

## Genetic Analysis of Somatic Cell Scores in US Holsteins with a Bayesian Mixture Model

P. J. Boettcher,<sup>\*1</sup> D. Caraviello,<sup>†2</sup> and D. Gianola<sup>†</sup>

<sup>\*</sup>Institute of Biology and Biotechnology of Agriculture, National Research Council, Segrate 20090, Italy

<sup>†</sup>Department of Dairy Science, University of Wisconsin, Madison 53706

### ABSTRACT

The objective of this study was to apply finite mixture models to field data for somatic cell scores (SCS) for estimation of genetic parameters. Data were approximately 170,000 test-day records for SCS from first-parity Holstein cows in Wisconsin. Five different models of increasing level of complexity were fitted. Model 1 was the standard single-component model, and the others were 2-component Gaussian mixtures consisting of similar but distinct linear models. All mixture models (i.e., 2 to 5) included separate means for the 2 components. Model 2 assumed entirely homogeneous variances for both components. Models 3 and 4 assumed heterogeneous variances for either residual (model 3) or genetic and permanent environmental variances (model 4). Model 5 was the most complex, in which variances of all random effects were allowed to vary across components. A Bayesian approach was applied and Gibbs sampling was used to obtain posterior estimates. Five chains of 205,000 cycles were generated for each model. Estimates of variance components were based on posterior means. Models were compared by use of the deviance information criterion. Based on the deviance information criterion, all mixture models were superior to the linear model for analysis of SCS. The best model was one in which genetic and PE variances were heterogeneous, but residual variances were homogeneous. The genetic analysis suggested that SCS in healthy and infected cattle are different traits, because the genetic correlation between SCS in the 2 components of 0.13 was significantly different from unity.

**Key words:** mastitis, mixture model, somatic cell count

### INTRODUCTION

The concentration of somatic cells in the milk of dairy cows tends to increase in response to bacterial infection of the mammary gland, because leukocytes are mobi-

lized to the udder to destroy invading pathogens. Data on SCC are routinely collected in the field and evaluated genetically to provide an indirect selection criterion for mastitis resistance (Interbull, 1996). Genetic evaluation and selection programs in most countries are based on SCC because data on mastitis incidence are not collected routinely. Cows out of sires that have a higher proportion of daughters with mastitis will tend to have a larger than average SCC, so selection indexes for udder health typically include negative weights on SCC (e.g., Boettcher et al., 1998). However, the presence of some leukocytes in a healthy udder is believed to be necessary for an initial response to infecting organisms. For this reason, some scientists have expressed concern that selection for low SCC could reduce the ability of cattle to respond to infection (e.g., Kehrl and Shuster, 1994). Several studies have attempted to evaluate the relationship between SCC in healthy udders and subsequent susceptibility to mastitis, but results have not been consistent. Some work has suggested that low SCC is associated with increased risk for infection (Elbers et al., 1998; Suriyasathaporn et al., 2000), but other studies have shown increased mastitis resistance with low SCC in the uninfected state (Beaudeau et al., 1998; Rupp and Boichard, 2000), or for daughters of sires with low estimated breeding values for SCC (Nash et al., 2000).

Current genetic evaluation procedures for SCC treat the records as homogeneous; that is, ignoring that there may be some hidden structure due to unknown disease status (Schutz, 1994). However, Detilleux and Leroy (2000) suggested that SCC could be a different trait, genetically, in infected and uninfected cows. If this were the case, considering the infection status of cattle may be of value when predicting breeding values for SCC. Unfortunately, this practice would not be straightforward, given the lack of data for mastitis incidence at the individual cow level in most countries. Even when mastitis is recorded, only data for the clinical form of the disease are obtained, whereas cattle may be infected subclinically, showing an increased SCC as the only symptom. To overcome these limitations, Detilleux and Leroy (2000) proposed the use of finite mixture models for analysis of SCC in the absence of information re-

Received April 21, 2006.

Accepted August 25, 2006.

<sup>1</sup>Corresponding author: pjoettcher@gmail.com

<sup>2</sup>Current address: Dow AgroSciences LLC, Indianapolis, IN 46268.

garding infection status. Such models are appropriate for analysis of heterogeneous data when observations are derived from different distributions, and are particularly useful for situations (like that of SCC) in which the distribution from which a given observation arose is unknown (McLachlan and Peel, 2000). Dettelleux and Leroy (2000) outlined a maximum likelihood approach for analysis of SCC with a finite mixture model. Gianola et al. (2004) refined this work and proposed algorithms for estimating parameters of interest. In addition, extensions to the model to allow for heterogeneity of variances were proposed; also, Gianola (2005) discussed issues connected with prediction of random effects in mixture models. Ødegård et al. (2003) developed a Bayesian approach for analysis of a 2-component mixture model for SCC with heterogeneous residual variances, and applied it to simulated data.

The model of Ødegård et al. (2003) considered heteroscedasticity of variances for residual effects only, and it was extended subsequently to derive a criterion suitable for selection against putative mastitis (Ødegård et al., 2005). If SCC were a trait that differs genetically between infected and uninfected cattle, allowing for heterogeneity of genetic and permanent environmental (PE) variances would be appropriate. The goal of this study was to extend the approach of Ødegård et al. (2003) to allow for heterogeneous variances of genetic and PE effects and to apply it to data on SCC collected in US Holsteins. Several models of increasing levels of complexity were compared for fit in an attempt to assess which model was most appropriate for use in genetic evaluation of SCC.

