A nonparametric Bayesian approach for genetic evaluation in animal breeding

A.L. Pretorius[#] and A.J. van der Merwe
Department of Mathematical Statistics, P.O.Box 339, University of the Free State, Bloemfontein, 9300, South Africa.

This article proposes the Bayesian approach to solve problems arising in animal breeding theory. General elements of Bayesian inferences, e.g. prior and posterior distributions, likelihood functions, and the solving of the random effects in the case of the mixed linear model are discussed. Since the random effects are typically assumed to be normally distributed in both the Bayesian and Classical models, a Bayesian procedure is provided which allows these random effects to have a nonparametric Dirichlet process prior distribution. In the case of the Dirichlet process, the Gibbs sampler is introduced to overcome some computational difficulties in solving the genetic parameters of the mixed linear model. To illustrate the application of these techniques, data from the Elsenburg Dormer sheep stud and data from a simulation experiment are utilized.

Keywords: Mixed linear model, Gibbs sampler, random effect, breeding value, Dirichlet Process Prior, genetics.
[#]Author to whom correspondence should be addressed; e-mail: pretoral@wwg3.uovs.ac.za

Introduction

Since the objective of this paper is to propose the Bayesian approach as a conceptual strategy to solve animal breeding problems, it is appropriate to define "Bayesian inference" and "Bayesian estimation" from the outset. Kendall & Buckland (1971) give the following definitions:

Bayesian inference: A form of inference which regards parameters as being random variables possessed of prior distributions reflecting the accumulated state of knowledge.

Bayes estimation: The estimation of population parameters by the use of methods of inverse probability.

In animal breeding problems, it is usually assumed that the data follows a mixed linear model. When the values of the variance components are not known, the classical approach to the problem of predicting linear combinations of fixed and random effects has been to estimate the variance components using restricted maximum likelihood (REML), and to proceed thereafter as if these estimates were the true values. However, the Bayesian approach has several practical advantages over the classical (REML) approach. Firstly, the estimates from the Bayesian approach for a variance are always positive and an interval estimate such as a highest posterior density region will not include negative values. This is in contrast to the REML estimates. Although REML estimates are defined inside the appropriate parameter space, their asymptotic distributions can generate interval estimates that include negative values (Gianola & Foulley, 1990). This potentially embarrassing phenomenon is often overlooked in discussion of likelihood-based methods. Secondly, highest posterior density regions using the Bayesian approach are never empty, whereas confidence intervals for the ratio of variances can be empty in the case of the classical approach. Finally, one can report the whole of the posterior probability distributions of the parameters in the Bayesian approach (Van der Merwe & Botha, 1993; Wright *et al*., 2000).

Thus, whilst for the REML estimates the variance components are fixed at a single value, ignoring uncertainty associated with estimating their values, the Bayesian approach incorporates this uncertainty by averaging over plausible values of the variance components. This is accomplished through the choice of appropriate prior distributions. Credibility intervals for the estimates can easily be obtained, and all the available information about the random effects to be predicted is contained in the posterior distribution of these random effects. Critics of the Bayesian approach have most often cited the following point: "Bayesian methodology is computer intensive. In many situations integration in several dimensions is required to obtain the posterior densities." This might have been a valid criticism in the past, but by using numerical integration techniques like the Markov Chain Monte Carlo Methods, and more specific Gibbs Sampling, these computational difficulties can be overcome. The joint posterior of all the parameters can now easily be obtained by using the given prior distributions and the likelihood, which generates the data. Characteristics such as the mean, mode, median, variance and phenotypic variance can be found from the marginal posterior distributions. Harville (1990) and Gianola & Foulley (1990) stated that "A more extensive use of Bayesian ideas by animal breeders is desirable and is more feasible from a computational standpoint".

In this communication, we introduce a Sire model to illustrate how the Bayesian approach, using a Dirichlet process prior for the random effects, and Gibbs sampling can be extended to variance component estimation in breeding problems. Estimates of the posterior densities of variance components and functions thereof, such as variance ratio and intraclass correlation are provided. We also propose a Dirichlet process prior for the distribution of the random effects (breeding values), and compare the estimated results with that of the classical results (REML).

Materials and Methods

The Mixed Linear Model is a commonly used statistical tool in animal breeding. As mentioned by Duchateau, Janssen & Rowlands (1998): "The emphasis in breeding experiments is on the variance components and on the prediction of particular random effects in the experiment." In the present study we fit a mixed linear model to the data, and provide a non-parametric Bayesian approach to solve the genetic parameters of the model.

