Bayesian Estimation of Parameters of a Structural Model for Genetic Covariances Between Milk Yield in Five Regions of the United States

R. Rekaya,* K. A. Weigel,† and D. Gianola†

*Department of Animal and Dairy Science,
University of Georgia, Athens, GA 30602

†Department of Dairy Science,
University of Wisconsin, Madison, WI 53706

ABSTRACT

Inference about genetic covariance matrices using multiple-trait models is often hindered by lack of information. This leads to imprecise estimates of genetic parameters and of breeding values. Patterns in a genetic covariance matrix can be exploited to reduce the number of parameters and to increase quality of inferences. A structural model for genetic covariances was developed and fitted to milk yield data in five regions of the United States. This was compared with a standard multiple-trait analysis using a deviance information criterion, a measure of quality of fit. Data consisted of 3,465,334 Holstein first-lactation records from daughters of 43,755 sires in five regions of the United States (Midwest, Northeast, Northwest, Southeast, Southwest). Parameters of the structural model included an intercept and effects of measures of genetic and of management similarity on genetic covariances. Genetic similarity depended on the number of records contributed by sires that were common to a pair of regions. Management similarity was a function of the quantity of concentrate used to produce 1000 kg of milk in each pair of regions. The structural and the multiple-trait models gave similar estimates of genetic covariances, but the number of parameters was 8 in the former vs. 15 in the latter. Hence, estimates of genetic covariances were more precise with the structural model. A deviance information criterion suggested a slight superiority of the multiple-trait model, although probably within sampling error. For both models, genetic correlations between milk yield in five regions of the United States were larger than 0.93.

(Key words: structural model, covariance component, genotype $\times$ environment interaction, Bayesian method)

Abbreviation key: DIC = deviance information criterion, GS = genetic similarity, MS = management similarity, MCMC = Markov chain Monte Carlo.

INTRODUCTION

Precise estimation of genetic covariances is difficult in multiple-trait models. Typically, this is due to insufficient statistical information. For example, when describing milk yield over several lactations, the number of genetic parameters is larger in a multi-trait model than in a univariate analysis. An obvious remedy is to use larger datasets, but this is not always practical, because collecting more data is time-consuming and expensive, especially in experimental settings. Also, increasing the number of records boosts computational demands and does not always lead to more information about genetic covariances, in a statistical sense. An alternative is to search for more parsimonious models.

In international genetic evaluation, for example, “traits” are defined according to country borders, e.g., milk yield in Italy and in the Netherlands is treated as a different trait. However, similarity in production systems between countries depends not only on geographical proximity but, importantly, on genetic ties and likeness of climatic conditions and management practices including quantity and quality of nutrient resources. Similarity between herds in different regions or countries can be used to describe genetic covariances. This would allow making better use of the available information, provided that the resulting model has fewer parameters and is not inferior in goodness of fit.

Information about climate, herd management practices, and genetic ties between regions (countries) can be exploited to explain patterns in a genetic covariance matrix using a structural model. If such model results in a lower degree of parameterization, it may lead to more precise inferences, because the same amount of information is used to estimate fewer parameters than in a standard multiple-trait analysis. In some animal breeding applications, patterns in covariance structure

Received May 21, 2001.

Accepted July 12, 2002.

Corresponding author: R. Rekaya; e-mail: rekaya@uga.edu.

are sometimes easy to identify (e.g., correlations between milk yields during lactation tend to be larger for adjacent test days).

The objective of this study was to present methodological developments for a structural model for genetic covariances, and to illustrate it via an application to first-lactation records of Holstein cows from five geographical regions of the United States.

MATERIAL AND METHODS

Multiple-Trait Model

Assume the following multiple-trait mixed linear model for t traits:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e} \quad [1]$$

where: $\mathbf{y} = \mathbf{y}_1', \mathbf{y}_2', \dots, \mathbf{y}_t'$ is the vector of records for the t traits; β and $\mathbf{u}$ are vectors of location effects and of additive genetic values, respectively; $\mathbf{X}$ and $\mathbf{Z}$ are the corresponding incidence matrices, and $\mathbf{e}$ is a residual vector. The data are assumed to be generated according to the stochastic process:

$$\mathbf{y}|\beta, \mathbf{u}, \mathbf{R} \sim N(\mathbf{X}\beta + \mathbf{Z}\mathbf{u}, \mathbf{R}) \quad [2]$$

where $\mathbf{R} = \mathbf{I} \otimes \mathbf{R}_0$, and $\mathbf{R}_0 = \{r_{ij}\}$ is a matrix of residual covariances between the t traits being analyzed.

