ADVANCES IN BAYESIAN METHODS FOR QUANTITATIVE
GENETIC ANALYSIS

D. Gianola 1, R. Rekaya 2, G.J.M. Rosa 3 and A. Sanches 4

1 Department of Animal Sciences, University of Wisconsin-Madison, WI 53706, USA
2 Department of Animal and Dairy Science, University of Georgia, Athens, GA 30602, USA
3 Departments of Animal Sciences and of Fisheries and Wildlife, Michigan State University,
   East Lansing, MI 48824-1225, USA
4 Departamento de Ciencias Exatas, FCAV/UNESP, 14870-000 Jaboticabal SP, Brazil

INTRODUCTION

Arguably, the “Bayesian controversy”, in the sense of Blasco (2001), is over or will be over
soon. Bayesian methods are employed now routinely in archaeology, artificial intelligence,
disease mapping, ecology, economics, forest management, physics, signal processing and last,
but not least, quantitative genetics and genomics. For example, Shoemaker et al. (1999) write :
“…Bayesian approaches can be easier to interpret and they have been employed in many
 genetic areas, including : the classification of genotypes and estimating relationships,
 population genetics and molecular evolution, linkage mapping (including gene ordering and
 quantitative risk analysis) and quantitative genetics [including quantitative trait (QTL)
 mapping]”. This should be augmented to include transcriptional analysis (Newton et al., 2001).
Walsh (2001), in a discussion of quantitative genetics in the age of genomics, has conjectured
 that, in the next 20 years, Bayesian procedures will likely displace their likelihood based-
counterparts. This is gratifying to some animal breeders that have been advocating the
Bayesian approach for 20 years or so (e.g. Gianola and Foulley, 1982; Gianola and Fernando,
1986). Further, and particularly in connection with Markov chain Monte Carlo, Bayesian
methods can be used for fitting any model, irrespective of its complexity (e.g. Gilks et al.,
1996). This gives these procedures an awesome power, in strictu sensu.

Bayesian methods still pose major difficulties, apart from the MCMC computations. In fact,
calculations will become heftier if perfect sampling methods transcend the research stage ;
here, the Markov chain is in the stationary distribution at the onset of iteration (Casella et al.,
2001). At any rate, one of the bottlenecks is the elicitation of priors when there is no
mechanistic basis for the probabilistic structure, or when the analyst is refractory to
introducing prior input into the inferences. On the other hand, the methods can cope with
a vast array of data driven challenges. Consider, for example, robustness of inferences in mixed
effects linear models. Animal breeders have mined the Gaussian assumption well, but have
also abused it. In our field, there seems to be a prevailing view that size (of the model) is what
matters, with minor concern about the distributional assumptions made. Is the multivariate
normal distribution always sensible ? This assumption is extremely, so there is a chance that it
will be violated in multiple-trait or in longitudinal analyses. In these cases, it makes sense
using robust distributions, at least at some level of a hierarchical model. While this can be a
horrendous task in a frequentist-likelihood setting, it is straightforward in a Bayesian setting. See Strandén and Gianola (1998; 1999) and Gianola (2001) for a discussion.

Our objective is to present a brief account of advances that are being made in connection with these difficulties. In the Second section, a discussion of the current state of research about priors in variance component models is presented. In the Third section, the power of the Bayesian methodology is highlighted via a review of some robust procedures for inferences in hierarchical linear and generalized linear models. These include thick-tailed distributions, error of measurement models, Bayesian model averaging (BMA) and semi-parametric processes. The Fourth section illustrates the flexibility of the Bayesian approach for modeling complexity through a discussion of Bayesian mixture models for analysis of somatic cell count (SCC) data. It is assumed that the reader is familiar with basic concepts of Bayesian inference. An account of Bayesian methods in quantitative genetics is in Sorensen and Gianola (2002). For an inventory of current research areas of activity in Bayesian analysis, see Berger (2002).

WHAT PRIOR?

