SORT 36 (1) January-June 2012, 69-80

Comparing and calibrating discrepancy measures for Bayesian model selection

Julián de la Horra[1] and María Teresa Rodríguez-Bernal[2]

Abstract

Different approaches have been considered in the literature for the problem of Bayesian model selection. Recently, a new method was introduced and analysed in De la Horra (2008) by minimizing the posterior expected discrepancy between the set of data and the Bayesian model, where the chi-square discrepancy was used. In this article, several discrepancy measures are considered and compared by simulation, and it is obtained that the chi-square discrepancy is reasonable to use. Then, an easy method for calibrating discrepancies is proposed, and the behaviour of this approach is studied on simulated data. Finally, a set of real data is analysed.

MSC: 62F15

Keywords: Bayesian model selection, discrepancy measure, calibration, posterior expected discrepancy.

1. Introduction

In recent years, many articles have been written about Bayesian model selection. Many of these articles have relied on Bayes factors or posterior odds; for instance, Spiegelhalter and Smith (1982), Aitkin (1991), O'Hagan (1995), Berger and Pericchi (1996). Other papers have considered a predictive approach; for instance, Geisser and Eddy (1979), San Martini and Spezzaferri (1984), Gelfand et al. (1992), Gelfand (1995), Laud and Ibrahim (1995), Gelfand and Ghosh (1998), Gutiérrez-Peña and Walker (2001), Trottini and Spezzaferri (2002), De la Horra and Rodríguez-Bernal (2005, 2006).

[1] Departamento de Matemáticas. Universidad Autónoma de Madrid. 28049 Madrid (Spain). julian.delahorra@uam.es

[2] Departamento de Estadística e Investigación Operativa. Universidad Complutense de Madrid. 28040 Madrid (Spain). mayter@mat.ucm.es

Received: October 2010
Accepted: February 2012

A different and easy method was proposed and analysed in De la Horra (2008). This method was based on the well-known property that, under the true model, the cumulative distribution function is distributed as a uniform distribution over the interval (0, 1). A suitable discrepancy measure between the sample and the Bayesian model is needed. In De la Horra (2008), the chi-square (χ^2) discrepancy was used but, of course, this is not the only discrepancy measure we may consider. The main aims in this article are:

1. To carry out a comparison between the performance of the χ^2 discrepancy and the performance of other possible discrepancy measures.
2. To calibrate the discrepancy we find between the set of data and the selected model.

The article is organized as follows:

In Section 2, the method introduced in De la Horra (2008) is briefly explained (for any discrepancy measure). In Section 3, several discrepancy measures are proposed. In Section 4, these discrepancy measures are compared by simulation, and it is found that the χ^2 discrepancy is reasonable to use.

Once we have decided to use a discrepancy measure and we have chosen a model, we have to remember that this model is not to be understood as the true model (because nobody knows the true model) but as the best model among several possible models. The discrepancy between the data and the model is just a number, and it is very important to decide if this number indicates either a small discrepancy or a large discrepancy:

- If the discrepancy is small, the model we have chosen is a good model for our data.
- If the discrepancy is large, the model we have chosen is not a good model for our data.

This problem of calibrating discrepancy measures has been previously studied, for instance, by McCulloch (1989), Soofi et al. (1995), and Carota et al. (1996). In Section 5, an easy procedure for calibrating the discrepancy between the set of data and the selected model is considered. Some examples (control cases) are analysed for illustrating and evaluating this method.

Finally, a set of real data is analysed in Section 6.

2. A method for model selection

An easy method for Bayesian model selection was proposed and developed in De la Horra (2008). This method was based on the use of a discrepancy measure and it is briefly explained here for the continuous case (although a modification was also given for its application to the discrete case).