## MATERIALS AND METHODS

### Data

Data were test-day records from primiparous Holstein cattle present in 105 large (>200 cows) well-managed herds in the upper Midwestern US (primarily Wisconsin) from January 2000 to March 2004. The SCC records had been converted to linear SCS, using the standard log 2 transformation (Ali and Shook, 1980). Because herds were well managed, the mean SCS was only 2.21 (SD = 1.84), much less than the US national average of approximately 3.00. Data were edited to require a known sire of cow, DIM from 5 to 305 d, and age at calving from 20 to 36 mo. In addition, all cows were required to have at least 1 record in the first 80 d of lactation. The final data set included 177,846 records from 31,040 cows, daughters of 3,082 different sires. An additive relationship file was created by tracing pedigrees at least 3 generations, including ancestors that were related to at least 2 animals with records. The pedigree file included 54,143 animals.

**Table 1.** Summary of the 5 models tested<sup>1</sup>

Feature	Model				
	1	2	3	4	5
Mixture components	1	2	2	2	2
Residual variance	Hom	Hom	Het	Hom	Het
Genetic variance	Hom	Hom	Hom	Het	Het
Permanent environment	Hom	Hom	Hom	Het	Het

<sup>1</sup>Hom = Homogeneous variance; Het = heterogeneous variance.

### Models

Data were analyzed with a series of 5 different models of increasing order of complexity. Model 1 was a standard test-day repeatability model, similar to that used by Reents et al. (1995) for the evaluation of SCS. Fixed effects of systematic nongenetic factors and random additive genetic and PE effects were fitted. The other 4 specifications were 2-component Gaussian mixture models differing according to the type of heterogeneity of variances considered (Table 1). All 3 variances (additive, PE, and residual) were homogeneous for model 2, whereas all variances were heterogeneous for model 5. Analyses were based on previous work of Ødegård et al. (2003), with some extensions to accommodate models 4 and 5.

For the mixture models, observations of SCS were assigned to 1 of the 2 components, assumed to be indicative of health status. Assignments were defined by a (unknown) vector  $\mathbf{z}$ , where  $z_i = 0$  for a record  $i$  from a "healthy" cow and  $z_i = 1$  for records from "infected" cows. Following the notation used by Ødegård et al. (2003), the equations for the various models can be written, given  $\mathbf{z}$ , as:

$$\mathbf{y} = \mathbf{X}_0\beta_0 + \mathbf{M}_z\mathbf{X}_1\beta_1 + (\mathbf{I}-\mathbf{M}_z)\mathbf{Z}_a\mathbf{a}_0 \quad [1]$$

$$+ \mathbf{M}_z\mathbf{Z}_a\mathbf{a}_1 + (\mathbf{I}-\mathbf{M}_z)\mathbf{Z}_p\mathbf{p}_0 + \mathbf{M}_z\mathbf{Z}_p\mathbf{p}_1 + \mathbf{e},$$

where  $\mathbf{y}$  = vector of  $n$  observations for test-day SCS;  $\beta_0$  = vector of fixed effects common to all records;  $\beta_1$  = vector of fixed effects corresponding to observations from infected cows;  $\mathbf{I}$  = identity matrix of order  $n$ ;  $\mathbf{M}_z$  = matrix with diagonal elements corresponding to vector  $\mathbf{z}$ ;  $\mathbf{a}_0$  = vector of random additive genetic effects on SCS in the healthy state;  $\mathbf{a}_1$  = vector of random additive genetic effects on SCS in the infected state;  $\mathbf{p}_0$  = vector of random PE effects in the healthy state;  $\mathbf{p}_1$  = vector of random PE effects in the infection state;  $\mathbf{e}$  = vector of residual effects; and  $\mathbf{X}_0$ ,  $\mathbf{X}_1$ ,  $\mathbf{Z}_a$ , and  $\mathbf{Z}_p$  = incidence matrices corresponding to fixed ( $\mathbf{X}$ ) and random ( $\mathbf{Z}$ ) effects, respectively.

The fixed effects in  $\beta_0$  included 3 regression coefficients for effects of DIM on SCS, 17 effects of age at

calving, and 3,361 herd-test-day effects. Regression coefficients for DIM were based on the Wilmink curve (Wilmink, 1987). Age-at-calving effects were one for each age from 20 through 36 mo. The  $\beta_1$  vector included a single element, the mean difference (shift) between components 1 (healthy) and 2 (diseased) for observation  $n$ . For the nonmixture model (model 1), all elements of  $\mathbf{M}_z$  were zero. For models with homogeneous genetic and PE variances (i.e., Models 1, 2, and 3),  $\mathbf{a}_0 = \mathbf{a}_1$  and  $\mathbf{p}_0 = \mathbf{p}_1$ . For these models,  $\mathbf{a}_0 \sim \text{MVN}(\mathbf{0}, \mathbf{A}\sigma_a^2)$ , where  $\mathbf{A}$  is the numerator relationship matrix and  $\sigma_a^2$  is the additive genetic variance, and  $\mathbf{p}_0 \sim \text{MVN}(\mathbf{0}, \mathbf{I}\sigma_{pe}^2)$ , where  $\mathbf{I}$  is an identity matrix of order 31,040. When genetic and PE effects were heterogeneous, expectations of  $\mathbf{a}_0, \mathbf{a}_1, \mathbf{p}_0, \mathbf{p}_1$  were all zero and

$$\text{Var} \begin{bmatrix} \mathbf{a}_0 \\ \mathbf{a}_1 \end{bmatrix} = \mathbf{G} \otimes \mathbf{A}, \quad [2]$$

where

$$\mathbf{G} = \begin{bmatrix} \sigma_{a0}^2 & \sigma_{a01} \\ \sigma_{a01} & \sigma_{a1}^2 \end{bmatrix}$$

is the variance-covariance matrix between additive genetic values under the “healthy” and “infected” statuses. Further,