In a Bayesian setting, prior information about unknown parameters or features of the model must be specified. For the location effects, the following vague normal prior will be used:

$$p(\beta|\lambda) \sim N(0, \mathbf{I}\lambda) \quad [3]$$

Here, λ is a positive, scalar hyper-parameter that is large enough so as to convey weak prior precision. For additive genetic values, the usual multiple-trait normal distribution will be adopted as prior:

$$\mathbf{u}|\Sigma_g \sim N(0, \mathbf{A} \otimes \Sigma_g) \quad [4]$$

where $\mathbf{A}$ is a matrix of additive relationships among animals, and $\Sigma_g = \{g_{ij}\}$ is a $t \times t$ positive-definite genetic covariance matrix, with typical element g_{ij} ($i, j=1, 2, \dots, t$). The priors for the elements of covariance matrices $\mathbf{R}_0$ and Σ_g were taken to be uniform (on $t \times t$ hyper-cubes), but bounded between minimum and maximum values, as follows:

$$p(\mathbf{R}_0) \propto k \quad r_{ij, \min} \leq r_{ij} \leq r_{ij, \max} \quad \forall \quad r_{ij} \quad [5]$$

$$p(\Sigma_g) \propto k \quad g_{ij, \min} \leq g_{ij} \leq g_{ij, \max} \quad \forall \quad g_{ij} \quad [6]$$

where k is a constant.

Prior independence between β , $\mathbf{u}$, $\mathbf{R}_0$, and Σ_g was assumed, but allowing for the dependence stated in [4]. Thus, the joint prior density has the form:

$$p(\beta, \mathbf{u}, \mathbf{R}_0, \Sigma_g|\lambda) \propto p(\beta|\lambda)p(\mathbf{u}|\Sigma_g) \quad [7]$$

with: $-\infty \leq (\beta, \mathbf{u}) \leq \infty$; $r_{ij, \min} \leq r_{ij} \leq r_{ij, \max}$ and $g_{ij, \min} \leq g_{ij} \leq g_{ij, \max}$.

In a standard multiple-trait analysis, all full conditional distributions needed for Bayesian implementation of a Markov Chain Monte Carlo (MCMC) algorithm, such as the Gibbs sampler, can be derived without difficulty (Varona et al., 1994; Wakefield et al., 1994; Wang et al., 1994).

Structural Model

The structural model for the genetic covariances involves an alternative parameterization of the multiple-trait model, aimed at explaining all possible $t(t-1)/2$ genetic covariances in terms of a fewer number of parameters. Here, the genetic covariance between each pair of traits is written as a linear function of a set of explanatory variables:

$$g_{ij} = \mathbf{k}_{ij}'\mathbf{b} \quad [8]$$

where $\mathbf{b}$ is a vector of effects explaining the genetic covariance between traits i and j , and $\mathbf{k}_{ij}$ is a corresponding row incidence vector of explanatory variables. In matrix notation, $\mathbf{g} = \mathbf{K}\mathbf{b}$, where $\mathbf{g}$ is a $t(t-1)/2 \times 1$ vector of genetic covariances, and $\mathbf{K}$ is a $t(t-1)/2 \times r$ incidence matrix, where r is the order of vector $\mathbf{b}$.

Under the new parameterization, the parameters of the structural model are: 1) β and $\mathbf{u}$, as defined before; 2) the diagonal elements of the genetic covariance matrix, g_{ii} ($i=1, 2, \dots, t$); 3) the r parameters ($\mathbf{b}$) of the structural model for the covariances, and 4) the residual variance-covariance matrix. The sampling model remains as defined previously in [2]. The joint posterior density, after assigning a prior distribution to each of the new parameters of the model, is:

$$p(\beta, \mathbf{u}, g_{11}, g_{22}, \dots, g_{tt}, \mathbf{b}, \mathbf{R}_0|\mathbf{y}, \lambda) \propto p(\mathbf{y}|\beta, \mathbf{u}, \mathbf{R}_0) \quad [9]$$

$$p(\beta|\lambda)p(\mathbf{R}_0) \times p(\mathbf{u}|\mathbf{b}, g_{11}, g_{22}, \dots, g_{tt})p(\mathbf{b})\prod_{i=1}^t p(g_{ii})$$

Priors for β and $\mathbf{u}$ are as in [3] and [4], respectively. Specific assumptions about the new parameters in [8] are:

$$p(\mathbf{b}) \sim N(0, \Phi) \quad [10]$$

where Φ is a diagonal matrix of order r , with sufficiently large elements so as to make the prior distribution of $\mathbf{b}$ diffuse (not very informative). Further,

$$p(g_{ii}|v_i, \sigma_i^2) \sim v_i \sigma_i^2 \chi_{v_i}^{-2} \text{ for } i=1,2,\dots,t \quad [11]$$

and, $p(\mathbf{R}_0) \propto k$ as in [5]. In [11], v_i and σ_i^2 are hyperparameters of a scaled inverted chi-square distribution.