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random sample of a continuous random variable X. We have to choose among m different Bayesian models, M_i, $i = 1, \ldots, m$. Each

Bayesian model consists of two components: a sampling density, $f_i(x|\theta)$ (where $\theta \in \Theta$), and a prior density, $\pi_i(\theta)$. For the sake of simplicity, we are assuming that the parameter space, Θ, is the same for all the models, but this is not necessary. The cumulative distribution function corresponding to $f_i(x|\theta)$ will be denoted by $F_i(x|\theta)$; this cumulative distribution function will have a relevant role in the method. In short, we can write:

$$M_i = \{f_i(x|\theta), \pi_i(\theta)\},\ i = 1, \ldots, m.$$

The method is based on the following idea:

Let us assume that $\mathbf{X} = (X_1, \ldots, X_n)$ is a random sample from a continuous random variable X with density function $f_i(x|\theta)$ and cumulative distribution function $F_i(x|\theta)$ (for θ fixed). It is well known that $(F_i(X_1|\theta), \ldots, F_i(X_n|\theta))$ can be considered as a random sample from a $U(0,1)$ (uniform distribution over the interval (0, 1)), because $F_i(X|\theta)$ follows a $U(0,1)$ distribution and, as a consequence, we hope that $(F_i(X_1|\theta), \ldots, F_i(X_n|\theta))$ will be well fitted by the $U(0,1)$ distribution.

We next describe the method in three steps:

(1) First of all, we measure the discrepancy between the sample we have obtained, $\mathbf{x} = (x_1, \ldots, x_n)$, and the distribution function $F_i(x|\theta)$ (for a fixed θ), by using a suitable discrepancy measure between $(F_i(x_1|\theta), \ldots, F_i(x_n|\theta))$ and the $U(0,1)$ distribution. This discrepancy will be denoted by $D_i(\mathbf{x}, \theta)$.

The idea behind this discrepancy is simple: if $F_i(x|\theta)$ (for a fixed θ) is a good model, $D_i(\mathbf{x}, \theta)$ will be close to zero; if $F_i(x|\theta)$ (for a fixed θ) is not a good model, $D_i(\mathbf{x}, \theta)$ will be far from zero.

(2) Of course, we are interested in evaluating the discrepancy between the sample we have obtained, $\mathbf{x} = (x_1, \ldots, x_n)$, and the whole Bayesian model, M_i. The Bayesian solution is easy; first of all, we compute the posterior density of the parameter,

$$\pi_i(\theta|\mathbf{x}) = \pi_i(\theta|x_1, \ldots, x_n) = \frac{f_i(x_1, \ldots, x_n|\theta)\pi_i(\theta)}{\int_\Theta f_i(x_1, \ldots, x_n|\theta)\pi_i(\theta)d\theta}$$
$$= \frac{f_i(x_1|\theta) \cdots f_i(x_n|\theta)\pi_i(\theta)}{\int_\Theta f_i(x_1|\theta) \cdots f_i(x_n|\theta)\pi_i(\theta)d\theta},$$

and then we evaluate the posterior expected discrepancy between the sample $\mathbf{x}$ and the model M_i:

$$D_i(\mathbf{x}) = \int_\Theta D_i(\mathbf{x}, \theta)\pi_i(\theta|\mathbf{x})d\theta.$$

(3) Finally, we only have to compare $D_1(\mathbf{x}), \ldots, D_m(\mathbf{x})$, and choose the Bayesian model having the smallest posterior expected discrepancy.

3. Some discrepancy measures

The χ^2 discrepancy was used and studied in De la Horra (2008). This discrepancy measure may be reasonable but, of course, it is not the only one we may consider.

In this section, several reasonable discrepancy measures are proposed. Remember that, in all the cases, we want to measure the discrepancy between $(F_i(x_1|\theta), \ldots, F_i(x_n|\theta))$ and the $U(0,1)$ distribution.

(1) $\boldsymbol{\chi^2}$ discrepancy

The discrepancy between $(F_i(x_1|\theta), \ldots, F_i(x_n|\theta))$ and the $U(0,1)$ distribution may be measured by the χ^2 discrepancy. For doing that, we partition the interval (0, 1) in k subintervals, $(0, 1/k), (1/k, 2/k), \ldots, ((k-1)/k, 1)$ and the χ^2 discrepancy is defined as usual:

$$D_i^1(\mathbf{x}, \theta) = \sum_{j=1}^{k} \frac{[O_{ij}(\theta) - n(1/k)]^2}{n(1/k)} = \sum_{j=1}^{k} \frac{[O_{ij}(\theta) - (n/k)]^2}{n/k},$$

where $O_{ij}(\theta)$ is the number of elements of $(F_i(x_1|\theta), \ldots, F_i(x_n|\theta))$ we have obtained in each subinterval.