$$\text{Var} \begin{bmatrix} \mathbf{p}_0 \\ \mathbf{p}_1 \end{bmatrix} = \mathbf{P} \otimes \mathbf{I}, \quad [3]$$

where

$$\mathbf{P} = \begin{bmatrix} \sigma_{p0}^2 & \sigma_{p01} \\ \sigma_{p01} & \sigma_{p1}^2 \end{bmatrix}$$

is the variance-covariance matrix between corresponding PE effects. Conditionally on the breeding values and PE effects, the variance matrix of the observation vector (residual variance matrix) was expressed as

$$\mathbf{R} = (\mathbf{I} - \mathbf{M}_z)\sigma_{e0}^2 + \mathbf{M}_z\sigma_{e1}^2, \quad [4]$$

where  $\mathbf{I}$  is an identity matrix of order  $n$ , and  $\sigma_{e0}^2$  and  $\sigma_{e1}^2$  are residual variances for observations from the first and second components, respectively. For models with homogeneous residual variance, (i.e., models 1, 2, and 4) equation [4] simplifies to  $\mathbf{R} = \mathbf{I}\sigma_{e0}^2$ .

### Bayesian Analysis

As mentioned previously, this study extended the work of Ødegård et al. (2003) by allowing for heteroge-

neity of animal (additive genetic and PE) and residual variances. Many aspects of the Bayesian structure of the model were very similar. Following the notation of Ødegård et al. (2003), the density of  $\mathbf{y}$ , conditional on  $\mathbf{z}$ , fixed and random effects, and residual variances, of the most complex model (with heterogeneity of all variances) can be expressed as

$$p(\mathbf{y} | \mathbf{z}, \beta_0, \beta_1, \mathbf{a}_0, \mathbf{a}_1, \mathbf{p}_0, \mathbf{p}_1, \sigma_{e0}^2, \sigma_{e1}^2) = p(\mathbf{y} | \beta_0, \mathbf{a}_0, \mathbf{p}_0, \sigma_{e0}^2, \{z_i = 0\}) \times p(\mathbf{y} | \beta_0, \beta_1, \mathbf{a}_1, \mathbf{p}_1, \sigma_{e1}^2, \{z_i = 1\}), \quad [5]$$

where  $\{z_i = 0\}$  and  $\{z_i = 1\}$  denote the observations assigned to the first and second components, respectively, of the mixture model. In more detail,

$$p(\mathbf{y} | \beta_0, \mathbf{a}_0, \mathbf{p}_0, \sigma_{e0}^2, \{z_i = 0\}) \propto \sigma_{e0}^2^{-\left(\frac{n_0}{2}\right)} \exp \left[ -\frac{1}{2\sigma_{e0}^2} (\mathbf{y} - \mathbf{X}_0\beta_0 - \mathbf{Z}_a\mathbf{a}_0 - \mathbf{Z}_p\mathbf{p}_0)' (\mathbf{I} - \mathbf{M}_z)(\mathbf{y} - \mathbf{X}_0\beta_0 - \mathbf{Z}_a\mathbf{a}_0 - \mathbf{Z}_p\mathbf{p}_0) \right] \quad [6]$$

for the  $n_0$  observations assigned to the first component. For observations assigned to the second component (i.e., when  $z_i = 1$ ), the conditional density is obtained by subtracting  $\mathbf{X}_1\beta_1$  to the quadratic form in equation [6], and replacing  $\sigma_{e0}^2$  with  $\sigma_{e1}^2$ ,  $n_0$  with  $n_1$ ,  $\mathbf{I} - \mathbf{M}_z$  with  $\mathbf{M}_z$  and  $\mathbf{Z}_a\mathbf{a}_0$  and  $\mathbf{Z}_p\mathbf{p}_0$  with  $\mathbf{Z}_a\mathbf{a}_1$  and  $\mathbf{Z}_p\mathbf{p}_1$ , respectively.

With regard to prior distributions, widely bounded uniform priors were assigned to all fixed effects. In addition,  $\beta_1$  was required to be greater than zero, to attain parameter identification. The multivariate normal distributions given above were used as priors for the random effects, conditionally on  $\mathbf{G}$  and  $\mathbf{P}$  for additive genetic and PE effects, respectively. Priors for  $\mathbf{G}$  and  $\mathbf{P}$  were inverted Wishart distributions, defined by 2 parameters, a degrees of freedom  $\nu$ , and a scale matrix  $\mathbf{V}$  of the same dimension of  $\mathbf{G}$  and  $\mathbf{P}$ . In all cases,  $\nu$  was set equal to 5. For the scale matrices, diagonal elements (variances) were systematically varied from 0.1 to 2.0 in different chains to examine the influence of changing these prior values. Off diagonal elements were held constant at a small, but nonzero value (0.001). The priors for  $\sigma_{e0}^2$  and  $\sigma_{e1}^2$  were scale-inverted  $\chi^2$  distributions. For example, for  $\sigma_{e0}^2$ , the prior was

$$p(\sigma_{e0}^2) \propto (\sigma_{e0}^2)^{-(\nu_{e0}+2)/2} \exp[-1/2(\nu_{e0}s_{e0}^2/\sigma_{e0}^2)], \quad [7]$$

where  $s_{e0}^2$  and  $\nu$  are hyper-parameters corresponding to the prior variance and degrees of freedom, respectively.