The conditional posterior distributions of β , $\mathbf{u}$, and $\mathbf{R}_0$ are as in the standard multiple-trait analysis, that is, normal for location parameters and breeding values, and scaled inverted Wishart for the residual variance covariance matrix (Wang et al., 1994, Wakefield et al., 1994; Rekaya, 1997). After augmenting the posterior distribution with “missing” records, whenever this is appropriate, the conditional posterior distributions of parameters of the structural model ($\mathbf{b}$) and of the diagonal elements of the genetic covariance (g_{ii}) follow from [9] by treating all other parameters as constant. This leads to:

$$p(\mathbf{b}|\beta, \mathbf{u}, g_{11}, g_{22}, \dots, g_{tt}, \mathbf{R}_0, \mathbf{y}) \propto p(\mathbf{u}|\mathbf{b}, g_{11}, g_{22}, \dots, g_{tt})p(\mathbf{b}|\Phi) \quad [12]$$

and:

$$p(g_{ii}|\beta, \mathbf{u}, \mathbf{b}, g_{11}, g_{22}, \dots, g_{(i-1)(i-1)}, g_{(i+1)(i+1)}, \dots, g_{tt}, \mathbf{R}_0, \mathbf{y}) \propto p(\mathbf{u}|\mathbf{b}, g_{11}, g_{22}, \dots, g_{tt}) \times p(g_{ii}|\sigma_i^2) \quad [13]$$

for $i=1,2,\dots,t$

These conditional distributions do not involve the sampling model [2]. All information about parameters of the structural model comes through the additive genetic values, and from the prior distribution of the parameters. Unfortunately, at least with this parameterization, conditional distributions [12] and [13] are not in closed form. Therefore, draws from these distributions (e.g., with Gibbs sampling) cannot be done directly. However, a rejection scheme (Gilks, 1992) or a Metropolis algorithm (Metropolis et al., 1953) can be used instead.

Model Comparison

The standard multiple-trait model and the structural model can be compared using a deviance information criterion (**DIC**), defined by Spiegelhalter et al. (1998) as:

$$DIC = \bar{D} + p_D \quad [14]$$

$$\text{where: } \bar{D} = E_{\theta|y}[D(\theta)] \quad [15]$$

is the posterior expectation of the Bayesian deviance

$$D(\theta) = -2 \log p(y|\theta) \quad [16]$$

and p_D , the “effective number of parameters” (see Appendix), is:

$$p_D = \bar{D} - D(\bar{\theta}) \quad [17]$$

where $\bar{\theta}$ is the posterior mean of parameters intervening in the sampling model [2].

A smaller value of *DIC* indicates a better fit of the model.

Model comparison is often based only on a measure of fit, or “deviance” between observed and predicted data. However, increasing model complexity, through adding more parameters, typically leads to a better fit. A fair comparison between models should consider both *measures of fit* and *model complexity*. From [14], it is seen that *DIC* engulfs these two criteria, as $\bar{D}$ and p_D are measures of fit and complexity, respectively. In our context, the Bayesian deviance in [16] is:

$$D(\theta) = (\mathbf{y} - \mathbf{X}\beta - \mathbf{Zu})'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\beta - \mathbf{Zu}) + N \log|\mathbf{R}_0| + \text{Constant}$$

where N is the order of $\mathbf{y}_i$ ($i=1,2,\dots,t$) assuming that all traits are recorded in all individuals. Note that the Bayesian deviance decreases as the sum of squares of errors decreases, the latter constituting the usual measure of fit.

Asymptotically, the effective number of parameters is approximately equal to the difference between the expected deviance and the deviance evaluated at the mean value of the parameters, as indicated in [17]. In the Appendix, a straightforward derivation of *DIC* is given. Using [17], equation [14] can be rewritten as:

$$DIC = D(\bar{\theta}) + 2p_D$$

illustrating that *DIC* consists of a measure of goodness of fit $D(\bar{\theta})$, with a penalty ($2p_D$) for increasing model complexity. Hence, if two models have the same goodness of fit, the *DIC* criterion would favor the one with fewer parameters. The *DIC* is easy to estimate in a MCMC implementation. At each iteration of MCMC, the Bayesian deviance, $D(\theta)$, is computed. At the end of iteration, the mean of the Bayesian deviance, $\bar{D}$, as well as the mean value of the model parameters, $\bar{\theta}$, is calculated, leading directly to [14].

Application to Holstein Data from Five Regions of the United States

Data. The objective was to illustrate the structural model using Holstein production records. Observations

Table 1. Summary of data.

Region ¹	Herd-year-season	Sires	Cows
1	158,835	25,660	997,686
2	104,948	20,194	792,124
3	24,266	21,886	943,938
4	37,152	16,428	551,789
5	11,337	10,856	179,797
Total	336,538	43,755 ²	3,465,334

¹1 = Midwest, 2 = Northeast, 3 = Northwest, 4 = Southeast and 5 = Southwest.

