MODEL SELECTION FOR PREDICTION OF BREEDING VALUES

D. Sorensen\(^1\) and R. Waagepetersen\(^2\)

\(^1\) Danish Institute of Agricultural Sciences, Dept. Anim. Breed. and Genetics, Denmark
\(^2\) Aalborg University, Dept. of Mathematical Sciences, Denmark

INTRODUCTION

Advances in computer-based inference methods have opened the way for fitting models to data with an enormous amount of flexibility. Unless there is a well established theory or mechanism underlying the problem at hand, agreement about the model to be used is more the exception than the rule. The scientist is then confronted to making a choice among the models entertained, and drawing inferences by selecting the most appropriate, in some sense, among these. Another route to follow is to base inferences on several of the models posed, accounting for the models’ posterior uncertainty (e.g. Hoeting et al., 1999).

There is a vast literature on model criticism and model choice, both from classical and Bayesian perspectives. Some of the methods are well founded theoretically, others are more exploratory and eclectic. Rather than providing a review of all methods, we concentrate on two approaches. Whether in a Bayesian or classical framework, one possibility is based on measures of model fit. For example, a widely used technique is to construct standardized residuals and to check whether their values fall within 95\% probability intervals, say, from the sampling or posterior distribution of the residual. If this is not the case, the particular data point is regarded as suspicious, and depending on the proportion of such outliers, this may be construed as an indication that the posited model does not fit well. There are not necessarily specific alternative models being compared here. In a second approach, whether couched in the Neyman-Pearson hypothesis testing framework and extensions of it (as in Cox, 1961, 1962), or in that advocated by Jeffreys (1961) and Savage (1962), several models are weighed against each. In this paper both approaches are implemented, since they illuminate different aspects of the model which are relevant from the point of view of prediction of breeding values. One of them provides a global measure of a model’s relative plausibility and is well founded theoretically. The other is more exploratory, but nonetheless is particularly apposite since it quantifies the predictive ability of the model.

A Bayesian perspective with a slight frequentist flavour in one of the methods is adopted here, and implementation is via Markov chain Monte Carlo (MCMC). Regardless of partisan issues, the Bayesian approach to inference is attractive when applied to model choice. Some of the attractions are: 1) it is nice to be able to make statements about the posterior probability of a model. This is more straightforward to interpret than “the probability of obtaining a value for the test statistic at least as extreme as the one observed, given the null hypothesis” under the sampling distribution of the test statistic in conceptual replications of the experiment carried out in identical conditions; 2) there are advantages in averaging over the distribution of (unknown) parameters rather than maximizing, especially in multi-modal, highly dimensional likelihoods; 3) contrary to frequentist tests, the Bayesian counterparts do not tend to reject null
hypotheses systematically in large samples, and tend to choose the correct model with probability 1 as the number of observations tend to infinity; 4) contrary to the likelihood ratio criterion, the Bayesian test favours the most parsimonious model, particularly as the amount of data increases, provided it fitted the data almost as well as the more complex model; 5) classical tests were developed for the comparison of two nested models. When the models are not nested, one must face the issues raised by the arbitrary designation of one of the two models as the null hypothesis. Further, with several models at play, there is the thorny problem of multiple frequentist tests. On the other hand, the Bayesian approach allows taking model uncertainty into account. Thus, the multiple testing problem is avoided and comparisons among models are straightforward: one chooses the most probable model on the basis of the data at hand; 6) With the Bayesian approach, there are no conceptual problems in testing a value of the parameter in the border of the parameter space. Last but not least, when used together with MCMC, Bayesian criteria are extremely flexible and easy to obtain. There are also limitations in some of the Bayesian methods, as we point out below.

This paper is organized as follows. In the next section, the two criteria for model choice are briefly and formally introduced, and we indicate how they can be implemented using MCMC. This is followed by an example using data from a large-scale selection experiment for litter size in pigs. Three models are fitted and compared using the criteria outlined below. The paper ends with a discussion and conclusions.

