however, we do not believe this to be a limitation. The amount of prior belief that an analyst wants to put in prior information from other breeds or from other countries will always be based on a somewhat subjective judgement.

When the amount of information provided in the data accumulates this information will dominate the information in the prior if the prior degree of belief is not very large.

From the Gibbs sampler it is also possible to obtain posterior estimates of sire breeding values on all the national scales included in the model. The estimates and the posterior variance of these estimates will take the uncertainty about (co) variance parameters into account.

Conclusion

The test example presented in the previous paragraphs have demonstrated that we have developed and implemented a Gibbs sampler that can run multivariate models with heterogeneous residual variances and varying degree of prior information. The method is implemented in a general software package and, therefore, also can be applied to estimated (co) variance matrices for use in MACE.

Table 1. Results of analysis of example data using REML (*Asymp s.e).* and Bayesian posterior means with varying prior information (*posterior s.e.*).

Analysis	Sire effect			Residual	
	σ_1^2	σ_{12}	σ_2^2	σ_1^2	σ_2^2
REML	302.1	14.5	602.3	12811.0	7009.47
	(920.2	*461.3*	*575.4*	*1471.0*	*789.5)*
Bayesian					
Prior	300.0	10.0	600.0	12000.0	7000.0
PDB* = 5	291.6	11.3	551.4	12752.5	7043.5
	(190.2	*158.4*	*270.2*	*1231.6*	*711.3)*
PDB = 10	302.4	10.5	584.9	12718.2	7012.2
	(143.6	*126.6*	*227.2*	*1212.7*	*695.3)*
PDB = 100	301.6	9.3	601.1	12489.3	7001.2
	(44.2	*41.8*	*84.0*	*1005.1*	*567.9)*

*) PDB = Prior degree of belief