²Unique sires.

were first-lactation milk yields of daughters of AI sires in five regions of the United States (Midwest = 1, Northeast = 2, Southeast = 3, Southwest = 4, and Northwest = 5). After edits (≥ 10 records per sire and $DIM \geq 275$) the data consisted of 3,465,334 lactations from 43,755 sires in 336,538 herd-year-seasons. Years represented were 1985 to 1997, and seasons were formed using 3-mo classes, starting by December-February, etc. Number of herd-year-season classes, sires and records, by region, are in Table 1. Most cows were from regions 1, 2, and 3. The pedigree file included 46,095 bulls.

Analyses. Two analyses were performed. The first one involved inferring genetic covariances between milk yield in the five regions using a standard multiple-trait approach, i.e., estimating directly the genetic variance-covariance matrix of order 5×5 . In the second analysis, several structural models were fitted to the genetic covariances and, hence, the genetic covariance matrix was estimated indirectly from the Gibbs sampling draws for g_{ii} ($i = 1, 2, \dots, t$) and $g_{ij} = \mathbf{k}_{ij}\mathbf{b}$. In both analyses, the following sire model was used:

$$y_{ijkl(r)} = HYS_i + AC_i + \gamma_r DIM_{ijkl(r)} + s_{k(r)} + e_{ijkl(r)}$$

where: $y_{ijkl(r)}$ is the milk yield of daughter l of sire k in region r ($r = 1, 2, \dots, 5$); HYS_i ($i = 1, 2, \dots, 336,538$) is the effect of herd-year-season i ; AC_j ($j = 1, 2, \dots, 4$) is the effect of age at calving j ; λ_r is a region-specific regression coefficient on days in milk (DIM); $s_{k(r)}$ is the transmitting ability of sire k in region r and $e_{ijkl(r)}$ is the residual term. Residual covariances between regions are zero, this leading to a simpler expression for $D(\theta)$.

Structural model for the genetic covariance. A three-parameter model was used to describe 10 genetic correlations (5 traits). Following equation [8], the additive genetic covariance (4 times the “sire” covariance) between regions i and j was described as:

$$g_{ij} = \mu + b_1 GS(i, j) + b_2 MS(i, j) \quad [18]$$

where μ is an intercept, common to all off diagonal elements of the covariance matrix; $GS(i, j)$ and $MS(i, j)$

Table 2. Genetic similarity (above diagonal) and management similarity (below diagonal) between the five geographical regions in the United States.

Regions ¹	1	2	3	4	5
1		0.44	0.42	0.38	0.41
2	0.85		0.41	0.41	0.40
3	1.03	1.21		0.37	0.46
4	0.93	1.09	1.10		0.35
5	1.13	1.34	0.91	1.22	

¹1 = Midwest, 2 = Northeast, 3 = Northwest, 4 = Southeast and 5 = Southwest.

are measures of genetic and management similarity between regions i and j , respectively, and b_1 and b_2 are regression coefficients. Following the notation in [8], $\mathbf{b} = (\mu, b_1, b_2)'$.

Definition of Genetic and Management Similarity

Genetic similarity between regions i and j was defined arbitrarily as the ratio between the number of daughters of bulls used in common in the two regions and the total number of daughters of all bulls used in the pair of regions:

$$GS(i, j) = \frac{\sum_{r=1}^2 \sum_{k=1}^{C(i,j)} ND_{kr}}{\sum_{r=1}^2 \sum_{k=1}^{T(i,j)} ND_{kr}}$$

where $C(i, j)$ is the number of bulls in common used in regions i and j , $T(i, j)$ is the total number of bulls used in the two regions, and ND_{kr} is the number of daughters of bull k in region r ($r = 1, 2$). As noted, the definition of genetic similarity is arbitrary; an alternative could have been $C(i, j) / T(i, j)$, without making reference to differential usage of sires.

Management similarity was arbitrarily defined as the ratio between the quantity of concentrate used to produce 1000 kg of milk (F^{1000}) in regions i and j .

$$MS(i, j) = \frac{F_i^{1000}}{F_j^{1000}} = MS(j, i)$$

To ensure symmetry of the genetic variance-covariance matrix, the restriction $MS(j, i) = MS(i, j)$ was imposed as noted. An alternative measure of similarity can be $MS(i, j) = MS(j, i) = |\log(\frac{F_i^{1000}}{F_j^{1000}})|$. Values of $GS(i, j)$ and $MS(i, j)$ for each pair of regions are in Table 2. The values ranged between 0.35 and 0.46 for genetic similarity, and 0.85 to 1.34 for management similarity.