Experimentation was done to test for the effects of different priors for  $s_{e0}^2$ , and chains eventually converged to similar posterior distributions. Therefore, a fixed prior was used. For all models,  $\nu$  was set to 5 and  $s_e^2$  was set to 1.0 for both mixture components. The elements of the  $\mathbf{z}$  vector were assumed to be independent and identically distributed as Bernoulli random variables, a priori, with their probabilities depending on  $P_m$ , the probability that an SCS is drawn from the “infected” status. The joint prior distribution of  $\mathbf{z}$  was

$$p(\mathbf{z} | P_m) = \prod_{i=1,2,\dots,n} P_m^{1-z_i}(1 - P_m)^{z_i} \quad [8]$$

Finally, the prior for the mixing proportion was a beta distribution with parameters  $\alpha$  and  $\beta$ . A value of 2 was assumed for both parameters, as in the example of Ødegård et al. (2003).

Based on the sampling distribution of  $\mathbf{y}$  and on the various prior distributions assigned, the joint posterior density of all unknown parameters was assumed to take the form

$$p(\beta_0, \beta_1, \mathbf{a}_0, \mathbf{a}_1, \mathbf{p}_0, \mathbf{p}_1, \mathbf{z}, \mathbf{G}, \mathbf{P}, \mathbf{R}, P_m | \mathbf{y}) \propto p(\mathbf{y}, \beta_0, \beta_1, \mathbf{a}_0, \mathbf{a}_1, \mathbf{p}_0, \mathbf{p}_1, \mathbf{z} | \mathbf{G}, \mathbf{P}, \sigma_{e0}^2, \sigma_{e1}^2, P_m) \quad [9]$$

$$p(\mathbf{G}) p(\mathbf{P}) p(\sigma_{e0}^2) p(\sigma_{e1}^2) p(P_m).$$

To implement a Gibbs sampler, realizations for each parameter of interest must be drawn from their conditional posterior distributions, given the most recent values for all other parameters in the model. For an element of the  $\beta$ ,  $\mathbf{a}$ , and  $\mathbf{p}$  vectors, the conditional posterior distribution was Gaussian. The mean was obtained by solving the mixed model equation corresponding to that element by inserting the most recent realizations for all other elements of  $\beta$ ,  $\mathbf{a}$ , and  $\mathbf{p}$  and forming an offset of the data vector (Wang et al., 1994). The variance was equal to the most recent realization of the residual variance divided by the diagonal element of the coefficient matrix that corresponded to the element of interest (Sorensen, 1999). The conditional posterior distributions of  $\mathbf{G}$  and  $\mathbf{P}$  were the inverted Wishart distributions:

$$\mathbf{G} \text{ and } \mathbf{P} \sim IW[(\mathbf{S}_f + \mathbf{V}_f^{-1})^{-1}, \nu_f + q_f] \quad [10]$$

where  $f = a$  or  $p$  and  $\mathbf{V}_f$  and  $\nu_f$  are scale and degree of freedom parameters of the corresponding prior distributions. For  $\mathbf{G}$

$$\mathbf{S}_f = \begin{bmatrix} \mathbf{a}_0' \mathbf{A}^{-1} \mathbf{a}_0 & \mathbf{a}_0' \mathbf{A}^{-1} \mathbf{a}_1 \\ \mathbf{a}_1' \mathbf{A}^{-1} \mathbf{a}_0 & \mathbf{a}_1' \mathbf{A}^{-1} \mathbf{a}_1 \end{bmatrix} \quad [11]$$

and  $q_f$  is the number of animals in the relationship matrix. For  $\mathbf{P}$

$$\mathbf{S}_f = \begin{bmatrix} \mathbf{p}_0' \mathbf{P}_0 & \mathbf{p}_0' \mathbf{P}_1 \\ \mathbf{p}_1' \mathbf{P}_0 & \mathbf{p}_1' \mathbf{P}_1 \end{bmatrix} \quad [12]$$

and  $q_f$  is the number of animals with records. The fully conditional posterior distributions of  $\sigma_{e0}^2$  and  $\sigma_{e1}^2$  were both scale-inverted  $\chi^2$  distributions. For  $\sigma_{e0}^2$ , the scale parameter was

$$(\mathbf{y} - \mathbf{X}_0 \beta_0 - \mathbf{Z}_a \mathbf{a}_0 - \mathbf{Z}_p \mathbf{p}_0)' (\mathbf{I} - \mathbf{M}_z) (\mathbf{y} - \mathbf{X}_0 \beta_0 - \mathbf{Z}_a \mathbf{a}_0 - \mathbf{Z}_p \mathbf{p}_0) + \nu_{e0} s_{e0}^2. \quad [13]$$

The scale parameter for  $\sigma_{e1}^2$  was similar to [13], with substitutions to account for the fact that different additive genetic and PE effects were present for observations in the second mixture component:

$$(\mathbf{y} - \mathbf{X}_0 \beta_0 - \mathbf{X}_1 \beta_1 - \mathbf{M}_z \mathbf{Z}_a \mathbf{a}_1 - \mathbf{M}_z \mathbf{Z}_p \mathbf{p}_1)' (\mathbf{y} - \mathbf{X}_0 \beta_0 - \mathbf{X}_1 \beta_1 - \mathbf{M}_z \mathbf{Z}_a \mathbf{a}_1 - \mathbf{M}_z \mathbf{Z}_p \mathbf{p}_1) + \nu_{e1} s_{e1}^2.$$

The elements of  $\mathbf{z}$  assigning individual records to 1 of the 2 mixture components had mutually independent Bernoulli conditional posterior distributions. Bernoulli distributions are fully defined by a parameter  $p$ , which is the probability that a certain binary outcome will be obtained. In this case,  $p_i$  was the probability that observation  $i$  would be assigned to the second (infected) component.

$$p_i = \frac{p(y_i | \theta, \mathbf{G}, \mathbf{P}, \sigma_{e1}^2, z_i = 1) P_m}{p(y_i | \theta, \mathbf{G}, \mathbf{P}, \sigma_{e0}^2, z_i = 0)(1 - P_m) + p(y_i | \theta, \mathbf{G}, \mathbf{P}, \sigma_{e1}^2, z_i = 1) P_m} \quad [14]$$

where  $\theta = [\beta_0' \beta_1' \mathbf{a}_0' \mathbf{a}_1' \mathbf{p}_0' \mathbf{p}_1']$ . The conditional densities for the  $y_i$  were as presented in equation [6], using the most recent realizations of the various parameters as true values. The fully conditional distribution of the mixing parameter  $P_m$  was a beta distribution, with parameters  $n_0 + \alpha$  and  $n_1 + \beta$ .