(2) Kolmogorov-Smirnov discrepancy

Let $G_0(y)$ denote the cumulative distribution function of the $U(0,1)$, and let $G_i(y|\theta)$ denote the empirical cumulative distribution function corresponding to the sample $(F_i(x_1|\theta), \ldots, F_i(x_n|\theta))$. The Kolmogorov-Smirnov discrepancy is defined as usual:

$$D_i^2(\mathbf{x}, \theta) = \sup_{y \in (0,1)} |G_i(y|\theta) - G_0(y)|.$$

(3) $\boldsymbol{L^1}$ discrepancy

Let $g_0(y)$ denote the density function of the $U(0,1)$, and let $g_i(y|\theta)$ denote some reasonable density estimator obtained from $(F_i(x_1|\theta), \ldots, F_i(x_n|\theta))$. The L^1 discrepancy is defined as usual:

$$D_i^3(\mathbf{x}, \theta) = \int_0^1 |g_i(y|\theta) - g_0(y)| dy.$$

In the next section, a density estimator with an Epanechnikov kernel will be used for $g_i(y|\theta)$ (see, for instance, Silverman (1986)). Of course, other density estimators may be used.

(4) Intrinsic discrepancy

Let us consider again $g_0(y)$ (defined over $\mathcal{X}_0 = (0,1)$) and $g_i(y|\theta)$ (defined over $\mathcal{X}_i \subset (0,1)$). Bernardo and Rueda (2002) defined the intrinsic discrepancy as follows (see also Bernardo (2005), Berger et al. (2009)):

$$D_i^4(\mathbf{x},\theta) = \min\left\{ \int_{\mathcal{X}_i} g_i(y|\theta) \log \frac{g_i(y|\theta)}{g_0(y)} dy \, , \, \int_{\mathcal{X}_0} g_0(y) \log \frac{g_0(y)}{g_i(y|\theta)} dy \right\}$$
$$= \int_{\mathcal{X}_i} g_i(y|\theta) \log \frac{g_i(y|\theta)}{g_0(y)} dy \, ,$$

where the last equality follows because $\mathcal{X}_i \subset \mathcal{X}_0$ and the second integral in the first line is not finite (for general properties of the intrinsic discrepancy, see Bernardo (2005)). In the next section, a density estimator with an Epanechnikov kernel will be again used for $g_i(y|\theta)$.

4. Comparing discrepancy measures

First of all, we will compare the performance of the four discrepancy measures proposed in Section 3. For doing that, we will proceed by simulation as follows:

(1) Fix m Bayesian models, $M_i = \{f_i(x|\theta), \pi_i(\theta)\}$, $i = 1, \ldots, m$.

(2) Simulate a very large number of random samples from the Bayesian model M_i. Apply the method described in Section 2 to these samples, for the four discrepancy measures proposed in Section 3, and record the percentage of correct classification with each discrepancy.

(3) Repeat Step (2) for each model M_i, $i = 1, \ldots, m$. Construct a double entry table with the percentages of correct classification with each discrepancy measure and each model.

(4) Finally, look for the discrepancy measure having the best performance.

This algorithm is next applied to two examples. These examples are simple to describe but quite interesting, as explained below.

Example 1. We consider the three following Bayesian models:

$$M_1 = \{f_1(x|\theta) \sim N(\theta, \sigma = 1); \pi_1(\theta) \propto 1\}$$
$$M_2 = \{f_2(x|\theta) \sim N(\theta, \sigma = 2); \pi_2(\theta) \propto 1\}$$
$$M_3 = \{f_3(x|\theta) \sim N(\theta, \sigma = 3); \pi_3(\theta) \propto 1\}$$

In the three models, the sampling model is the normal distribution (with different standard deviations) and the prior density is the reference prior. We are considering three similar Bayesian models because, if the method has a good performance when similar models are compared, the performance of the method will be still better when the models are quite different.