Prior distributions. In the two analyses, we used $\lambda = 10^8$, so the prior distribution of β in [3] was

$$p(\beta) \sim N(0, 10^8)$$

In this application, the residual variance-covariance matrix ($\mathbf{R}_0$) was diagonal and we used the following proper, independent, priors for each of its diagonal elements:

$$r_{ii} \sim U[0, 2 \times 10^6] \text{ for } i=1,2,\dots,5.$$

In the standard multiple-trait analysis, the prior for each of the elements of the genetic variance covariance matrix was:

$$g_{ii} \sim U[g_{ij,\min}, g_{ij,\max}]$$

where $g_{ij,\max} = 10^6$ for all i and j , and

$$\begin{cases} g_{ij,\min} = 0 & \text{if } i = j \\ g_{ij,\min} = -10^6 & \text{if } i \neq j \end{cases}$$

In the structural model, the priors for g_{ii} ($i=1,2,\dots,5$), the diagonal elements of the genetic covariance matrix, were as for the standard multiple-trait approach. Each of the elements of $\mathbf{b}$ was assigned a normal prior with mean 0 and variance 10^8 . This takes an a priori stance of “no effect” of either genetic or management similarity on g_{ij} , but with a sufficiently large uncertainty, such that their effects range with high probability between -5×10^4 and 5×10^4 in the prior distribution.

Proper priors were used throughout, thus ensuring propriety of the posterior distribution. It must be noted that the structural model confers more flexibility in the assignment of priors to components of the genetic covariance matrix than the standard multiple-trait approach.

Implementation

For the standard multiple-trait model, all conditional posterior distributions needed for Gibbs sampling were in closed form, as pointed out earlier. In the structural model, the conditional posterior distributions of $\mathbf{b} = (\mu, b_1, b_2)'$ and those of the diagonal elements of the genetic (co)variance matrix, g_{ii} , ($i=1,2,\dots,5$), were not in a recognizable form. The $\mathbf{b}$ parameters were updated via the Metropolis algorithm (Metropolis et al., 1953), while the Metropolis-Hastings algorithm (Hastings, 1970) was used for updating the diagonal elements, g_{ii} . For the parameters in $\mathbf{b}$, a normal proposal density was used, and the updating was as follows. A random sample was drawn from the proposal distribution, centered at the current value of the Markov chain, and with an arbitrary variance such that an acceptance rate of 30 to 50% could be obtained, as recommended by Carlin

Table 3. Posterior means, standard deviations and quantiles of genetic covariances between milk yield (kg^2) in five regions: standard multiple-trait analysis.

Parameter ¹	Mean	SD	Quantiles		
			2.5%	50%	97.5%
g_{11}	90,365	1,572	87,434	90,330	93,413
g_{22}	101,710	1,873	97,933	101,740	105,100
g_{33}	90,188	1,702	87,217	90,045	94,045
g_{44}	83,712	2,156	79,328	83,707	87,998
g_{55}	108,430	2,371	103,020	108,570	112,690
g_{12}	93,679	1,503	90,642	93,657	96,487
g_{13}	86,621	1,346	83,766	86,662	89,101
g_{14}	83,415	1,478	80,454	83,470	86,213
g_{15}	95,030	1,402	92,326	94,979	97,907
g_{23}	89,891	1,423	87,399	89,735	93,089
g_{24}	89,104	1,399	86,340	89,146	91,761
g_{25}	100,750	1,684	97,378	100,810	103,990
g_{34}	82,981	1,456	80,076	82,960	85,826
g_{35}	96,165	1,574	92,972	96,215	99,135
g_{45}	88,584	1,487	85,919	88,504	91,777

¹ g_{ij} = genetic (co)variance between milk yields in regions i and j (1 = Midwest, 2 = Northeast, 3 = Northwest, 4 = Southeast, 5 = Southwest).

and Louis (1996). The variances of the proposal distributions for μ, b_1, b_2 were 900, 200, and 70, respectively. We calculated the ratio between the posterior density evaluated at the candidate value and at the current value. If the ratio was ≥ 1 , we accepted the candidate value. If the ratio was < 1 , the chain moved to the next step with probability equal to this ratio. For the diagonal elements of the genetic covariance matrix, a scaled inverted chi-square distribution was used as proposal. Its mean was equal to the current value of the Markov chain, and its variance was equal to the square root of the current value multiplied by 10. When draws for the genetic variance-covariance matrix were obtained (for both models), it was ensured that the matrices were positive-definite and consistent with the boundaries introduced by the prior distributions.

RESULTS

Table 3 presents estimates of genetic variances and covariances for milk yield between the five regions obtained with the standard multiple-trait analysis. Genetic covariances were positive and large, relative to their posterior standard deviations. Genetic variances were similar across regions, with slightly higher values for the Northeast and Southwest. Given the large number of records and the strong genetic ties between regions, the posterior precision of the genetic covariance components was high, as indicated by small posterior standard deviations and narrowness of the posterior distributions between the 2.5 and 97.5% quantiles. In fact, the posterior coefficients of variation ranged be-

Table 4. Posterior means and standard deviations of effects of parameters of the structural model for genetic covariances (kg²).

Parameter ¹	Mean	SD
μ	65,636	620
b_1	45,402	3463
b_2	5,184	935

¹ μ = intercept, b_1 = regression on genetic similarity and b_2 = regression on management similarity.

tween 1.7 and 2.6% for genetic variances in the Midwest and Southeast, respectively. Similar, high, levels of posterior precision were obtained for the genetic covariances.