The Bayesian structure of the more simple models is obtained by removing terms that refer to effects and (co)variances for the second component when the variance for a given factor is homogeneous. When the variances of genetic and PE effects were homogeneous, prior and fully conditional distributions were scale-inverted  $\chi^2$  distributions, rather than inverted Wishart.

### The Gibbs Sampler

The Gibbs sampler was run as follows: 1) Initial values for all fixed and random effects were zero. Variances

and covariances were set to the values used in the corresponding prior distributions. 2) Observations were randomly assigned to 1 of the 2 mixture components. 3) Fixed effects were sampled piecewise from univariate normal distributions in the following order: a) regressions on DIM, b) mean of the second component, c) age-at-calving, d) herd-test-day. The mean of the second component was forced to be  $\geq 0$ , so only positive values of  $\beta_1$  were accepted in the sampling process. 4) Random effects were sampled from univariate normal distributions. Genetic effects were sampled first, followed by PE effects. 5) The PE covariance matrix was sampled from an inverted-Wishart (or scale-inverted  $\chi^2$ ) distribution. 6) The genetic covariance matrix was sampled from an inverted-Wishart (or scale-inverted  $\chi^2$ ) distribution. 7) The residual variances were sampled from scale-inverted  $\chi^2$  distributions. Variance of the "healthy" component was sampled first. 8) Group membership variables (i.e., elements of  $\mathbf{z}$ ) were sampled from Bernoulli distributions. 9) The mixing proportion  $P_m$  was sampled from a beta distribution.

The steps for the appropriate models were repeated as needed at each chain of the Gibbs sampler. Five sampling chains of 205,000 cycles each were generated for each model. For each chain, the first 5,000 cycles were discarded as burn-in period so that a total of 1,000,000 posterior samples were available for each model. Convergence was assessed by the approach of Gelman et al. (2004), which involves calculating the square root of the sum of the within and across chain variances, divided by the variance within chains. Convergence was declared when this value was less than 1.1. Model 5 was the slowest model to converge, with a ratio of 1.09; the value of this ratio was  $< 1.02$  for all other models. Posterior distributions of (co)variances were assessed based on sampling every 20th cycle. Posterior means for breeding values were obtained by averaging realizations from every 500th cycle.

### Comparison of Models

**Model Fit.** The models were compared based on the deviance information criterion (**DIC**; Spiegelhalter et al., 2002). The DIC is one of a family of methods, including Akaike's information criterion (Akaike, 1970) and the Bayesian information criterion (Spiegelhalter et al., 2002), which consider both the fit and the complexity of a model. The DIC can be expressed as

$$\text{DIC} = D + p_D, \quad [15]$$

where  $D$  is the posterior expectation of the Bayesian deviance and  $p_D$  is a measure of the effective number of parameters in the model. The  $p_D$  is obtained based

on the difference between the posterior mean of the deviance and the deviance evaluated at the posterior means of the parameters. The model with the lowest DIC is considered to be the most appropriate model statistically. The DIC has previously been used for evaluation of Bayesian models applied to livestock data (Rekaya et al., 2003).

The quantities needed to calculate the DIC can be obtained readily from the Gibbs sampling iteration. To obtain  $D$ , the expected Bayesian deviance, one needs to calculate the Bayesian deviance:

$$= -2 \log p(\mathbf{y} | \boldsymbol{\theta}, \mathbf{R}) \quad [16]$$

every  $k$ th cycle of the Gibb's sampler ( $\boldsymbol{\theta}$  is a vector of the realized values of all parameters of interest at that cycle of the sampling chain), and then average these values over all samples taken. The deviance evaluated at the posterior means of the parameters of interest is then calculated after the Gibb's chain is completed, by substituting  $\bar{\boldsymbol{\theta}}$  for  $\boldsymbol{\theta}$  in equation [16] where  $\bar{\boldsymbol{\theta}}$  is a vector of posterior means of all parameters entering into the deviance. Here,  $D$  was evaluated every 2,000 cycles, whereas posterior means were calculated based on samples obtained every 500th cycle.

**EBV.** In addition to comparing models for statistical appropriateness, the EBV resulting from the different models were evaluated for similarity. The posterior means of additive genetic effects (calculated by sampling every 500th cycle) were used as EBV. Two approaches were used to examine similarity of EBV. First, Pearson correlation coefficients were calculated between all pairs of the 7 sets of animal solutions (1 set of EBV from each of the 3 models (1, 2, and 3) with homogeneous genetic variance and 2 sets each from the 2 models (4 and 5) with heterogeneous genetic variance). Correlation coefficients were calculated for 2 sets of animals: 1) all animals,  $n = 54,143$ , and 2) only sires with at least 10 offspring ( $n = 541$ ).

Second, to examine changes in rank, all sires with at least 10 offspring were sorted in ascending order based on each of the 7 sets of EBV. Then, the top and bottom 50 sires were identified for each set. Finally, the number of animals in common between each pair (high ranking with high ranking and low vs. low) of these sets was observed. Low numbers of mismatches were assumed to indicate high similarity among evaluation models.