In a hierarchical model, frequentists, likelihoodlums and Bayesians fanatics probably agree in that it is sensible to view additive genetic effects as drawn from a multivariate normal random process. The reason is the infinitesimal construct of Fisher (1918), giving a mechanistic basis; Bayesians view this distribution as a statement of prior uncertainty. For other parameters, e.g. “fixed” effects and (co)variance components, the classical schools do not introduce any prior (probabilistic) input, whereas Bayesians encounter the problem of eliciting a prior. One of the criticisms made to Bayesian inference is the effect that a possibly subjective, arbitrary or misguided prior can have on inferences. Hence, efforts have been directed at arriving at "objective priors", that is, prior distributions that say "little" relative to contributions made by the data. The effect of the prior dissipates as sample size increases, so the problem is one of finite sample inferences. The state of prior knowledge can range from almost complete ignorance about the value of the parameters to a fair bit of information.

In a situation of ignorance, arriving at objective or non-informative priors is extremely difficult, and consensus seems to be lacking. For details, see Bernardo (1979) and Berger and Bernardo (1992). Some \textit{ad hoc} approaches used in quantitative genetics have included improper uniform priors, vague distributions and Jeffreys rules. The latter fail in multi-parameter situations. For example, in an analysis of variance model with several “treatments”, Jeffreys rule does not recognize that degrees of freedom must be lost as the number of treatments increases. The most promising approach, albeit technically complex, is called “reference” analysis (Bernardo, 1979; Berger and Bernardo, 1992). It is based on information-theoretic arguments, and it is inextricable of the experimental design, of the decision target (the parameter of interest) and of the nuisance parameters.

If some knowledge is available, one must be tempted to assign an informative prior distribution to the unknown parameter, with hyper-parameters elicited in some manner. For example, adopt an inverted Gamma distribution for the residual variance. However, prior knowledge may be such that one can go up to specifying only some features of the prior, but
not the entire distribution. Hence, if one adopts an inverse Gamma distribution when, say, only the mean and variance of the prior can be specified, more prior input is introduced than what is dictated by available knowledge. Here, the issue is to specify a prior not conveying information beyond that truly available a priori. The solution is the maximum entropy formalism, developed by Jaynes, a physicist. Cantet et al. (1992) were the first ones that suggested maximum entropy priors for Bayesian analysis of quantitative genetic parameters.

**Reference analysis.** The amount of information that an experiment $e$ is expected to provide about a parameter $\theta$ is defined as:

$$I[e, p(\theta)] = \int \left\{ \int \log \left[ \frac{p(\theta | y)}{p(\theta)} \right] p(0 | y) d\theta \right\} g(y) dy$$

The outer expectation is taken with respect to the marginal (in the Bayesian sense) distribution of the observations, but conditionally on the experimental design adopted. The $p(.)$ denotes prior or posterior densities. The notation $I[e,p(\theta)]$ makes explicit that this measure of information is actually a functional of $p(\theta)$, that is, a function that depends on another function, this being the density $p(\theta)$. Suppose now that the experiment is replicated $K$ times, and that all data vectors are concatenated successively. If the experiment could be replicated an infinite number of times, one would be in a situation of perfect or complete information about the parameter. Thus, when $K \to \infty$, the resulting $I[e,p(\theta)]$ measures, in some sense, the missing information about $\theta$ expressed as a function of the prior density $p(\theta)$. A "non-informative" prior would then be that maximizing the missing information. The prior density $p(\theta)$ which maximizes the functional above can be found via calculus of variations and, typically, the solution is not explicit. However, when an asymptotic approximation to the posterior exists, the reference prior is Jeffreys rule for a single parameter situation. When there are nuisance parameters (typical case), one must establish an ordered parameterization (with the parameter of interest singled out), and then follow this procedure : 1) find the reference prior for the nuisance parameter, conditionally on the parameter of interest. 2) Integrate the likelihood over the preceding conditional reference distribution. 3) Apply a single-parameter reference prior algorithm. When the asymptotic approximations exist, the density of the reference prior distribution can be written from Fisher’s information matrix (all unknowns with proper prior distributions must be integrated out first). The algebra is formidable. Once the reference prior is found, the posterior analysis is carried out in the usual manner. The reference prior is typically improper and non-conjugate, thus complicating MCMC computations.