The mixed linear model postulated that the observable random variable is a linear combination of fixed effects and random effects plus a random residual. For illustration purposes, we consider 879 weaning weights, from the progeny of 17 sires, of the Dormer sheep stud at the Elsenburg College of Agriculture near Stellenbosch. The season of birth, age of dam, sex of lambs and birth status effects were included as fixed effects in the model (p = 17). The q = 17 different sires were included as the random effects. An appropriate mixed linear model for the Dormer sheep stud is thus given by

$$y_i = X_i b + z_i u_i + e_i \qquad (1.1)$$

where y_i is n_i x 1, the vector of weaning weights for the lambs (progeny) of the i^{th} sire; X_i: known incident matrix of order n_i x p_i; z_i: vector of n_i elements 1; β: p x 1 vector of uniquely defined fixed effects; u_i: the random effect of the i^{th} sire and ε_i: n_i x 1 vector of random residuals with $\varepsilon_i \sim N(0;\sigma_\varepsilon^2 I_{n_i})$. It is standard in implementation of the model to assume that the random residuals and random effects are normally distributed. From (1.1) it follows that the conditional distribution that generates the data (likelihood function) is

$$y_i \mid b, u_i, s_e^2 \sim N_n\left(X_i b + z_i u_i, I_{n_i} s_e^2\right) \qquad (1.2)$$

where I_{n_i} represents a n_i x n_i identity matrix and $N_n(\underline{\mu}, \Sigma)$ denotes the n-dimensional multivariate normal distribution with mean vector $\underline{\mu}$ and variance-covariance matrix Σ. An integral part of Bayesian analysis is now the assignment of a prior distribution to the unknown parameters in the model. The information contained in the prior distribution is combined with the information supplied by the data into the conditional posterior distribution of the parameters, which is known as the posterior distribution. All inferences about the model parameters are based on the posterior distribution.

As in Kleinman & Ibrahim (1998), we will present a mixed linear model for which the random effects have a nonparametric Dirichlet process prior distribution, i.e. $u_i \sim G$ where $G \sim DP(M \cdot G_o)$. Such a model assumes that the prior distribution G (base measure) itself is uncertain and drawn from a Dirichlet process. The parameters of the Dirichlet process are $G_o = N(0,\sigma_u^2)$, a probability measure, and M. Although it may be hard to quantify, M is a positive scalar that is related to how "clumpy" the data are (often called a precision parameter). Clumpy data occur when the different sires are concentrated into a few clusters (Pretorius & Van der Merwe, 2000a). In practice it is difficult to select appropriate values for this parameter. Instead, it is suggested to place a prior distribution on this parameter, and simulate it given the data. West (1992) assumed that $M \sim G(a,b)$ a gamma prior with $a > 0$ and scale $b > 0$. We may also extend this idea to include a reference prior (uniform for log (M)) by letting $a \rightarrow 0$ and $b \rightarrow 0$. In the final sections of the paper we use the latter, which means that $p(M) \propto M^{-1}$ and $M > 0$. However, a range of possible values is chosen for M and utilized in the analysis.

The foundation of Dirichlet process technology is discussed in Ferguson (1973), West *et al*. (1994) and Walker *et al*. (1999), where the process and its usefulness as a prior distribution are discussed. Useful applications can also be found in Doss (1994), MacEachern (1994), and Lui (1996). Mixture of Dirichlet process priors (MDP) can be of great importance in animal breeding experiments as shown in the rest of the paper. Note that the complex derivations of the posterior distributions are omitted because of the nature of the paper. However, these derivations can be found in the technical reports listed in the reference list.

The nonparametric Bayesian approach for sampling the random effects is to implement the Dirichlet process prior on these effects. It can be described as follows: The draw of the first random effect $u_{i=1}$ is always from its full conditional density

$$h(u_i / b, s_u^2, s_e^2, y_i) = N\{(z_i' z_i + \frac{s_e^2}{s_u^2})^{-1} z_i'(y_i - X_i b); (z_i' z_i + \frac{s_e^2}{s_u^2})^{-1} s_e^2\}, i = 1 \qquad (1.3)$$

The value of the next random effect, u_i, can then either be set equal to one of the other random effects, u_j , $i \neq j$ with a certain probability,

$$\frac{f(y_i \mid X_i b + z_i u_j; s_e^2 I_{n_i})}{I_i + \sum_{j=1; j\neq i}^{q} f(y_i \mid X_i b + z_i u_j; s_e^2 I_{n_i})} \tag{1.4}$$

or could be drawn from (1.3) with probability

$$\frac{I_i}{I_i + \sum_{j=1; j\neq i}^{q} f(y_i \mid X_i b + z_i u_j; s_e^2 I_{n_i})} \tag{1.5}$$

where

$$I_i = M(2p)^{-\frac{n_i}{2}}(s_u^2)^{-\frac{1}{2}} Q_i^{\frac{1}{2}}(s_e^2)^{-\frac{n_i}{2}} \exp\left\{\frac{1}{2s_e^2}(y_i - X_i b)'\Omega_i(y_i - X_i b)\right\} \tag{1.6}$$

$$Q_i = \left[\frac{1}{s_e^2}(z_i' z_i) + \frac{1}{s_u^2}\right]^{-1} \quad \text{and} \quad \Omega_i = \left(\frac{1}{s_e^2} z_i Q_i z_i' - I_{n_i}\right)$$

Thus, each summand in the conditional posterior distribution of u_i is separated into two elements. The first element is a mixing probability, and the second is a distribution to be mixed. The values of the probabilities in (1.4) and (1.5) change with each new draw of a random effect.