## RESULTS

Posterior means for selected parameters of interest for the 5 models are in Table 2. Most models produced similar estimates of the mixing proportion ( $p_2$ ), with around 5% of the observations in the second component

**Table 2.** Posterior means (standard deviations) for various parameters according to the different models

Parameter <sup>1</sup>	Model				
	1	2	3	4	5
$p_2$	—	0.054 (0.006)	0.040 (0.005)	0.083 (0.002)	0.047 (0.001)
$\mu_1$	1.80 (0.019)	1.61 (0.017)	1.66 (0.017)	1.66 (0.027)	1.64 (0.017)
$\mu_2$	—	5.65 (0.029)	6.09 (0.023)	2.36 (0.228)	5.74 (0.042)
$\sigma_{G1}^2$	0.16 (0.024)	0.23 (0.023)	0.10 (0.018)	0.12 (0.019)	0.31 (0.040)
$\sigma_{G2}^2$	—	—	—	2.41 (0.251)	0.52 (0.151)
$\sigma_{G12}$	—	—	—	0.07 (0.047)	0.39 (0.092)
$\sigma_{PE1}^2$	1.14 (0.023)	0.90 (0.021)	1.09 (0.019)	1.23 (0.022)	0.90 (0.033)
$\sigma_{PE2}^2$	—	—	—	3.22 (0.202)	0.76 (0.388)
$\sigma_{PE12}$	—	—	—	0.19 (0.056)	0.39 (0.092)
$\sigma_{e1}^2$	1.63 (0.006)	0.95 (0.004)	0.95 (0.005)	0.86 (0.004)	0.90 (0.006)
$\sigma_{e2}^2$	—	—	1.20 (0.025)	—	1.00 (0.025)

<sup>1</sup> $p_2$  = Proportion of observations assigned to the second component,  $\mu_i$  = mean of observations from component  $i$ ,  $\sigma^2$  = variance,  $\sigma$  = covariance, G = genetic, PE = permanent environmental, e = residual.

(presumably associated with mastitis), and 95% in the “healthy” group. For models 2, 3, and 5, the means of the 2 components were similar across models and quite distinct from each other, ranging from 1.61 to 1.66 for the low group and 5.65 to 6.09 for the high group. Results from model 4, however, were strikingly different. First, the proportion of records assigned to the second (high) component was much greater, at about 8%, vs. around 5% for the other 3 mixture models. Also, the difference in means of the 2 groups was much less. At 1.66 SCS, the mean of observations in the first component of model 4 was similar to those for the other 3 mixture models. However, the mean of the second component was much lower, at 2.36, vs. around 6.00 for the other 3 models. The posterior standard deviation was also much greater for model 4, in spite of the fact that the posterior standard deviation of the mixture proportion was lower for model 4 than for the other mixture models.

All mixture models (2 to 5) had much lower residual variance than did the standard linear model (model 1). Residual variance was generally around 1.00 for the mixture models (with the exception of a residual variance of 1.20 for the second component of model 3) vs. 1.60 for the linear Gaussian model. This difference is due to the variability in means between the 2 components in the mixture models, which is unaccounted for in the linear model specification. When heterogeneous variance was allowed, the residual variance estimate was somewhat larger for the “infected” component of the mixture.

No obvious trend was observed for genetic variance when comparing the standard model with the 4 mixture models. In model 4, the estimates of genetic and PE variances for the second component were much larger than the variances obtained by either of the other 3

mixture models. The genetic variance of the second component was 2.41 in model 4 vs. 0.52 for model 5; corresponding PE variances were 3.22 and 0.76, respectively.

Allowing for genetic and PE variances to be heterogeneous across components had a much greater effect on the analysis than did allowing for heterogeneous residual variance (because differences between models 2 and 3 were much less than between models 2 and 4). In addition, the genetic effects for the 2 components were distinctly different in model 4, and the correlation between genetic effects of the 2 components was only 0.13. In contrast, when genetic, PE, and residual variances were all allowed to be heterogeneous (model 5), the genetic effects of the 2 components essentially differed only in variance ( $r = 0.97$ ).

Based on the results from Table 2, clear differences among the models existed. However, results did not reveal per se which model was more strongly supported by the data. The DIC values (based on means across 5 sampling chains per model) for the 5 models are in Table 3. In the light of the empirical standard errors of DIC statistics, based on between-chain variability, all models can be considered as producing statistically significantly different results, roughly. According to the DIC, model 4 was favored, by far; recall that a model with the lowest DIC is preferred. The average DIC for

**Table 3.** Deviance information criteria for the different models

Model	Mean <sup>1</sup>	SEM
1	955,354	98
2	562,278	42
3	564,112	36
4	509,907	23
5	539,131	294

<sup>1</sup>Means across 5 sampling chains.

**Table 4.** Correlations<sup>1</sup> between predicted breeding values for SCS from the different models, for all animals<sup>2</sup> (above diagonal) and sires with at least 10 daughters<sup>3</sup> (below diagonal)

Model <sup>4</sup>	Model						
	1	2	3	4		5	
				C1	C2	C1	C2
1		0.93	0.94	0.94	0.32	0.92	0.92
2	0.96		0.94	0.93	0.23	0.97	0.96
3	0.97	0.98		0.97	0.21	0.90	0.88
4 C1	0.95	0.96	0.98		0.14	0.90	0.89
4 C2	0.48	0.37	0.37	0.28		0.28	0.29
5 C1	0.95	0.98	0.95	0.94	0.43		>0.99
5 C2	0.95	0.98	0.94	0.93	0.45	>0.99	

<sup>1</sup>All correlations were significantly different from zero ( $P < 0.0001$ ).