We have examined the reference prior in some variance component models, and so have Du Plessis and van der Merwe (1996). In a one-way layout where primary parameter was heritability, we found essentially no differences between the reference analysis and one in which Jeffreys rule was applied separately to the mean, to the residual variance and to heritability. Also, in a simulation with a model containing several “fixed” effects, individual additive genetic effects and a residual term (2 variance components), there were no sizable differences between the posterior distributions obtained with 3 different analyses : 1) a reference prior in which heritability was the primary parameter, 2) bounded flat priors for fixed
effects and variance components, and 3) flat priors for the fixed effects and scaled inverted chi-square distributions for the variance components.

**Maximum entropy priors.** An account of maximum entropy as a means of allocating probabilities is in an unfinished book by the late Jaynes. Suppose that a prior distribution is to be assigned to some unknown quantity. This prior should take into account whatever information is available, but not more. For example, knowledge of average values (or of other aspects of the distribution) will give a reason for favoring some possibilities over others but, beyond this, the distribution should be as uncommitted as possible. Further, no possibilities should be ruled out, unless dictated by prior knowledge. The information available defines constraints that fix some properties of the prior distribution, but not all. Jaynes argued that the only measure of uncertainty represented by a probability distribution is entropy. Then, he proposed that a distribution maximizing entropy, subject to the constraints imposed by the information available, would represent the "most honest" description of what is known about a set of propositions. The entropy of a prior distribution with density \( p(\theta) \) is defined as:

\[
H[p(\theta)] = -\int [\log p(\theta)] p(\theta) \, d\theta
\]

This must be maximized subject to constraints imposed by available information \( I \). The integral above is a functional so this is also a calculus of variation problem. For example, if one specifies the first and second moments of the prior distribution, there would be 3 constraints. The first one ensures propriety of the prior distribution, and the other two fix the values of the moments. In general, if \( M-1 \) moments are known, the density of the maximum entropy distribution (relative to a uniform measure) has the form:

\[
p(\theta) = \frac{\exp\left(1 + \sum_{i=1}^{M} \lambda_i \theta_i\right)}{\int \exp\left(1 + \sum_{i=1}^{M} \lambda_i \theta_i\right) \, d\theta}
\]

The \( \lambda \)’s are Lagrange multipliers introduced to obey the constraints posed by the information \( I \); typically, these must be calculated numerically. It can be shown that if only the prior mean is specified, the maximum entropy distribution is exponential. If the mean and variance are known, the maximum entropy distribution is normal. For example, suppose that one goes to some of the thousands of tables published in Animal Breeding Abstracts (CAB has been printing these tables in anticipation of formal Bayesian analyses), and computes the mean and variance of estimates of the heritability of some trait. The maximum entropy prior distribution follows immediately: since heritability is positive (this being part of the information \( I \) available), the maximum entropy distribution is truncated normal. Subsequently, this can be combined with the rest of the Bayesian structure of the problem, and the joint posterior density follows. Typically, the maximum entropy distribution does not lead to recognizable conditional posterior distributions, complicating the MCMC process somewhat.