There is some plausible intuition behind this above mixture scheme. If the breeding value, u_i of sire i has a relatively large residual using sire j's breeding value, then u_j is relatively less likely to be chosen as the breeding value of sire i. Conversely, if the breeding value of sire i has a relatively small residual using sire j's breeding value, then the random effect u_j is relatively more likely to be chosen as the breeding value of sire i. On the other hand, the greater the residual for sire i, the greater the probability p (in 1.5) that sire i will get a new value from $h(\cdot,\cdot)$ in (1.3). This scheme results in what MacEachern (1994) calls a "cluster structure" among the different sires. This cluster structure partitions the q different sires into k groups, where $0 < k \leq q$. Thus, all the sires in a specific cluster will have identical breeding values and sires in different clusters will have different breeding values. This may sound farfetched, but since the Gibbs sampler is repeated several times, the algorithm leads to reduced variation and hence faster convergence of the estimated random effects to their true values. The average of the simulated values for each breeding value is then computed, thus every sire will have its own unique breeding value.

In a Bayesian sense, a fixed effect is a random variable on which prior knowledge is diffuse or vague. Thus, a constant prior will often be used for these effects (Gianola & Fermnando, 1986). In the model mentioned in the previous section and for our data set, "flat" or uniform priors distributions are assigned to the fixed effects, β and the model variance, σ_ε^2, as to represents this lack of prior knowledge. Therefore

$$p(\beta, \sigma_\varepsilon^2) = p(\beta)\, p(\sigma_\varepsilon^2) \propto \text{constant} \tag{1.7}$$

and for the sire variance,

$$p(\sigma_u^2) \propto \text{constant} \tag{1.8}$$

The fully conditional posterior density for each of the above unknowns is obtained by regarding all other parameters in the joint posterior as known. The required conditional posterior for β is

$$p(b / u, s_e^2, y) = N_p\left\{b^*, s_e^2 \left(\sum_{i=1}^{q} (X_i' X_i)\right)^{-1}\right\} \tag{1.9}$$

where

$$b^{*} = (\sum_{i=1}^{q}(X_i'X_i))^{-1}\sum_{i=1}^{q}X_i'(y_i - z_iu_i)$$

For σ_ε^2 we have

$$p(s_e^2 / b, u, y) = K_e\{\prod_{i=1}^{q}(\frac{1}{s_e^2})^{\frac{n}{2}}\}\exp\left\{-\frac{1}{2s_e^2}\sum_{i=1}^{q}(y - X_ib - z_iu)'(y - X_ib - z_iu)\right\}; s_e^2 > 0 \quad (1.10)$$

an Inverse Gamma distribution, where

$$K_e = \left\{\frac{\sum_{i=1}^{q}(y - X_ib - z_iu)'(y - X_ib - z_iu)}{2}\right\}^{\frac{1}{2}(n-2)}\frac{1}{\Gamma(\frac{n-2}{2})}$$

Also $u = (u_1, u_2,\ldots,u_q)^T$, $y = (y_1^T, y_2^T,\ldots,y_q^T)^T$ and $\sum n_i = n$ the sample size.

Finally, for the sire variance σ_u^2 we have

$$p(s_u^2 \mid b, u, s_e^2, y) = K_u\left(\frac{1}{s_u^2}\right)^{\frac{q}{2}}\exp\left\{-\frac{1}{2s_u^2}u'\mathbf{A}^{-1}u\right\}, s_u^2 > 0 \quad (1.11)$$

also an Inverse Gamma distribution, where

$$K_u = \left\{\frac{u'\mathbf{A}^{-1}u}{2}\right\}^{\frac{q-2}{2}}\frac{1}{\Gamma\left(\frac{q-2}{2}\right)}$$

and A the numerator relationship matrix with elements reflecting the genetic relationship among the sires.

From (1.3) - (1.5) it is clear that this relationship matrix is not taken into account when the random effects are estimated. Indeed, when considering A in the estimation process, it means that some of the sires are considered to be more equal than others because they are related. The Dirichlet process on the other hand chooses the random effects in such a way that the sires will be grouped into clusters or groups. Thus, taking the aforementioned into account, surely one can argue that the Dirichlet process fulfills, in a certain sense, the same role as the relationship matrix, and by omitting it, cannot be seen as a serious limitation (Kleinman & Ibrahim, 1998).

All the necessary posterior distributions are now known for the model parameters, thus the Gibbs sampler can be implemented. The Gibbs sampler is a technique for generating random variables from a marginal distribution indirectly, without having to calculate the density. Suppose we are given a joint density f(x,y) and are interested in obtaining characteristics of the marginal density

$$f(x) = \int f(x, y)dy \quad (1.12)$$

such as the mean or variance. Perhaps the most natural and straightforward approach would be to calculate f(x) analytically or even numerically, and then use this result to obtain the desired characteristics. However, there are some, perhaps many, cases where the marginal distributions cannot be found. Rather than compute or approximate the marginal distribution directly, the Gibbs sampler allows us to effectively simulate a random sample

$$X_1;X_2;\ldots;X_n \sim f(x)$$

without requiring f(x). By simulating a large enough sample, the mean and variance or any other characteristic of f(x) can be calculated to the desired degree of accuracy. The Gibbs sampler is therefore a method of generating a sample from the marginal distribution by sampling from the full conditional distributions f(x/y) and f(y/x) which are usually known linear statistical models. The ultimate value of the Gibbs sampler lies in its simplicity and practical potential. For further reading, refer to articles by Geman & Geman (1984), Gelfand & Smith (1990), Casella & George (1992) and Wang *et al*. (1993).