Table 4 shows posterior means and standard deviations of effects of parameters (**b**) of the structural model for the genetic covariances. All posterior means were positive, with most of the posterior probability in the positive part of the real line. This indicates positive true genetic covariances, and an increase in covariance as genetic and management similarity increase. For example, using Tables 2 and 4, the expected genetic covariance between regions 1 and 2 would be given, after rounding, by: $65,636 + 45,402 \times 0.44 + 5,184 \times 0.85 = 90,019 \text{ kg}^2$.

Table 5 shows estimates of genetic variances and covariances between milk yield in different regions using the structural model. The latter values were computed using the draws for the parameters of the genetic covariance model, **b**, as mentioned above. We noted a lower genetic variance for Northeast and Southwest, and larger genetic variances for the Midwest, Northwest, and Southeast relative to estimates found using

Table 5. Posterior mean, standard deviations and quantiles of genetic covariances between milk yield (kg²) in five regions: structural model.

Parameter ¹	Mean	SD	Quantiles		
			2.5%	50%	97.5%
g_{11}	92,617	711	90,840	92,724	94,357
g_{22}	96,539	1086	93,917	96,496	98,812
g_{33}	93,134	806	90,582	93,082	95,011
g_{44}	89,487	1,274	86,805	89,389	91,956
g_{55}	96,622	1,353	93,381	96,681	99,623
g_{12}	90,279	626	88,883	90,342	91,643
g_{13}	90,077	617	88,691	90,063	91,412
g_{14}	88,081	624	86,872	88,111	89,744
g_{15}	90,143	731	88,638	90,188	91,762
g_{23}	90,268	609	88,965	90,307	91,607
g_{24}	90,046	658	88,756	89,994	91,542
g_{25}	91,124	667	89,584	91,036	92,313
g_{34}	88,195	593	86,972	88,213	89,574
g_{35}	91,260	712	89,614	91,307	93,007
g_{45}	87,819	728	86,348	87,772	89,023

¹ g_{ij} = genetic (co)variance between milk yields in regions 1 and j (1 = Midwest, 2 = Northeast, 3 = Northwest, 4 = Southeast, 5 = Southwest).

Table 6. Genetic correlations between five regions of the USA using the standard multiple-trait model (above diagonal) and a structural model (below diagonal).

Region ¹	1	2	3	4	5
1	...	0.97	0.96	0.96	0.94
2	0.94	...	0.94	0.97	0.95
3	0.97	0.95	...	0.95	0.97
4	0.95	0.95	0.97	...	0.93
5	0.93	0.94	0.96	0.94	...

¹1 = Midwest, 2 = Northeast, 3 = Northwest, 4 = Southeast, 5 = Southwest.

the standard multiple-trait model. The differences in genetic variance between the two models were between 2.5 to 10%. No pattern was seen in these differences, perhaps reflecting Monte Carlo error. Genetic covariances were slightly lower in the structural model in five out of 10 comparisons, suggesting that differences are probably within Monte Carlo error. However, there was a dramatic increase in precision of estimates of the elements of the genetic covariance matrix when using the structural model. The posterior standard deviations were reduced to more than 1/2 of those found with the standard model. This result is not surprising, given the fewer number of parameters in the structural model. The same was observed for the posterior distributions, with the 95% credibility regions being narrower for the structural model. Genetic correlations (Table 6) were high throughout and very similar for both models. The lowest genetic correlations were between the Southwest and the Midwest (0.93 with the structural model, although the estimate was 0.94 with the multiple trait model), and between the Southeast and Southwest (0.93 with the standard multiple-trait approach and 0.94 with the structural model).

Posterior means and standard deviations of residual variances of milk yield in the five regions are shown for both models in Table 7. Results were similar. In the Midwest, Northeast, and Southeast, residual variances were slightly lower with the standard multiple-trait

Table 7. Posterior means and standard deviations of residual variances (kg²) in five regions of the United States using the standard multiple-trait model and a structural model.

Region ¹	Standard		Structural	
	Mean	SD	Mean	SD
1	1,174,750	2,157	1,181,817	1,979
2	1,265,673	2,562	1,266,762	2,398
3	1,407,540	2,406	1,398,124	2,647
4	1,155,217	3,105	1,163,719	2,991
5	1,476,500	5,215	1,460,255	4,801

¹1 = Midwest, 2 = Northeast, 3 = Northwest, 4 = Southeast, 5 = Southwest.

Table 8. Posterior means and standard deviations of heritability of milk yield in five regions of the United States using the standard multiple-trait model and a structural model.

Region ¹	Standard		Structural	
	Mean	SD	Mean	SD
1	0.308	0.0049	0.316	0.0044
2	0.321	0.0051	0.304	0.0047
3	0.256	0.0048	0.271	0.0048
4	0.289	0.0078	0.297	0.0062
5	0.292	0.0073	0.279	0.0054

¹1 = Midwest, 2 = Northeast, 3 = Northwest, 4 = Southeast, 5 = Southwest.

model, and the opposite was true in the Northwest and Southwest. There were no large differences in the posterior standard deviations of the residual variances, but there was a tendency for these to be smaller in the structural model. At any rate, residual variances were estimated with high precision in the two models, given the large amount of data used in the analysis. In general, differences in residual variances were negligible.