CRITERIA FOR MODEL COMPARISON

Posterior probability of a model and Bayes factors. Suppose the investigator poses \( K \) competing models. A prior probability \( \Pr(M_i), i = 1,2,...,K, \) is assigned to model \( i \), where \( \sum_{i} \Pr(M_i) = 1 \). The models do not need to have a nested structure. After observing data \( y \), the posterior probability of model \( i \) is obtained applying Bayes theorem:

\[
\Pr(M_i | y) = \frac{\Pr(M_i) p(y | M_i)}{\sum_{i} \Pr(M_i) p(y | M_i)}.
\]

(1)

where \( p(y | M_i) \) is the marginal or prior predictive density of the data, given model \( i \). It can be interpreted as the probability of obtaining the observed data under model \( i \), before these data became available. The posterior odds ratio of model \( i \) relative to model \( j \) is equal to

\[
\frac{\Pr(M_i | y)}{\Pr(M_j | y)} = \frac{\Pr(M_i) \Pr(y | M_i)}{\Pr(M_j) \Pr(y | M_j)}.
\]

(2)

The Bayes factor is defined to be:

\[
B_{ij} = \frac{p(y | M_i)}{p(y | M_j)} = \frac{\Pr(M_i | y) \Pr(M_i)}{\Pr(M_j | y) \Pr(M_j)}.
\]

(3)
the ratio between the posterior odds of the models and the prior odds, which is equal to the
former when the prior probabilities of the models are equal. It provides a measure of whether
the data have increased or decreased the odds on \( M_i \) relative to \( M_j \). The interpretation of (3) is
immediate: a value of \( B_{ij} \) equal to 100 say, means that the observed data is 100 times more
likely to have been generated under model \( i \) than under model \( j \). Jeffreys (1961) suggested a
way of calibration which can provide a rough descriptive statement about standards of
evidence of one model (hypothesis) versus another. In general though, interpretation may
depend on the context. In terms of (2), if the prior probability of the models are equal, then a \( B_{ij} \)
equal to 100 means that model \( i \) is 100 times more probable than model \( j \), or if only two models
are involved, that the posterior probability of model \( i \) is 0.99 and of model \( j \) is 0.01.

The numerator of (3) is equal to:

\[
p(y \mid M_i) = \int p(y \mid \theta_i, M_i) p(\theta_i \mid M_i) \, d\theta_i = E_{\theta_i \mid M_i} \left[ p(y \mid \theta_i, M_i) \right],
\]

which is an averaged likelihood, with the prior density \( p(\theta_i \mid M_i) \) as the weighting function.
Unless the parameters of models \( i \) and \( j \) are fully specified, the Bayes factor is influenced by
prior information. Further, improper prior distributions leads to impropriety of (4). Several
ways of dealing with this problem have been suggested. O’Hagan (1994) and Kass and Raftery
(1995) provide an overview as well as a comprehensive review of Bayes factors.

Schwarz (1978) proposed an approximation to (3) that brings insight, does not require eliciting
prior distributions and is easy to compute. It is known as the Bayesian information criterion
\((BIC)\), and is derived using a Taylor series approximation to (4), which results in:

\[
BIC \approx 2 \ln B_{ij} \approx 2 \ln \left( \frac{p(y \mid \hat{\theta}_i, M_i)}{p(y \mid \hat{\theta}_j, M_j)} \right) - (p_i - p_j) \ln n,
\]

where \( p(y \mid \hat{\theta}_i, M_i) \) is the maximized likelihood under model \( i \), \( p_i \) is the number of parameters
of model \( i \) and \( n \) is the length of vector \( y \). When \( p_i \geq p_j \) the \( BIC \) is smaller than twice the log-
likelihood ratio, so the adjustment favours parsimony.

*Computation of the Bayes factor using MCMC.* Despite the possible dependence on the prior,
Bayes factors are one of several useful tools for model comparison. The influence of prior
information on inferences can (and should) be studied via a sensitivity analysis. With the
advent of MCMC methods, many methods for computing Bayes factors have been proposed in
the literature. A recent comparative review is in Han and Carlin (2001). Here we use the
method proposed by Newton and Raftery (1994). The Monte Carlo consistent estimator is
equal to:
\[ \hat{p}(y | M_i) = \left[ \frac{1}{m} \sum_{j=1}^{m} p^{-1}(y | \theta_i^{(j)}, M_i) \right]^{-1}. \]  

(5)

In (5), \( m \) is the number of MCMC samples and \( \theta_i^{(j)} \) is the \( j^{th} \) MCMC draw from the distribution \( [\theta_i | y, M_i] \).