We start to construct a stochastic process that has the desired posterior distribution as its stationary distribution and then simulate the process. We begin with a set of starting values for the model parameters, $\beta^{(0)}$, $u^{(0)}$, $\sigma_\varepsilon^{2\,(0)}$, $\sigma_u^{2\,(0)}$, and the successively generate values from the conditional posterior distribution of each of the parameters, conditioning on the most recently generated values of the other parameters of each step.

Thus, the Gibbs sampler for $p(\beta, u, \sigma_\varepsilon^2, \sigma_u^2 \mid D)$ is:

(0) Select starting values for $u^{(0)}$, $\sigma_\varepsilon^{2\,(0)}$, $\sigma_u^{2\,(0)}$. Set i = 0
(1) Sample $\beta^{(i+1)}$ from (1.9)
(2) Sample $\sigma_\varepsilon^{2\,(i+1)}$ from (1.10)
(3.1) Sample $u_1^{(i+1)}$ according to the mixture scheme, (1.3) - (1.5)
(3.2) …
…
(3.q) Sample $u_q^{(i+1)}$ according to the mixture scheme, (1.3) - (1.5)
(4) Sample $\sigma_u^{2\,(i+1)}$ from (1.11)
(5) Set i=i+1 and return to step (1)

MATLAB software has been developed to generate the samples that enabled us to obtain the marginal posterior densities for the model parameters, using the Gibbs sampler. The full conditional posteriors are updated after every iteration. We run multiple chains (m=101 000) of the Gibbs sampler to obtain these draws from the posterior distributions given the data. The first 1 000 draws of each chain are discarded, and then every 10^{th} draw is saved. By saving every 10^{th} draw, the chain yielded a posterior sample of 1 000 approximately uncorrelated draws. All posterior analyses are based on these 1 000 draws, giving us a full Bayesian solution to all the mixed linear model parameters.

Results and Discussion

The weaning weights of 879 lambs from the progeny of 17 sires of the Elsenburg Dormer sheep stud are utilized in the study. An experimental data set is then used to compare Bayesian and REML results of the genetic parameters. REML estimates (MTDFREML program, Boldman *et al*., 1995) were compared with estimates obtained with the Dirichlet process (MDP). A range of values (M = 5, 50 and 500) was used for the precision parameter of the Dirichlet process prior. The variance components obtained are reported in Table 1. The estimated marginal posterior density of the sire variance, σ_u^2 is given in Figure 1, and variance ratio coefficient, $v = \frac{s_u^2}{s_e^2}$ in Figure 2.

A commonly derived statistic, the heritability (h^2) of the trait, which is a function of the two variance components is also calculated and reported in Table 1. It describes the proportion of the total variation in the environment of the study attributable to genetics. In this formula $h^2 = \frac{4s_u^2}{s_u^2 + s_e^2}$, σ_u^2 is multiplied by four in the numerator to account for the fact that lambs from the same sire are half siblings and the sire accounts for half of the inherited genetic component, and $\sigma_u^2 + \sigma_\varepsilon^2$ is the phenotypic variance. The higher the heritability, which lies between zero and one, the higher the proportion of the total variation that can be assumed to be genetic in origin.

Table 1 Estimated values for the variance components, σ_u^2, σ_ε^2 and h^2 as obtained by the Gibbs sampler.

	REML	MDP, M = 5	MDP, M = 50	MDP, M = 500
σ_u^2	3.08	3.21	3.42	3.53
σ_ε^2	21.11	21.62	21.31	21.24
h^2	0.51	0.52	0.55	0.57

From the results in Table 1, it is clear that the Dirichlet process that we have presented is certainly valuable in the case of variance estimation in animal breeding. With the MDP (M = 500) reported the highest proportion of

total variation that can be assumed to be genetic (57.0%). A 95% credibility interval for σ_u^2 in the case of MDP, M = 500, is given by [2.2070 ; 4.4306], and for σ_ε^2 by [20.5333 ; 21.9641].

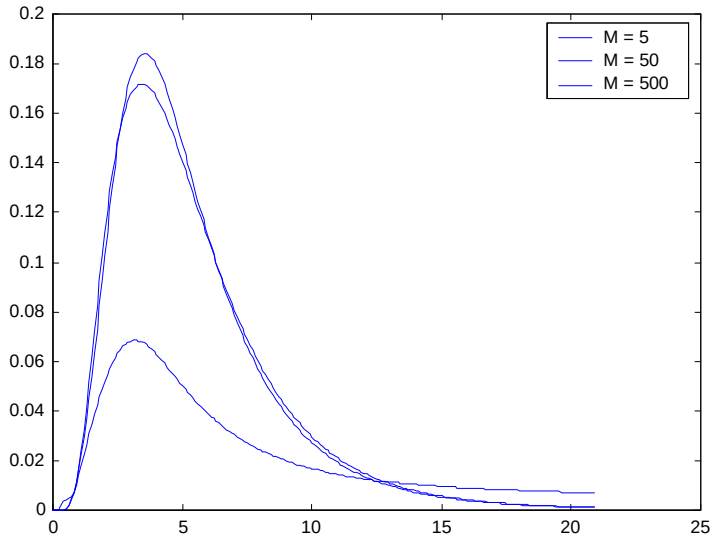

Figure 1 Estimated marginal posterior density of the sire variance $ó_u^2$. Posterior modes are : for M = 5 : 3.21; M = 50 : 3.42 and M = 500 : 3.50