**ROBUST INFERENCE**

**Thick-tailed distributions.** One is often confronted with observations that are strange in some sense, apart from recording errors. In dairy cattle, discrepancies between “expected” and estimated breeding values have been attributed to outliers resulting from preferential treatment.
of cows (Strandén and Gianola, 1998; 1999). Methodologies for robust inference have been suggested, e.g. thick-tailed and skewed distributions for residuals in regression models (Lange and Sinsheimer, 1993; Fernández and Steel, 1998). For binary and polychotomous responses, thick-tailed distributions (such as the Student-t process) have been proposed for robust inferences in threshold models (Albert and Chib, 1993; Gianola and Sorensen, 1996); applications are in Kizilkaya and Tempelman (personal communication) and in Chang et al. (2002). Alternatives to the t distribution, such as the slash and the contaminated normal processes, have also been implemented in cross-sectional and longitudinal hierarchical linear models (Rosa, 1999; Rosa and Gianola, 2001). Rodriguez-Zas (1998) used multivariate and univariate-t processes in longitudinal models for somatic cell score (SCC) in Holsteins. The normal distribution is a special case of the so-called normal/independent distributions, corresponding to \( t \) or slash distributions with a non-finite robustness parameter, \( \nu \), or to a contaminated normal distribution with no contaminants. The robust models are more general and flexible than the Gaussian. The normal/independent distributions provide automatic outlier detection, and the models can be fitted via MCMC.

Error distributions (and perhaps the distribution of genetic effects) may be asymmetric. Fernández and Steel (1998) proposed a procedure for introducing skewness into symmetric distributions. von Rohr and Hoeschele (1999) discussed the skewed Student-t distribution, and illustrated the methodology. Rosa et al. (2001) fitted skewed Gaussian models to data from a fertility experiment with Holstein cows.

**Bayesian model averaging.** Consider a survival analysis of dairy cows. The information may consist of covariates such as herd, sire, year-season of birth, molecular markers and last known survival status, since censoring is common. The objective may be to assess effects of explanatory variables, or to predict survival time of the future progeny of some sires. Hence, one searches for some survival model, and finds that a proportional hazards model \( M_1 \) fits well, and that it gives sensible parameter estimates. However, another proportional hazards model \( M_2 \) also fits well, but it gives different estimates and predictions. Which model should be used at the end? Typically, models are selected in some *ad hoc* manner, and inferences are based on the model chosen eventually, as if there were no uncertainty about it. In Bayesian analysis, however, the model can be treated as an item subject to uncertainty. The posterior distribution of the "model random variable" is used to obtain inferences that take into account automatically the relative plausibility of the models under consideration. This is called Bayesian model averaging (BMA); Hoeting *et al.* (1999) give the details. These authors argue as follows: since part of the evidence must be used for model selection, ignoring the uncertainty about the model leads to an overstatement of precision. Theoretical arguments show that BMA can be used to enhance the predictive ability of an analysis, e.g. of lactation curves (Jensen, 2001)

**Measurement error models.** Some authors have used measurement error models for binary outcomes, such as Rekaya *et al.* (2001), in a study of fertility in dairy cattle, and Rosa *et al.* (2001), in genetic map construction when genotypes are miscoded. In the last study, results
suggested that the approach provides more reliable estimates of genetic maps, unless there is a strong reason to believe that genotypes are ascertained without error.

**Dirichlet process priors.** There is also uncertainty associated with the distributions of effects in hierarchical models (Escobar and West, 1998). A nonparametric model avoiding the specification of such distributions is a logical alternative, and Dirichlet process priors are the cornerstone of nonparametric Bayesian modeling. A Dirichlet process prior on a distribution places a distribution on the space of all distribution functions (Escobar and West, 1998). Two hyper-parameters govern a Dirichlet process. One of these defines the location of the prior, and the other one is a “precision parameter”, defining the concentration of the prior around the location. Kleinman and Ibrahim (1998), Escobar and West (1998) and Pretorius and van der Merwe (2000) give an account of how Dirichlet processes can be incorporated in linear models. We had 3214-test day milk yield records from 341 Holstein cows with complete lactations. Wood’s nonlinear curve with 3 parameters was fitted in a hierarchical linear model in which the effects on trajectory parameters were herd-year-seasons (13 classes), age of cow (5 classes) and the additive genetic values of cows. The pedigree file included 703 animals. A standard Bayesian model under normality and homogeneous residual variance was compared to one in which a Dirichlet process was fitted to the trajectory parameters, with the rest of the model being the same. Results pointed away from the normality assumption, and differences in parameter estimates were large. For example, the posterior mean of the genetic correlation between $a$ and $c$ parameters of Wood’s curve was −0.44 and 0.28 in standard and Dirichlet process specifications, respectively. The results suggest that inferences can be sensitive to the functional form assumed for at least some distribution in the hierarchy.