Posterior means of heritability (Table 8) were similar for both models, but precision was greater for the structural model, especially for the Southwest (fewer sires). Here, the posterior variance of heritability with the structural model was 50% of the variance obtained with the multiple-trait analysis (0.0073 vs. 0.0053). This illustrates that imposing a parsimonious structure on the genetic covariance can also enhance the precision of inferences about heritability. In general, heritability of milk yield was about 0.30, with the highest point estimates being for the Northeast and Midwest (0.31 to 0.32), and the lowest one for the Northwest, at around 0.26 with both models.

The DIC was slightly lower for the standard multiple-trait model (0.003%), but differences were within the limits of Monte Carlo error. Table 9 presents the DIC values and its components in the two models. The multiple-trait model had a slightly better fit, but received a slightly larger penalty from having a larger “effective” number of parameters. The goodness of fit component accounted for about 98% of the DIC in the two models.

Table 9. Expectation of the Bayesian deviance, $\bar{D}(\theta)$, Bayesian deviance evaluated at the posterior mean of model parameters, $D(\theta)$, effective number of parameters (p_D) and deviance information criterion (DIC) for the standard multiple-trait model and the structural model.

Model	Standard	Structural
$\bar{D}(\theta)$	38490642	38500483
$D(\theta)$	38131602	38141504
p_D	359038	358979
DIC	38849680	38859462

The structural model is expected to give better results when the amount of information is small and when genetic ties between traits (regions) are weak, which was not the case here. Also, the performance of the structural model depends on the possibility of detecting patterns in the structure of the genetic covariances, and of explaining these patterns with quantitative variables, e.g., management similarity in this study.

The results of the two analyses (high estimates of genetic correlations and overlap between posterior distributions of genetic variances) suggest that it is reasonable to treat data from different geographical regions of the United States as a single trait, although perhaps with heterogeneous residual variance. This has been the approach followed in official genetic evaluation for milk yield in dairy cattle in the United States.

CONCLUSIONS

The structural model presented here, illustrated with dairy cattle data, is an appealing and flexible technique for making inferences about genetic covariances in some animal breeding applications. Its advantages may be even clearer in situations in which there is a limited amount of information on genetic covariances. Typically, insufficient information is due to few genetic ties between traits (e.g., regions in this study). External information regarding management, genetic similarity and climate can be used to explore patterns in genetic covariances. The structural model has fewer parameters than a standard multi-trait model, and can produce a large increase in precision of parameter estimates. In our example, genetic covariances and heritabilities were estimated much more precisely than in the standard multiple-trait analysis, without any adverse effects on the values of the genetic correlations and without compromising model fit. In international genetic evaluation of dairy cattle, genetic ties between countries are often weak. Here, a standard multiple-trait analysis, although technically feasible, would give estimates of genetic correlations that may be of little practical value, due to large standard errors or posterior standard deviations. On the other hand, a structural model with a few parameters explaining genetic covariances is expected to be advantageous in such a situation. The performance of the structural model, however, depends on our ability of detecting existing patterns in the covariance structure, and on finding a right set of variables or parameters to explain such patterns. In our study, information from the dataset (genetic similarity) and from public databases (management similarity) was easy to obtain. In this example, even with a large amount of data and with strong genetic ties between traits (regions), the structural model gave results that

were similar to those from the standard multiple-trait model, with dramatically larger precision, and without any obviously adverse effect on model fit.

The structural model is an extension to the multiple-trait domain of models for heterogeneous variances described by Foulley et al. (1990, 1992). An appealing further extension would consist of imposing a structure both on the genetic covariances and on the residual dispersion components. The latter are known to be heterogeneous in dairy cattle breeding and such heterogeneity has been related to explanatory factors (Weigel et al., 1993).