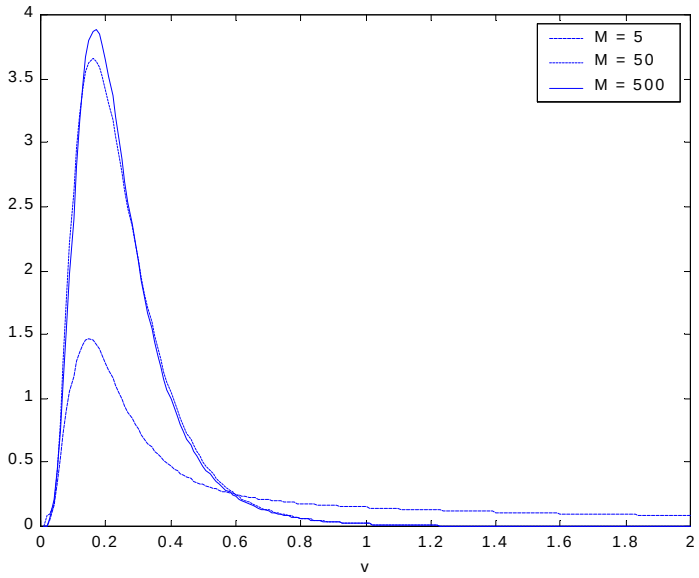

Figure 2 Estimated marginal posterior density of the variance ratio $v = \frac{ó_u^2}{ó_{å}^2}$

Apart from estimating heritability and sire variance components, an animal breeder might be interested in the breeding value of specific sires since the objective of a breeding experiment is to determine which sires should be retained for future selection. As seen before, the Gibbs sampler can obtain such predictions. In Table 2 the estimated breeding values for the 17 sires are reported. Also, note that the posterior modes of the estimated posterior distributions are reported for M = 5. The estimated REML values are also included in the table.

It is evident from this table that the estimates for larger values of M are quite close to the REML estimates. One must still remember that some differences may be ascribed to the number of iterations used in the Gibbs

sampler. The question which immediately arises from the above, is "why are the posterior modes used for M = 5?" In answering this question, lets look at the estimated marginal posterior densities for different values of M, of the breeding value of sire ID41019 with breeding value u_3. These posterior densities are presented on Figures 3-5.

Table 2 The estimated breeding values of the 17 sires in the Dormer sheep stud. REML estimates and their standard errors (SE's) are also reported.

Sire_ID	REML	SE's for REML	MDP, M = 5	MDP, M = 50	MDP, M = 500
41004	0.139	0.92	-0.535	0.567	0.661
41019	3.330	0.99	4.547	4.026	3.723
41037	0.578	1.06	2.377	1.224	1.241
43002	-1.181	1.18	-0.649	-1.019	-1.059
44042	-1.256	0.95	-1.267	-1.247	-1.247
44170	-0.170	1.18	-1.705	-1.336	-0.287
44174	-0.569	1.34	-0.746	-0.697	-0.532
45070	-0.963	0.93	-0.901	-0.475	-0.902
45135	-0.537	1.10	-0.890	-0.601	-0.244
46015	-1.709	0.96	-1.590	-1.282	-1.652
46037	-0.842	0.91	-0.785	-0.375	-0.779
48014	-0.853	0.97	-1.361	-0.853	-1.018
48052	-0.302	1.00	-0.781	-0.023	-0.287
48140	-1.256	1.10	-1.783	-0.749	-1.199
49046	0.406	1.41	2.336	0.277	0.561
49053	0.463	1.31	2.482	0.319	0.992
49134	0.795	1.34	2.807	1.070	1.232

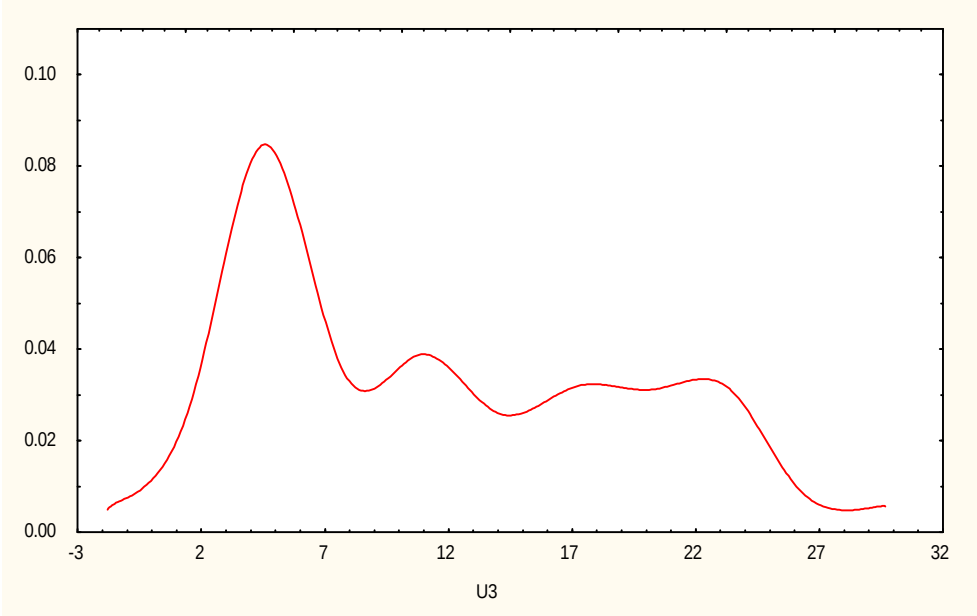

Figure 3 Estimated marginal posterior density of the breeding value for sire ID41019, with M = 5