**MODELLING WITH BAYESIAN MIXTURES**

To illustrate the flexibility of the Bayesian approach, consider modeling SCC as an indicator of mastitis. Detilleux and Leroy (2000), referred to as DL hereinafter, suggested a 2-component mixture model for SCC in dairy cattle. They stated that it is not sensible to view SCC as a draw from some single distribution. Selection for low SCC may be harmful, as neutrophils intervene in protection mechanisms against infection. Also, perhaps a high SCC plays a role in protecting the mammary gland. DL indicate that a mixture model can account for the effect of infective status on SCC, and that it can give estimates of prevalence, as well as of a probability of status (infected versus non-infected) for a cow, given the data. A theoretical advantage of a mixture model is that it can represent a complex distribution in situations where a single parametric family fails. The approach of DL can be extended in several directions using the Bayesian approach. Our suggested extensions must not be construed as criticism of DL. We regard their paper as an important contributions made towards the analysis of SCC. Rather, the extensions highlight that Bayesian inference can enrich the specification in a number of respects.

First, it is known that there is differential expression of genes in diseased and healthy individuals. Mixture models are flexible enough to allow for such differential expression. This can be done by introducing a genetic correlation in the mixture model, much in the same way that one can think of a genetic covariance between ovulation rate and scrotal circumference in sheep, or between direct and maternal effects in cattle. In the context of mastitis, and in a 2-
component mixture, a sufficient condition for identification of the genetic correlation is that some healthy animals have relatives that contract the disease. Second, as noted by DL, the assumption that genetic and environmental variances are the same for each of the two components of the mixture can be relaxed. Third, their estimates of conditional probabilities of membership for each cow do not take into account the uncertainty associated with parameter estimates. In mixture models, it is not always clear if asymptotic properties of the estimates hold as soon as one departs from standard settings, even under normality. Estimates of conditional probabilities based on maximum likelihood estimates of parameters, acting as if these were the true values, must be viewed with caution. Although they assumed homogeneous genetic and residual variances in each of the two populations to ensure a global maximum of the likelihood function, this is at the expense of realism. One would expect variances to be larger in the "infected" component of the mixture, merely from scaling considerations. If the model calls for heterogeneous dispersion parameters, likelihood-based inference would run the risk of placing the (conceptual) asymptotic distribution at a local stationary point, whereas the Bayesian analysis would integrate over the entire parameter space. Fourth, the approach of DL does not include hierarchical modeling of the probability of membership. In their model, they assume that the probability is the same for every observation. However, the chance of infection is known to vary between years, herds and sire families. Also, and in the context of longitudinal information, the evidence indicates that the incidence of mastitis is higher around calving than at other stages of lactation. It would be sensible to allow for a time-varying probability of membership. More generally, for the mixture model to be useful in selection, it would be desirable to express these probabilities as a function of the breeding values or transmitting abilities, as recognized by DL. Fifth, mastitis is a complex disease, and more than two components may be needed for adequate modeling. There are several different types of bacterial agents, perhaps leading to different distributions of SCC. Hence, a mixture model with an unknown number of components may be in order, although the difficulty of implementing it is clear. A Bayesian approach with an unknown number of components may require reversible jump Markov chain Monte Carlo, and the difficulties of tuning this procedure are notorious. Assuming that models with \(2, 3, \ldots, K\) components are equi-probable \textit{a priori}, one can choose one using the Bayes factor or, even better, infer breeding values or future observations via BMA.

REFERENCES
Bernardo and Smith (1994) "Bayesian Theory", Wiley, New York, USA.