**Model comparison using posterior predictive distributions.** Bayes factors, like the likelihood ratio criterion, provide overall measures of the relative adequacy of a model. However, a model can be the most plausible within a set of competing models, and yet be unable to predict the data at hand well, or to give reasonable predictions of specific features of the data or of future observations. This consideration is relevant if the model will be used for selection purposes. Here, one may also be interested in studying whether the various models rank candidates for selection differently. The method that we propose addresses several aspects of the fit of the model and was advocated by Guttman (1967). Many variations and extensions have been proposed recently, especially with a view towards exploiting the power of MCMC. The general idea is as follows: if the model fits, then replicated data simulated under the model should produce something similar to the observed data, at least similar in “relevant ways”, as Rubin (1984) expresses it. This replication is at the core of frequentist inference; the difference here is that the simulated data are generated conditional on the observed data \( y \), from the posterior predictive distribution with density:

\[ p(y_i^* | y, M_i) = \int p(y_i^* | \theta_i, M_i) p(\theta_i | y, M_i) d\theta_i, \]  

(6)

where \( y_i^* \) is the vector of simulated data under model \( i \), \( \theta_i \) is the vector of parameters of model \( i \) and \( p(\theta_i | y, M_i) \) is the posterior distribution of \( \theta_i \). A variant of this method based on cross-validation has been formalized recently by setting it in a decision-theoretic framework (Gelfand and Ghosh, 1998)

**Simulation from the posterior predictive distribution.** Draws from (6) can be obtained as follows: 1) obtain the \( j^{th} \) draw \( \theta_i^{(j)} \) from \( [\theta_i | y, M_i] \) using some MCMC algorithm; 2) draw \( y_i^{(j)} \) from \( [y_i^* | \theta_i^{(j)}, M_i] \). Repeat steps 1) and 2). The set of \( y_i^{(j)} \), \( j=1,2,..., \), constitute extractions from (6). Imagine that one wishes to test a specific feature of the observed data. Label this \( h(y) \), a function of the data (and perhaps of \( \theta_i \)) that maps the data space to the real numbers. One can then study whether \( h(y) \) falls in a region of high posterior density in the simulated distribution \( [h(y_i^*) | y, M_i] \). The function \( h(y_i^*) \) is constructed from the Monte Carlo draws \( y_i^{(j)} \). This is repeated for all the models under investigation. Some criterion such as 95% probability intervals from \( [h(y_i^*) | y, M_i] \) can be used to discriminate among models.
AN EXAMPLE

Results from work in progress that illustrate the two approaches described above are presented. The data originate from a large scale selection experiment for litter size in pigs that was conducted from 1989 to 1992, in which selection of high intensity was applied once only. Details can be found in Sorensen et al. (2000). Briefly, selection was based on predicted genetic values from a repeatability additive genetic model, using data from registered breeding herds. One selection and one control line were kept in a common research farm. The data file consisted of 10,060 litter size records from 4,175 sows. The pedigree file consisted of 6,437 individuals; the oldest were born in 1981.

The models. First the full model is defined, labeled model F hereinafter. This model was introduced in the genetics literature by SanCristobal et al. (1998), who extended previous work by, e.g., Foulley et al. (1992). SanCristobal et al. (1998) chose a likelihood implementation via the EM algorithm. Here, a Bayesian MCMC approach is followed. The sampling distribution of data \( y \) (total number of born piglets per litter) is assumed to be as follows:

\[
y | \beta, a, p, \sigma^2_{e_y} \sim N(X\beta + Za + Wp, I \sigma^2_{e_y}). \tag{7}
\]