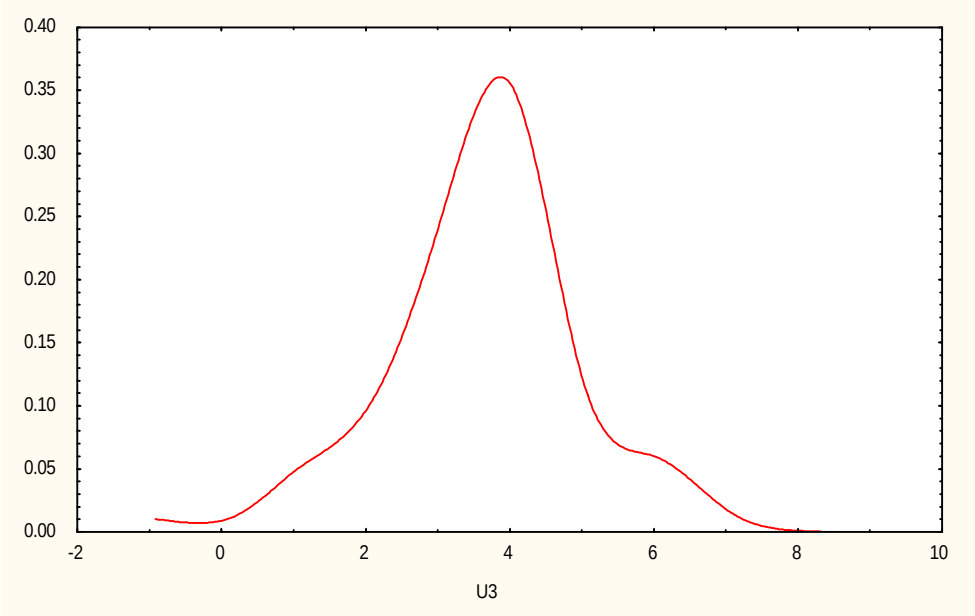

Figure 4 Estimated marginal posterior density of the breeding value for sire ID41019, with M = 50

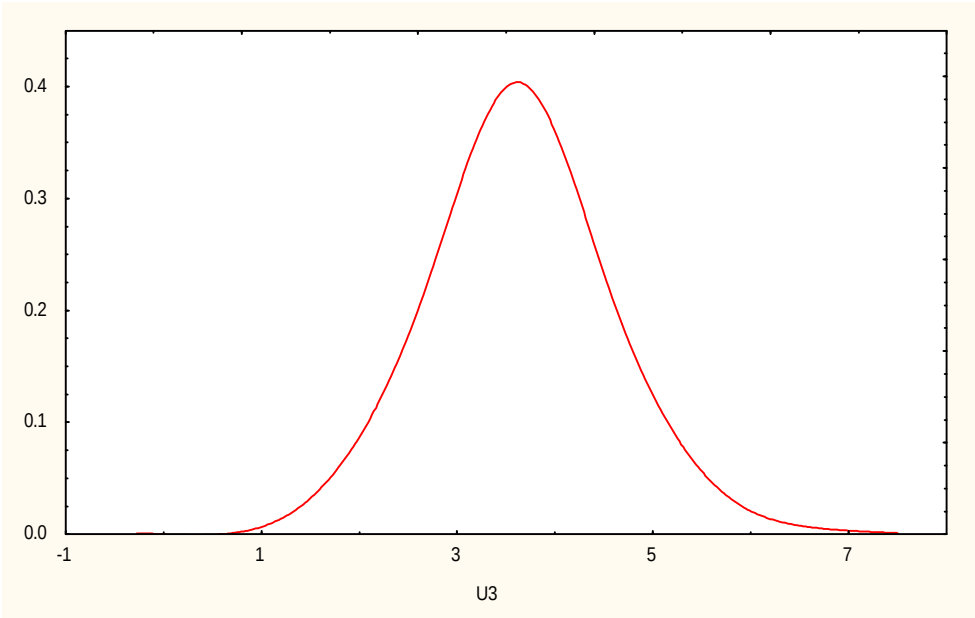

Figure 5 Estimated marginal posterior density of the breeding value for sire ID41019, with M = 500.

It is clear from Figure 3 that the density of the breeding value u_3, reflects a large departure from normality when M = 5. Moreover, this figure also reflects uncertainty about the underlying prior for u_3. The Dirichlet process prior therefore becomes more parameter free resulting in this "strange" shape of the estimated marginal posterior density. There is also some uncertainty about the exact location of the mean and height in this density. As M increases a moderate departure is noticed in Figure 4 (shape of the density tends to become more bell-shaped), whereas no departure of normality is evident in Figure 5, i.e. when M = 500. This is to be expected, as the random effects are directly affected by the relaxation of the normal assumptions when M is reasonably small. Thus it would be more appropriate to find the posterior modes of these densities, rather to calculate the means.

It is very important to note that the resulting breeding values are quite sensitive to the choice of this precision parameter, M. Much care should go into the selected values for M since different results may be obtained when the choice of M is not appropriate. For more discussion on the choice of M, see Escobar (1994).

