# GENETIC ANALYSIS OF YEARLING WEIGHT IN SIMMENTAL CATTLE USING GAUSSIAN MIXTURE RESIDUALS

**H.N. Oliveira[1], G.B. Santos[1], L.F.A. Marques[2] and G.J.M. Rosa[3]**

[1] Department of Animal Breeding and Nutrition, FMVZ, UNESP, Botucatu, SP, Brazil
[2] Veterinary School, Universidade Federal Fluminense, Niterói, RJ, Brazil
[3] Depts. of Animal Sciences and of Fisheries & Wildlife, MSU, East Lansing - MI, USA

## INTRODUCTION

The Gaussian distribution is often adopted for statistical modeling of genetic evaluations of the Simmental breed in Brazil. For some traits, however, the Gaussian process does not present a good approximation of the sampling model because of heterogeneity of the data and/or the presence of outliers. Some thick-tailed distributions have been suggested in the literature as alternatives to the normal distribution, such as the Student-t process (Strandén and Gianola, 1999; Rosa, 1999). Here, a mixture of Gaussian distributions is discussed within the mixed-model and animal-breeding context. A Bayesian framework is adopted and Markov chain Monte Carlo is used to carry out the posterior analysis. The methodology is applied to variance component estimation and genetic evaluation of yearling weight in Simmental breed.

## THE GAUSSIAN MIXTURE ANIMAL MODEL

Consider the animal model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\varepsilon}$ , where $\boldsymbol{X}$ and $\boldsymbol{Z}$ are incidence matrices, β is the vector of fixed effects, $\boldsymbol{u}$ is the vector of breeding values, and ε is the vector of residuals. Here, the usual Gaussian assumption for the residuals is replaced by a mixture model, where each element of the residual vector is expressed by $\varepsilon_i = e_i / \sqrt{w_i}$ , where $e_i \sim N(0, \sigma_e^2)$ , and $w_i$ is a discrete random variable with probability distribution given by:

$$p(w_i \mid \lambda, \gamma) = \begin{cases} \gamma & , \text{if } w_i = \lambda \\ 1-\gamma & , \text{if } w_i = 1 \end{cases}$$

with $0 < \lambda < 1$ and $0 \leq \gamma < 1$.

The joint distribution of $\boldsymbol{y}$ and of the unobservable random vector $\boldsymbol{w}$, given the parameters, is $p(\boldsymbol{w}, \boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{u}, \sigma_e^2, \lambda, \gamma) = p(\boldsymbol{y} \mid \boldsymbol{w}, \boldsymbol{\beta}, \boldsymbol{u}, \sigma_e^2) \prod_{i=1}^{m} p(w_i \mid \lambda, \gamma)$ . To complete the specification of the model in a Bayesian context, it is assumed that the joint prior distribution of the unknowns $\boldsymbol{\beta}$ , $\boldsymbol{u}$, $\sigma_g^2$, $\sigma_e^2$, $\lambda$ and $\gamma$ has density $p(\beta, \boldsymbol{u}, \sigma_g^2, \sigma_e^2, \lambda, \gamma) = p(\boldsymbol{\beta}) p(\boldsymbol{u} \mid \sigma_g^2) p(\sigma_g^2) p(\sigma_e^2) p(\lambda) p(\gamma)$ . For the fixed effects, flats priors are often assumed in animal breeding. The breeding values are assumed to have a multivariate normal distribution with null mean vector and covariance matrix $\boldsymbol{A}\sigma_g^2$ , where $\boldsymbol{A}$ is the numerator relationship

matrix and $\sigma_g^2$ is the additive genetic variance. The prior distributions of the variance components are inverse-Gamma distributions. A uniform distribution is used as a prior for $\lambda$, and an independent Beta distribution is adopted as prior for $\gamma$, because of conjugacy.

The joint posterior density of all unobservables is proportional to the product of the augmented sampling model and the priors. This joint posterior distribution is analytically intractable, but MCMC methods such as the Gibbs sampler and the Metropolis-Hastings algorithm can be used to draw samples, from which features of marginal distributions of interest can be inferred. The conditional posterior distributions deriving from it are needed for MCMC implementation. The posterior distribution of $\beta$ and $\boldsymbol{u}$, given all other parameters, is multivariate normal, where the vector $\boldsymbol{w}$ can be interpreted as a 'weight' assigned to $\boldsymbol{y}$ in the analysis. The conditional posterior distributions for each variance component is inverse-Gamma. The fully conditional distribution of each $w_i$ is equivalent to a Bernoulli process. The fully conditional posterior density of the parameter $\gamma$ is Beta. For the parameter $\lambda$, its conditional distributional does not have a close form. A Metropolis-Hastings algorithm, as proposed by Rosa (1999) was tailored for sampling from the distribution of $\boldsymbol{w}$ and $\lambda$, simultaneously. Candidate values are simulated for $\boldsymbol{w}$ and $\lambda$, according to a candidate generator density. The algorithm moves from the current state $\boldsymbol{T} = \{\lambda, w_1 \ldots w_m\}$ to $\boldsymbol{T}^* = \{\lambda^*, w_1^* \ldots w_m^*\}$, with an acceptance probability given as a ratio of densities.

## MATERIAL AND METHODS

**Data.** 13,414 records of yearling weight of Simmental, sired by 1,065 bulls and 5,257 cows, were analyzed and distributed within 1,739 contemporary groups. The pedigree file consisted of 24,271 animals. Variance components and breeding values were estimated by three different approaches : 1) REML and BLUP methodology for a Gaussian (**GML**) model ; 2) MCMC Bayesian methodology for a Gaussian (**BG**) model; and 3) MCMC Bayesian methodology for a mixture (**BM**) model.