<sup>2</sup> $n = 54,143$ .

<sup>3</sup> $n = 541$ .

<sup>4</sup>C1 and C2 are components 1 and 2, respectively.

model 4 was 509,907 or about 6% lower than the DIC of the second-ranking model, model 5. The DIC for model 5 was about 10% lower than for models 2 and 3. As one can observe by comparing models 4 and 5, the assumption of fixed residual variance had profound effects on both the parameter estimates and the fit of the model as evaluated by DIC. The proportion of observations in the second or “infected” component was much greater than in any other model and the difference in SCS level for the 2 classes was much less than for any other model. One conclusion that can be drawn is that models 1 to 3 were underparameterized relative to model 4 (because DIC were larger) and model 5 was overparameterized (larger and more variable DIC). Table 3 also shows that any of the 4 mixture models used was superior to the standard linear model (model 1) for analysis of SCS data. The DIC of model 1 was nearly twice as large as for any of the mixture models. Boettcher et al. (2005) observed a similar advantage when applying a simple mixture model (i.e., similar to model 2) to SCS data from goats.

The correlations among EBV from the different pairs of models were all about 0.90 (Table 4), except for the pairs that included the second (high) component of model 4. Correlations with the EBV for the second component of model 4 ranged from 0.14 to 0.32 when all animals were considered, and from 0.28 to 0.48 when only sires with  $\geq 10$  daughters were considered. The lowest correlations between sets of EBV were between the first and second components of model 4. At the same time, the EBV from the first component of model 4 were similar to the EBV from the other models (generally around 0.95). The highest correlations ( $< 0.99$ ) were between the 2 components of model 5.

Despite high correlations among EBV, the degree of sire reranking among models (Table 5) indicates that the use of a mixture model would lead to real changes

in sire selection if applied instead of the linear model. For all mixture models (models 2 through 5), the top 50 sires (low SCS) differed by at least 10 sires ( $> 20\%$ ) from the top 50 identified by the linear model (model 1). Slightly less reranking was observed among the bottom 50 sires. As expected, the rankings of sires were widely different between the 2 components of model 4. Eleven sires were in common among the top 50, and 13 were in common among the bottom 50.

## DISCUSSION

Based strictly on statistical considerations, mixture models are more appropriate for analysis of SCS data of dairy cattle than the standard linear model. Four different mixture models were applied in this study, and all had a markedly lower DIC than the linear model, indicating superior fit to the data while accounting for the increased complexity of the mixture model. Correlations of EBV from the mixture models with those from the linear model were generally  $\geq 0.90$ , but shuffling in order of the highest ranked sires was observed, demonstrating that practical differences would be realized with the adoption of a mixture model for genetic evaluation. Differences between the linear and mixture models may be even more marked in an analysis with all herds included, because this study included primarily data from well-managed herds with low mean SCS.

The superiority of the mixture model may also increase if a more complex model is applied. The primary objective of this research was to compare models that differed according to the distributions of random effects in the models and heterogeneity of variance of effects from each component. Therefore, very simple assumptions were made about the mixing proportions and all observations had the same prior of falling in each of

**Table 5.** Number of animals in common among the 50 highest (above diagonal) and lowest (below) sires<sup>1</sup> for each pair of breeding value estimates from the various models

Model <sup>2</sup>	Model							
	1	2	3	4		5		
				C1	C2	C1	C2	
1		38	39	37	17	36	37	
2	41		44	40	12	44	42	
3	43	44		44	12	40	38	
4 C1	42	41	44		11	35	35	
4 C2	17	14	15	13		13	14	
5 C1	39	46	41	40	39		48	
5 C2	39	46	40	39	15	49		

<sup>1</sup>Sires with at least 10 daughters.

<sup>2</sup>C1 and C2 are components 1 and 2, respectively.

the 2 components, regardless of possible effects of differences in herds, ages, and genetics on this probability. Ødegård et al. (2005) have outlined an approach to accomplish this by including a “liability” variable into the mixing proportion.

The “best” mixture model (based on having the lowest DIC) was one that allowed for heterogeneous genetic and PE variance, but assumed homogeneous residual variance. Correlations between the EBV of the 2 components of the model with heterogeneous genetic effects were low, suggesting that SCS for infected and healthy cows may be different traits.

Although the statistical evidence supporting the use of mixture models is strong, questions remain about the biological ramifications of applying a mixture model, and about the precise meaning of the different EBV resulting from a mixture model with heterogeneous genetic effects. One possible way to approach this question may be to analyze a set of data for which the true infection status was known, considering SCS from healthy and infected cows as different traits, similar to the approach of Heringstad et al. (2006), and then comparing these results with those from the application of a mixture model to the same data. One might also consider a more complex model that includes more components based on the type of pathogen. Some research (M. M. Schutz, Purdue University, West Lafayette, IN; personal communication) has suggested that SCS in response to a contagious infection is under greater genetic control than SCS during an environmental infection. Of course, prospects for such studies are limited by lack of data availability. Another issue is how a genetic evaluation for SCS can be translated into a selection criterion, as discussed in Ødegård et al., (2005).

Finally, the statistical analysis of SCS and IMI should be complemented with biological research to determine whether high or low SCS is favorable in both healthy and infected states. As mentioned in the intro-

duction, contradictory reports have been presented on the relationship between SCS in the seemingly healthy udder and subsequent susceptibility to infection. The relationships between SCS in the infected udder and elimination of the pathogen should also be examined. High SCS in this state could be beneficial, leading to a fast return to health, but a very strong immune response; high SCS could also trigger more damage to the udder and cause greater loss in milk yield.