In (7), the vector \( \beta \) contains parameters for the effects of the 9 parities, 4 seasons, 70 herds and the two types of insemination (artificial and natural service). The vector \( a \) contains 6,437 additive genetic values and the vector \( p \) contains 4,175 permanent environmental effects. Matrices \( X, Z \) and \( W \) are known incidence matrices and \( I \) is the identity matrix of order 10,060×10,060. The scalar \( \sigma^2_{e_y} \) is the residual variance component associated with parity \( i \) and sow \( j \). The following structure is assumed for \( \sigma^2_{e_y} \):

\[
\sigma^2_{e_y} = \exp(k_{p(i)} + \tilde{a}_j), \quad i=1,2,\ldots,nj; \quad j=1,2,\ldots,q. \tag{8}
\]

where \( p(i) = 1 \) if \( i = 1 \) and 2 otherwise, \( k_{p(i)} \) models the effect (on the variance) of parity 1 if \( p(i) = 1 \), and any other parity other than 1 if \( p(i) = 2 \), \( \tilde{a}_j \) is the additive genetic value and \( n_j \) is the number of parities of individual \( j \). The model postulates different residual variances for first parity and the rest, and assumes that this residual variance is under genetic control. Vectors \( a \) and \( \tilde{a} \), each with \( q \) elements, are assigned the following Gaussian distribution:

\[
a | A, G_0 \sim N\left(0 \bigg| G_0 \otimes A \right), \tag{9}
\]

where \( A \) is the known numerator relationship matrix of order \( q \times q \) and the \( 2 \times 2 \) unknown matrix \( G_0 \) has elements.
In (10), $\sigma^2_a$ is the additive genetic variance component contributing variation in litter size records, and $\sigma^2_p$ is the additive genetic variance component contributing variation to the conditional variance of a litter size record, given $b$, $a$ and $p$. The scalar $\rho$ is the additive genetic correlation. The prior for the vector of permanent environmental effects $p$ is:

$$p \mid \sigma^2_p \sim N(0, \sigma^2_p I)$$

where $\sigma^2_p$ is the permanent environmental variance. The vector $b$ was assigned a normal distribution a priori, with mean 0 and known variance, $10^6$. All variance components where assigned independent scaled inverted chi-square distributions, and the correlation in (10) was a priori uniform between –1 and 1. The second model, labeled model $H$ (for heterogeneous variance), is similar to the previous one, except that the variance (8) is not assumed to be under genetic control. In model $H$,

$$\sigma^2_{e_i} = \exp(k_{pi(i)}), i=1,2,\ldots,n_i; j=1,2,\ldots,q.$$  

Thus model $H$ is the usual heterogeneous variance model. Finally, the third model labeled model $S$ for standard, assumes no structure for the residual variance (8), and it is therefore the common additive genetic repeatability model with homogeneous residual variance.

**MCMC implementation.** Very briefly, the MCMC algorithm applied was as follows. Location parameters associated with the vector $b$ were sampled from their fully conditional posterior distribution in one pass. The remaining of the location parameters $p$, $(k_1,k_2)$, $(a,\tilde{a})$ were sampled in turn, conditional on the other parameters, using a so-called Langevin-Hastings algorithm. All variance components were sampled singly, conditional on the value of the remaining parameters, using Metropolis-Hastings with a random walk proposal. In order to improve mixing a reparameterization was used to transform random effects $(a,\tilde{a})$ into independent standard normal random effects using the square root of $G_0 \otimes A$.

**Results.** Table 1 shows marginal posterior means (second row) and 95% probability intervals (third row, in brackets) of selected parameters for model $F$. There is clear evidence for residual variance heterogeneity associated with parity one versus the rest, and for genetic variation contributing to environmental sensitivity. The sign and magnitude of the genetic correlation (~0.58) indicates that individuals which genetically produce larger litters, tend to show a more homogenous litter size across parities. The 95% posterior probability intervals indicate that all parameters shown in the table have probability mass a long way from zero.
Table 1. Posterior mean and 95% probability intervals of selected parameters of model F

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Posterior Mean</th>
<th>95% Probability Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma^2_\rho$</td>
<td>0.63</td>
<td>(0.33; 0.95)</td>
</tr>
<tr>
<td>$\exp(k_1)$</td>
<td>5.31</td>
<td>(4.69; 5.98)</td>
</tr>
<tr>
<td>$\exp(k_2)$</td>
<td>7.52</td>
<td>(6.74; 8.35)</td>
</tr>
<tr>
<td>$\sigma^2_\nu$</td>
<td>1.54</td>
<td>(1.16; 1.96)</td>
</tr>
<tr>
<td>$\sigma^2_\eta$</td>
<td>0.11</td>
<td>(0.08; 0.15)</td>
</tr>
<tr>
<td>$\rho$</td>
<td>-0.58</td>
<td>(-0.72; -0.40)</td>
</tr>
</tbody>
</table>

See text for definition of symbols. 95% posterior intervals in brackets.