We now apply the algorithm described at the beginning of this section. For doing that, we generate, for instance, 1000 random samples with 50 elements each from one of the sampling densities in Model 1 (for instance, from the $N(0, \sigma = 1)$ distribution). The improper prior in Model 1 is used for obtaining the posterior density:

$$\pi(\theta|x_1, \ldots, x_{50}) \sim N(\bar{x}; \sigma = 1/\sqrt{50}\,) \qquad (\text{model } M_1).$$

The same procedure is then carried out for Model 2 and Model 3. For these models, posterior densities are:

$$\pi(\theta|x_1, \ldots, x_{50}) \sim N(\bar{x}; \sigma = 2/\sqrt{50}\,) \qquad (\text{model } M_2)$$
$$\pi(\theta|x_1, \ldots, x_{50}) \sim N(\bar{x}; \sigma = 3/\sqrt{50}\,) \qquad (\text{model } M_3)$$

The percentages of correct classification for each discrepancy measure and each model are shown in Table 1.

In these examples, χ^2 discrepancies are computed by partitioning the interval (0, 1) into $k = 4$ subintervals (the number of subintervals must not be too small, but each subinterval must contain a reasonable number of observations). ■

Table 1: *Percentages of correct classification in Example 1.*

	χ^2	$K-S$	L^1	Intrinsic
M_1	100%	100%	100%	100%
M_2	99%	96%	77%	76%
M_3	92%	89%	43%	45%

Example 2. We consider the three following Bayesian models:

$$M_1 = \{f_1(x|\theta) \sim N(\theta, \sigma = 1); \pi_1(\theta) \propto 1\}$$
$$M_2 = \{f_2(x|\theta) \sim N(\theta, \sigma = 5); \pi_2(\theta) \propto 1\}$$
$$M_3 = \{f_3(x|\theta) \sim N(\theta, \sigma = 10); \pi_3(\theta) \propto 1\}$$

In the three models, the sampling model is again the normal distribution (with different standard deviations) and the prior density is again the reference prior, but now the models are more different than in Example 1, because the standard deviations are more different. So, it is expected that the percentages of correct classification will be better than in Example 1 for all the discrepancy measures.

We now apply the algorithm described at the beginning of this section, in a similar way to Example 1. The percentages of correct classification with each discrepancy measure and each model are shown in Table 2. ■

Table 2: *Percentages of correct classification in Example 2.*

	χ^2	$K-S$	L^1	Intrinsic
M_1	100%	100%	100%	100%
M_2	99%	100%	100%	100%
M_3	98%	98%	79%	66%

Main conclusions

- The global performance of the χ^2 discrepancy is the best one in these examples.
- Of course, another discrepancy measure may have a better performance in other cases, but the point here is that the χ^2 discrepancy is reasonable to use. As a consequence, the χ^2 discrepancy will be used in the following sections.

5. Calibrating the discrepancy

Now, we have to choose among m different Bayesian models, M_i, $i = 1, \ldots, m$, trying to find the best model for our data. It is important to remark that the model we choose is not to be understood as the true model (nobody knows the true model), but as the best model for our data we can find among several possible models.

Therefore, we have to answer the following question: is the model we have finally chosen good enough for our data? A reasonable procedure for answering this question is given next.

Once we have chosen M_i as the best model among $M_1,\ldots,M_m$, $D_i(\mathbf{x})$ is just a number giving the posterior expected discrepancy between our data, $\mathbf{x}$, and the model, M_i. Now, it is important to calibrate this number:

- If the discrepancy is small, the model we have chosen is a good model for our data.
- If the discrepancy is large, the model we have chosen is not a good model for our data.

For deciding if the discrepancy, $D_i(\mathbf{x})$, between our data, $\mathbf{x}$, and the model, M_i, is either large or small, we may proceed as follows:

(1) Simulate a very large number of random samples from the Bayesian model, M_i, and compute the posterior expected discrepancies between each of these samples and M_i.

(2) Compare $D_i(\mathbf{x})$ to the posterior expected discrepancies we have computed in Step (1), for obtaining in what percentile $D_i(\mathbf{x})$ is placed.

This procedure is applied next to some examples with simulated data. The aim of these examples is to evaluate the behaviour of the procedure in these control cases.

Example 3. A random sample is simulated from a $N(0,\sigma=2)$ distribution.