## ACKNOWLEDGMENTS

The authors would like to thank the producers from whom the data for the study were collected. Support by the Wisconsin Agriculture Experiment Station, and by grants NRICGP/USDA 2003-35205-12833 and NSF DEB-0089742 is acknowledged.

## REFERENCES

- Akaike, H. 1970. Statistical predictor identification. *Ann. Inst. Stat. Math.* 22:203–217.
- Ali, A. K. A., and G. E. Shook. 1980. An optimum transformation for somatic cell concentration in milk. *J. Dairy Sci.* 63:487–490.
- Beaudeau, F., H. Seegers, C. Fourichon, and P. Hortet. 1998. Association between milk somatic cell counts up to 400 000 cells/mL and clinical mastitis occurrence in French Holstein cows. *Vet. Rec.* 143:685–687.
- Boettcher, P. J., J. C. M. Dekkers, and B. W. Kolstad. 1998. Development of an udder health index that includes milking speed. *J. Dairy Sci.* 81:1157–1168.
- Boettcher, P. J., P. Moroni, G. Pisoni, and D. Gianola. 2005. Application of a finite mixture model to somatic cell scores of Italian goats. *J. Dairy Sci.* 88:2209–2216.
- Detilleux, J., and P. L. Leroy. 2000. Application of a mixed normal mixture model for the estimation of mastitis-related parameters. *J. Dairy Sci.* 83:2341–2349.
- Elbers, A. R. W., J. D. Milrenburg, D. Delange, A. P. P. Crauwells, H. W. Barkema, and Y. H. Schukken. 1998. Risk factors for clinical mastitis in a random sample of dairy herds from the southern part of the Netherlands. *J. Dairy Sci.* 81:420–426.
- Gelman, A., J. B. Carli, H. S. Stern, and D. B. Rubin. 2004. Bayesian data analysis. 2nd ed. Chapman and Hall, Boca Raton, FL.
- Gianola, D. 2005. Prediction of random effects in finite mixture models with Gaussian components. *J. Anim. Breed. Genet.* 122:145–160.



- Gianola, D., J. Ødegård, B. Heringstad, G. Klemetsdal, D. Sorensen, P. Madsen, J. Jensen, and J. Detilleux. 2004. Mixture model for inferring susceptibility to mastitis in dairy cattle: A procedure for likelihood-based inference. *Genet. Sel. Evol.* 36:3–27.
- Heringstad, B., D. Gianola, Y. M. Chang, J. Ødegård, and G. Klemetsdal. 2006. Genetic associations between clinical mastitis and somatic cell score in early first lactation cows. *J. Dairy Sci.* 89:2236–2244.
- Interbull. 1996. Sire evaluation procedures for nondairy-production, and growth and beef production traits practiced in various countries. *Interbull Bull.* 13.
- Kehrli, M. E., Jr., and D. E. Shuster. 1994. Factors affecting milk somatic cells and their role in health of the bovine mammary gland. *J. Dairy Sci.* 77:619–627.
- McLachlan, G., and D. Peel. 2000. *Finite mixture models*. 1st ed. John Wiley and Sons, New York, NY.
- Nash, D. L., G. W. Rogers, J. B. Cooper, G. L. Hargrove, J. F. Keown, and L. B. Hansen. 2000. Heritability of clinical mastitis incidence and relationships with sire transmitting abilities for somatic cell score, udder type traits, productive life, and protein yield. *J. Dairy Sci.* 83:2350–2360.
- Ødegård, J., J. Jensen, P. Madsen, D. Gianola, G. Klemetsdal, and B. Heringstad. 2003. Detection of mastitis in dairy cattle by use of mixture models for repeated somatic cell scores: A Bayesian approach via Gibbs sampling. *J. Dairy Sci.* 86:3694–3703.
- Ødegård, J., J. Jensen, P. Madsen, D. Gianola, G. Klemetsdal, and B. Heringstad. 2005. A Bayesian liability-normal mixture model for analysis of a continuous mastitis-related trait. *J. Dairy Sci.* 88:2652–2659.
- Reents, R., J. Jamrozik, L. R. Schaeffer, and J. C. Dekkers. 1995. Estimation of genetic parameters for test day records of somatic cell score. *J. Dairy Sci.* 78:2847–2857.
- Rekaya, R., K. A. Weigel, and D. Gianola. 2003. Bayesian estimation of parameters of a structural model for genetic covariances between milk yield in five regions of the United States. *J. Dairy Sci.* 86:1837–1844.
- Rupp, R., and D. Boichard. 2000. Relationship of early first lactation somatic cell count with risk of subsequent clinical mastitis. *Livest. Prod. Sci.* 62:169–180.
- Schutz, M. M. 1994. Genetic evaluation of somatic cell scores for United States dairy cattle. *J. Dairy Sci.* 77:2113–2129.
- Sorensen, D. 1999. *Gibbs Sampling in Quantitative Genetics*. Danish Inst. Anim. Sci., Tjele, Denmark.
- Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. van der Linde. 2002. Bayesian measures of model complexity and fit. *J. R. Stat. Soc., B, Stat. Methodol.* 64:583–639.
- Suriyasathaporn, W., Y. H. Schukken, M. Nielen, and A. Brand. 2000. Low somatic cell count: A risk factor for subsequent clinical mastitis in a dairy herd. *J. Dairy Sci.* 83:1248–1255.
- Wang, C. S., J. J. Rutledge, and D. Gianola. 1994. Bayesian analysis of mixed linear models via Gibbs sampling with an application to litter size in Iberian pigs. *Genet. Sel. Evol.* 26:91–115.
- Wilmink, J. B. M. 1987. Adjustment of test-day milk, fat and protein yield for age, season and stage of lactation. *Livest. Prod. Sci.* 16:335–348.