Some of the estimated fixed effects are presented in Table 3. It is clear from the values of these estimates that the choice of the precision parameter M does not influence the fixed effects, nor is it affected by the relaxation of the normal assumption when M is small. These minor differences are attributable to the little mass of the Dirichlet process on the fixed effects.

Table 3 Estimated values for selected fixed effects in the model.

	MDP, M = 5	MDP, M = 50	MDP, M = 500	REML
β_{14}	3.669	3.716	3.666	3.603
β_{15}	9.601	9.512	9.502	9.527
β_{16}	3.09	2.977	3.01	2.949

Much of the current research focused on the distributional properties of Bayesian models compared to classical models. At present one would suggest the results of these models should be compared by means of partial F-tests, residual analysis, cross validation (data-splitting) or tests of overall model adequacy. However, model validation involves an assessment of how the fitted models will perform in practice, i.e. how successful it will be when applied to new or future data. An experiment will thus be conducted where all the genetic parameters in the mixed linear model are known, and the dependent variable will be built. These parameters, using REML and Bayesian methods (Dirichlet process) will then be estimated.

Again, consider the following mixed linear model $y_i = X_i b + z_i u_i + e_i$

where $\varepsilon_i \sim N(0;\sigma_v^2 I_n)$ and for which the random effects have a nonparametric Dirichlet process prior distribution, i.e. $u_i \sim G$ where $G \sim DP(M \cdot G_o)$. The parameters of the Dirichlet process are $G_o = N(0,\sigma_u^2)$, a probability measure, and M = 100. The values for $\sigma_\varepsilon^2 = 4.88$ and $\sigma_u^2 = 0.7211$. The only fixed effect in the model is sex ($\beta_1 = 0.705$). Thus, the male lambs weigh on average 0.705 kg more than female lambs. A total of 200 sires are added as random sire effects to the model. For each sire, 10 male and 10 female weaning weights are generated using the Dirichlet process.

Table 4 reports the estimated variance components used for the experimental data set. A small difference between the estimates and the actual values of these variance components are observed, indicating that the Bayesian approach using the Gibbs sampler is certainly valuable and worthwhile in the context of animal breeding and selection.

Table 4 Estimated variance components for the experimental data set using the Dirichlet process and REML

	MDP, M = 100	MDP, Sim M	REML
σ_u^2	0.765	0.768	0.772
σ_ε^2	4.932	4.935	4.929

Table 5 contains the estimated breeding values of the first 10 sires in the data set along with their rankings. The second column (EXP) in the table is the actual breeding values of the sires. The fourth and sixth columns contain the results when a Dirichlet process is implemented in the Gibbs sampler. For the fourth column, the precision parameter M is set equal to the true value, i.e. 100. The next step is to simulate this parameter from the data (Pretorius & van der Merwe, 2000b). The results are reported in the sixth column (MDP, Sim M*).

Table 5 Estimated breeding values of 10 of the 200 sires in the experimental data set and posterior rankings.

Rank	EXP	Sire_ID	MDP, M = 100	Sire_ID	MDP, Sim M*	Sire_ID	REML	Sire_ID
1	1.4702	9	1.413	9	1.4188	9	1.44597	9
2	1.3975	2	1.3183	2	1.3239	2	1.32166	2
3	0.381	5	0.913	10	0.8727	10	0.87709	10
4	0.2997	1	0.2711	5	0.256	5	0.26953	6
5	0.146	10	0.2381	6	0.2142	6	0.2679	5
6	-0.082	6	-0.0503	7	-0.1292	7	-0.02923	7
7	-0.2163	8	-0.1544	3	-0.2231	3	-0.15932	3
8	-0.2835	11	-0.2535	1	-0.2281	1	-0.16895	1
9	-0.3977	4	-0.2812	11	-0.3324	11	-0.26275	11
10	-0.4936	7	-0.4713	8	-0.4889	8	-0.44905	8

From the table we can show that the Dirichlet process in Bayesian inference regarding breeding experiments is a very promising method. According to the experimental data, it is known that sires 9,2,5,1 and 10 are ranked as the five best sires in the model. The Dirichlet process ranked sires 9,2,10,5 and 6 as the best animals. This is an 80% success rate in the ranking procedure. Sires 9,2,10,6 and 5 are also ranked as the best sires by the REML analysis. In the next section dealing with the model adequacy, the SSE (sum square errors) is calculated and reported in Table 6 below:

Table 6 The calculated sum square errors for the different models

	REML	MDP, M = 100	MDP, Sim M
SSE	46.145	46.440	44.833

Note that $SSE = \sum_{i=1}^{200} (u_i - \hat{u}_i)^2$ where u_i is the actual breeding value, and $\hat{u}_i$ the estimated breeding value.

From the results in Table 6, it is believed that the Bayesian nonparametrics, using the Gibbs sampler, have as much to offer as the REML analysis. Since the posterior densities resulting from the Gibbs sampler can easily be used to construct confidence intervals for the model parameters, the potential mathematical consequences of the toolkit that is explored here in the world of the animal breeder is evident.