**GML analysis**. Variance components were estimated by restricted maximum likelihood methodology considering a Gaussian animal model, with the random effects of breeding values and the fixed effects of the contemporary groups. This analysis was implemented using MTDFREML software (Boldman *et al*., 1993).

**Bayesian analyses**. The same model adopted for the **GML** analysis was also considered within the Bayesian context. Flat priors were assumed for variance components and fixed effects. Marginal densities of the variance components and other genetic parameters were estimated from the Gibbs output. Graphical inspection and the Gibanal program (VanKaam, 1998) were used for assessing convergence to the equilibrium distribution, the joint posterior. A burn-in period of 1,000 iterations was adopted, followed by 250,000 iterations with a thinning interval of 250. Hence, nominal sample size for post-Gibbs analyses was 1,000. The set of MTGSAM programs for mixed model analysis (Van Tassel and Van Vleck, 1995) was used for running Gibbs sampling for the **BG** model.

The **BM** analysis considered the same conditional expectation of the previous models, but the Gaussian distribution of the residuals was replaced by a scale mixture of two Gaussian distributions. A FORTRAN code developed by Dr. D. Sorensen was modified for implementing the analysis under these settings.

**RESULTS AND DISCUSSION**

The **GML** estimates of variance components and heritability and the *posterior* means for the **BG** and **BM** models are presented in Table 1. The estimates of the parameters $\lambda$ and $\gamma$ of the **BM** model were 0.1364 and 0.5943, respectively. This means that about 60% of the animals are from a population that has a residual variance 7.33 times greater than that of the other population (the 40% lasting animals). The **GML** estimates of the additive genetic and residual variances resembled the *posteriori* means of the **BG** model, and the weighted average of the two populations in the **BM** model. The **BM** model seems to be more appropriate for situations where there is heterogeneity of the residual variance but homogeneity of the genetic variance (Pereira *et al*., 2001).

**Table 1. Estimates of variance components and heritability by the GML model, and the *posterior* means, by the BG and BM models**

| Model | $\sigma_g^2$ | $\sigma_e^2$ | $h^2$ |
|---|---|---|---|
| GML | 624.00 | 1612.97 | 0.28 |
| BG | 628.11 | 1611.15 | 0.28 |
| BM[A] | 623.76 | 342.17 | 0.65 |
| BM[B] | 623.76 | 2508.59 | 0.20 |
| BM[C] | 623.76 | 1614.11 | 0.28 |

[A] First Population; [B] Second Population; [C] Weighted average of the two populations

Correlations among predicted breeding values from the three approaches are presented in Table 2. There were very high correlations among all predictions for the whole population. However, when animals are sorted by the predicted breeding values from the **GML** model, and only the first percentile is considered, the correlations involving the results from the **BM** model are lower. From the 242 animals at the first percentile for **GML** and for **BG**, there were, respectively, only 169 and 168 animals that were also in the first percentile for the **BM**. These results suggest that, although for the whole populations there was an apparent agreement among the predictions from the three statistical methods, important changes in rank may be found for the selected population, especially for the males, where only a small part of the population is selected. It is interesting to notice that the correlations between **GML** and **BM** predictions are a little higher than the correlations between **BM** and **BG** results.

**Table 2. Pearson and Spearman correlation coefficients among predicted breeding values obtained by the GML model, and by the *posterior* means for the BG and BM models, for the whole population and for the first percentile[A,B]**

| | Pearson Correlations | | | Spearman Correlations | | |
|---|---|---|---|---|---|---|
| | GML | BG | BM | GML | BG | BM |
| GML | - | 0.99 | 0.98 | - | 0.99 | 0.98 |
| BG | 0.98 | - | 0.98 | 0.98 | - | 0.98 |
| BM | 0.65 | 0.64 | - | 0.57 | 0.56 | - |

[A] Correlations for the whole population and for the first percentile are above and below diagonals, respectively.
[B] First percentile refers to animals with higher genetic values predicted by REML.

## CONCLUSION

The scale mixture of Gaussian distributions is an appealing and flexible alternative to the Gaussian process often used in animal breeding. The Gaussian mixed-effects model is a particular case of the mixture model discussed in this paper, and the adequacy of the normality assumption can be indicated in the analysis of the mixture model. The MCMC implementation of the model is relatively easy, as almost all conditional posterior distributions have a closed form ; and the methodology may be implemented using publicly available software with minor modifications. The results suggest that, although the computations are a little more taxing for the mixture model, it may present more robust inferences in situations with either heterogeneity of variances or presence of outliers.

## REFERENCES

Boldman, K. G., Kriese, L. A., Van Vleck, L. D., VanTassell, C. P. and Kachman, S. D. (1995) U.S. Department of Agriculture Research Service. 115 p.

Pereira, I. G., Rosa, G. J. M.and Oliveira, H. N. (2001) *J. Anim. Sci*, Suppl. 1. **79** : 342.

Rosa, G. J. M. (1999) *Proc. of the Internacional Symposium on Animal Breeding and Genetics*. Viçosa, MG, Sept. 21-24, p.133-159.

Strandén, I. J. and Gianola, D., (1999) *Genet. Sel. Evol*. **31** : 25-42.

VanTassell, C. P. and VanVleck, L. D. A. (1995) Department of Agriculture, Agricultural Research Service. 91 p.

VanKaam, J. B. C. H. M. (1998) http://www.student.wau.nl/~janthijs/breedingsite/eadgibanal.html