MCMC estimates of the log-prior predictive density of the data for the three models are shown in the second line of table 2. The third line in the table shows the 95% predictive coverage, i.e. the percentage of all the 10,060 litter size records that fall within their 95% posterior prediction interval.

Table 2. MCMC estimates of the log-marginal posterior density of the data and of 95% predictive coverage using all the data

| Model | $\ln p(y | M_i)$ | % Predictive coverage $^1$ |
|-------|----------------|---------------------------|
| Model F | -24,159 | 98.5 |
| Model H | -24,427 | 97.6 |
| Model S | -24,511 | 97.4 |

$^1$95% predictive coverage. See text for definition of model symbols.

The following conclusions can be drawn from the results in table 2. The values of the log-prior predictive densities confer an overwhelming relative advantage to model F. It is easy to show from the figures in the table that the posterior probability of model F is practically 1 for any reasonable set of prior probabilities chosen. In contrast, the percentage predictive coverage does not distinguish clearly among the 3 models. In all cases, the figures tend to be too large. This may be due to some degree of overfitting, since few records are available for the three random effects that are fitted to each sow with records in the case of model F.

In order to study the behaviour of the models in the tails of the distribution, 95% posterior predictive coverage was computed for the 50 largest order statistics. The sizes of the litters in this group were 18, 19, 20 and 21. The results are summarized in table 3.

Table 3. MCMC estimates of 95% predictive coverage for the 50 largest order statistics and 95% predictive intervals for the maximal litter size

<table>
<thead>
<tr>
<th></th>
<th>Model F</th>
<th>Model H</th>
<th>Model S</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Predictive coverage $^2$</td>
<td>2</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>95% Predictive interval $^2$</td>
<td>(21.4; 26.2)</td>
<td>(21.4; 25.4)</td>
<td>(21.0; 24.8)</td>
</tr>
</tbody>
</table>

$^2$95% predictive interval for the maximal record (maximal observed record: 21 piglets).

The coverage percentages are far from 95% for all three models. In all cases, the 50 largest observed records tend to fall to the left of their 95% posterior intervals. This can be partly explained by the discreteness of litter size: only 4 categories exist among the 50 largest litters.
The ranking of the models is different of that under the marginal likelihood criterion, although discrimination here is a little vague. The last line in the table shows the range of the 95% predictive interval for the maximal litter size. The average range of the prediction intervals for the 50 highest order statistics was 3.3, 2.7 and 2.3 for models $F$, $H$ and $S$, respectively. Other comparisons, e.g. based on the posterior probability of additive genetic values being among the 50 largest are being studied and will be reported at the congress.

**DISCUSSION AND CONCLUSIONS**
In typical statistical problems in the biological sciences one is almost certain a priori that the model implemented is incorrect in a non-trivial way. In the case of the litter size experiment reported here, a continuous sampling model was postulated for discrete data, and a model with an infinite number of genes acting additively was assumed a priori. However, the concern is to decide whether the model(s) can be useful for some specific purpose. This decision should be guided by the purpose(s) for which the model will be used. Different models may well be chosen under different criteria tested with the same data. MCMC offers enormous flexibility to study model adequacy. Here we focused on the posterior probability of the model and related quantities (Bayes factors, prior predictive distributions), and on features derived from posterior predictive distributions. Especially for prediction purposes, one would in general also like to study the performance of BMA. This requires the posterior probability of the model as input. In the experiment reported here there is no room for BMA, because the posterior probability of model $F$ is practically 1.

The analyses of the litter size experiment produced conflicting results. We should not be disturbed by this, but rather by the absence of conclusive statements. More features from posterior predictive distributions are being studied; we have not left the scientific drawing board yet.

**REFERENCES**