Conclusion

In this paper we have applied a general technique for Bayesian nonparametrics to an important class of models, the mixed linear model for animal breeding experiments. An easy method to calculate posterior densities for the genetic parameters of this model has been presented. The important contribution of this paper revolves around the nonparametric prior distribution, the Dirichlet process prior, for the random effects in the case of the mixed linear model, and to correctly model and interpret these estimated random effects. It is also clear from the example that the choice of the precision parameter M largely influences the posterior densities of these effects. The choice of M will determine whether the estimates from the Dirichlet process behave like the nonparametric Bayes estimator or like the parametric empirical estimator. Further work suggested by this paper includes allowing for the simulation of M from the data as implemented in the experimental data set in the previous section. In choosing a prior distribution for M based on the number of clusters among the sires, one can propose a mixture of gamma distributions for the posterior distribution of M.

As far as Bayesian approaches are concerned, the very real advantage of being able to input broad prior ideas of the model parameters is noted, as well as the much richer and more tractable forms of inference that are been made possible by the Gibbs sampler-based approach to computations of genetic parameters for animal breeding problems. It is suggested that future research should be concerned with the study of intraclass correlation in the mixed linear model.

Acknowledgment

The Institute for Animal Improvement and Production (IRENE) supplied the Dormer data. Also, a special word of thanks to Prof. JB van Wyk for conducting the REML analysis.

References

Boldman, K.G., Kriese, L.A., van Vleck, L.D., van Tassell, C.P. & Kachman, S.D., 1995. A manual for use of MTDFREML. A set of programs to obtain estimates of variances and covariances. US Dept of Agriculture, Agricultural Research Service

Cassela, G., & George, E.I., 1992. Explaining the Gibbs sampler. Amer. Statist. 46, 167-174.

Doss, H., 1994. Bayesian nonparametric estimation for incomplete data via successive substitution sampling. Annals Statist. 22, 1763-1786.

Duchateau, L., Janssen, P., & Rowlands,G.L., 1998. Linear Mixed Models. An introduction with applications in veterinary research. ILRI, Nairobi, Kenya, 159-170.

Escobar, M.D., 1994. Estimating Normal Means with a Dirichlet Process Prior. J.A.S.A., 89(425), 289-275.

Ferguson, T.S., 1973. A Bayesian analysis of some nonparametric problems. Ann. Statist., 1, 209-230.

German, S., & German, D., 1984. Stochastic relaxation, Gibbs distribution, and the Bayesian Restoration of image. I.E.E.E. Trans. Pat Anal. and Mach. Intel. 6, 721-741.

Gelfand, A.E., & Smith, A.F.M., 1990. Sampling-based approaches to calculating marginal densities. J.Amer. Statist. Assoc. 85, 398-409.

Gianola, D., & Fernando, R.L., 1986. Bayesian Methods in Animal Breeding Theory. J. Anim. Sci. 63, 217-244.

Gianola, D., & Foulley, J.L., 1990. Variance Estimation from Integrated Likelihoods (VEIL).Genet. Sel. Evol. 22, 403-417.

Harville, D.A., 1990. BLUP and beyond. In: Advances in Statistical Methods for Genetic Improvement of Livestock. Eds. Gianola, D., and Hammond, K., Springer-Verlag, NY-Heidelberg-Berlin. pp. 239-376.

Kendall, M.G., & Buchland, W.R., 1971. A Dictionary of Statistics terms. Hafner, New York.

Kleinman, K.P., & Ibrahim, J.G., 1998. A semi-parametric Bayesian approach to the random effect model. Biometrics 54, 921-938.

Liu, J., 1996. Nonparametric hierarchical Bayes via sequential imputations. Annals Statist. 24, 911-930.

MacEachern, S.N., 1994. Estimation normal means with a conjugated style Dirichlet process prior. Commun. Statist. – Simula, 23,727-741.

Pretorius, A.L. & van der Merwe, A.J. 2000a. Semi-parametric Bayes inference for the mixed linear model. Technical report No. 274. Department Mathematical Statistics, University of the Free State, RSA.

Pretorius, A.L. & van der Merwe, A.J. 2000b. Bayesian estimation in animal breeding using Dirichlet process prior. Technical report No. 277. Department Mathematical Statistics, University of the Free State, RSA.

Van der Merwe, A.J., & Botha, T.J., 1993. Bayesian Estimation in Mixed Linear Models using the Gibbs Sampler. South African Statist. J., 27, 149-180.

Wang, C.S., Rutledge, J.J., & Gianola, D., 1993. Marginal Inference about Variance Components in a Mixed Linear Model using Gibbs Sampling. Genet. Sel. Evol. 25, 41-62.

Walker, S.G., Damien, P., Laud, P.W., & Smith, A.F.M., 1999. Bayesian nonparametric inference for random distributions and related functions. JR Statist. Soc. B, 61, 485-527.

West, M., 1992. Hyperparameter Estimation in Dirichlet Process Mixture Models. Technical Report 92-A03. Duke University, ISDS.

West, M., Muller, P., & Escobar, M.D., 1994. Hierarchical prior and mixture models, with applications in regression and density estimation. In: Aspects of uncertainty: A tribute to D.V. Lindley, Eds. Smith, A.F.M. and Freeman, P.R. Wiley, London.

Wright, D.R., Stern, H.S., & Berger, P.J., 2000. Comparing Traditional and Bayesian Analysis of Selection Experiments in Animal Breeding. J Agri. Bio. Environ. Statist. 5, 240-256.

.