REFERENCES

- Carlin, B. P., and T. A. Louis. 1996. Bayes and Empirical Bayes Methods for Data Analysis. Chapman and Hall, London.
- Foulley, J. L., D. Gianola, M. San Cristobal, and S. Im. 1990. A method for assessing extent and sources of heterogeneity of residual variances in mixed linear models. *J. Dairy Sci.* 73:1612–1624.
- Foulley, J. L., M. San Cristobal, D. Gianola, and S. Im. 1992. Marginal likelihood and Bayesian approaches to the analysis of heterogeneous residual variances in mixed linear Gaussian models. *Comput. Statist. Data Analysis* 13:1–15.
- Gilks, W. R. 1992. Derivative-free adaptive rejection sampling for Gibbs sampling. Pages 169–194 in *Bayesian Statistics 4*, J. M. Bernardo, J. O. Berger, A. D. Dawid and A. F. M. Smith, ed. Clarendon Press, Oxford, UK.
- Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57:97–109.
- Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953. Equations of state calculations by fast computing machines. *J. Chem. Physics* 21:1087–1091.
- Rekaya, R. 1997. Analisis Bayesiano de datos de produccion en los dias del control para la seleccion de caracteres lecheros. Ph.D. Universidad Politecnica de Madrid, Escuela Tecnica Superior de Ingenieros Agronomos. Madrid, Spain.
- Searle, S. R. 1971. *Linear Models*. Wiley, New York, NY.
- Spiegelhalter, D. J., N. G. Best, and B. P. Carlin. 1998. Bayesian deviance, the effective number of parameters and the complexity of arbitrarily complex models. Technical Report, Medical Research Council, Biostatistics Unit, Cambridge, UK.
- Varona, L., C. Moreno, L. A. Garcia Cortes, and J. Altarriba. 1994. Estimacion multi-caracter de componentes de varianza y covarianza en vacuno lechero mediante muestreo de Gibbs. *Rev. Port. Zootec.* 1:185–195.
- Wakefield, J. C., A. F. M. Smith, A. Racine-Poon, and A. E. Gelfand. 1994. Bayesian analysis of linear and nonlinear population models using the Gibbs sampler. *Appl. Statist.* 43:201–221.
- Wang, C. S., J. Rutledge, and D. Gianola. 1994. Bayesian analysis of mixed linear models via Gibbs sampling with application to litter size in Iberian pigs. *Genet. Select. Evol.* 26:91–115.
- Weigel, K. A., D. Gianola, B. S. Yandell and J. F. Keown. 1993. Identification of factors causing heterogeneous within-herd variance components using a structural model for variances. *J. Dairy Sci.* 76:1466–1478.

APPENDIX

Let the entire vector of parameters of the model be θ with p being the number of elements it possesses. From [17] we have:

$$p_D = \bar{D} - D(\bar{\theta})$$

Expand now $D(\theta)$ up to the second order around $E_{\theta|y} = \bar{\theta}$:

$$D(\theta) = D(\bar{\theta}) + (\theta - \bar{\theta})' \frac{\partial D(\theta)}{\partial \theta} \Big|_{\theta=\bar{\theta}} + \frac{1}{2} (\theta - \bar{\theta})' \frac{\partial^2 D(\theta)}{\partial \theta \partial \theta'} \Big|_{\theta=\bar{\theta}} (\theta - \bar{\theta}) \quad [19]$$

Recall that $D(\theta) = -2 \log p(y|\theta) = -2L$. Then:

$$\frac{\partial D(\theta)}{\partial \theta} = -2L' \text{ and } \frac{\partial^2 D(\theta)}{\partial \theta \partial \theta'} = -2L'' \quad [20]$$

where L is the log-likelihood, and L' and L'' are the first and second derivatives of the log-likelihood with respect to θ , respectively. Using [20] in [19]:

$$D(\theta) = D(\bar{\theta}) - 2(\theta - \bar{\theta})' L_{\bar{\theta}}' - (\theta - \bar{\theta})' L_{\bar{\theta}}'' (\theta - \bar{\theta}) \quad [21]$$

Asymptotically, it is well known that:

$$p(\theta|y) \sim N(\hat{\theta}, [-L_{\hat{\theta}}'']^{-1}) \quad [22]$$

where $\hat{\theta}$ is the maximum likelihood estimator. From [21], and replacing θ by $\hat{\theta}$,

$$D(\theta) = D(\hat{\theta}) - (\theta - \hat{\theta})' L_{\hat{\theta}}'' (\theta - \hat{\theta}) = D(\hat{\theta}) + (\theta - \hat{\theta})' (-L_{\hat{\theta}}'') (\theta - \hat{\theta}) \quad [23]$$

Under [22] it can be shown that $(\theta - \hat{\theta})' (-L_{\hat{\theta}}'') (\theta - \hat{\theta})$ has an asymptotic chi-square distribution with p degrees of freedom (Searle, 1971). From [22] one can write:

$$(\theta - \hat{\theta})|y \sim N(0, [-L_{\hat{\theta}}'']^{-1}) \quad [24]$$

Then, $(\theta - \hat{\theta})' (-L_{\hat{\theta}}'') (\theta - \hat{\theta})$ is chi-square on p degrees of freedom, if and only if, $L_{\hat{\theta}}'' (L_{\hat{\theta}}'')^{-1}$ is idempotent with rank p . This is the case, as $L_{\hat{\theta}}'' (L_{\hat{\theta}}'')^{-1}$ is the identity matrix.

Finally, taking expectations in [23] with respect to the posterior distribution of θ leads to the asymptotic results:

$$E_{\theta|y}[D(\theta)] = D(\hat{\theta}) + p$$

and

$$p = E_{\theta|y}[D(\theta)] - D(\hat{\theta})$$

as in [17]. Recall that asymptotically, $\hat{\theta} = \bar{\theta}$.