Consider, as possible models, the three following Bayesian models:

$$M_1 = \{f_1(x|\theta) \sim N(\theta,\sigma=1); \pi_1(\theta) \propto 1\}$$
$$M_2 = \{f_2(x|\theta) \sim N(\theta,\sigma=2); \pi_2(\theta) \propto 1\}$$
$$M_3 = \{f_3(x|\theta) \sim N(\theta,\sigma=3); \pi_3(\theta) \propto 1\}$$

We apply the algorithm described in Section 2. The model M_2 is chosen, because the smallest posterior expected discrepancy, $D_2(x) = 8.05$, is obtained from M_2. Now, we have to calibrate this value, so we simulate 1000 random samples from model M_2. It is found that the discrepancy $D_2(x) = 8.05$ is between percentiles 12 and 13. Therefore, in this case, model M_2 is a very good model for our data. This is a very reasonable result because all the data did come from M_2. ■

Example 4. A random sample is simulated in which 5% of the elements come from a $N(0,\sigma=1)$ distribution, 90% from a $N(0,\sigma=2)$ distribution, and 5% from a $N(0,\sigma=3)$ distribution.

Consider again, as possible models, the three Bayesian models given in Example 3, and apply the algorithm described in Section 2. The model M_2 is again chosen, because

the smallest posterior expected discrepancy, $D_2(x) = 12.16$, is obtained from M_2. To calibrate this value we simulate 1000 random samples from model M_2, and it is found that the discrepancy $D_2(x) = 12.16$ is between percentiles 56 and 57. Therefore, in this case, model M_2 is still a good model for our data, although the discrepancy is larger than in Example 3. This is again a very reasonable result because, in this case, almost all the data came from M_2. ■

Example 5. Finally, we simulate a random sample in which 33% of the elements come from a $N(0, \sigma = 1)$ distribution, 34% from a $N(0, \sigma = 2)$ distribution, and 33% from a $N(0, \sigma = 3)$ distribution.

Again the three Bayesian models given in Example 3 are considered and the algorithm described in Section 2 is applied. The model M_2 is chosen again, because the smallest posterior expected discrepancy, $D_2(x) = 17.58$, is obtained from M_2.

Calibrating as before, we simulate 1000 random samples from model M_2, and it is found that the discrepancy $D_2(x) = 17.58$ is between percentiles 92 and 93. Therefore, in this case, model M_2 is not a good model for our data and, once more, this is a very reasonable result. ■

Main conclusions

- This method for calibrating the discrepancy shows a good behaviour in these controlled situations.
- As a consequence, the method can be applied to a set of real data with reasonable confidence. This is carried out in the next section.

6. Application to real data

A set of 30 failure times for air conditioners on an airplane was introduced by Proschan (1963). This set of real data was analysed first by Berger and Pericchi (1996) and then by Gutiérrez-Peña and Walker (2001). They consider three Bayesian models for explaining this set of real data, with exponential, lognormal and Weibull densities as sampling densities:

$$M_1 = \left\{ f_1(x|\theta) = \frac{1}{\theta} \exp\left\{ -\frac{x}{\theta} \right\}; \pi_1(\theta) \propto \frac{1}{\theta} \right\}$$

$$M_2 = \left\{ f_2(x|\mu, \sigma^2) = \frac{\exp\{-(\log x - \mu)^2/(2\sigma^2)\}}{\sqrt{2\pi} x \sigma}; \pi_2(\mu, \sigma^2) \propto \frac{1}{\sigma^2} \right\}$$

$$M_3 = \left\{ f_3(x|\alpha, \beta) = \beta x^{(\beta-1)} \alpha^{-\beta} \exp\left\{ -(x/\alpha)^\beta \right\}; \pi_3(\alpha, \beta) \propto \frac{1}{\alpha\beta} \right\}$$

The corresponding posterior distributions are:

$$\pi_1(\theta|x) \sim (2n\bar{x})\chi^{-2}_{2n}$$

$$\pi_2(\mu,\sigma^2|x) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+1} \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}\log^2 x_i - n(\overline{\log x})^2\right)\right\} \exp\left\{-\frac{n}{2\sigma^2}(\mu-\overline{\log x})^2\right\}$$

$$\sim NIG\left(n-1, \frac{1}{n-1}\left(\sum_{i=1}^{n}\log^2 x_i - n(\overline{\log x})^2\right), n, \overline{\log x}\right)$$

$$\pi_3(\alpha,\beta|x) \propto \beta^{n-1}\left[\prod_{i=1}^{n} x_i\right]^{\beta-1} \alpha^{-n\beta-1}\exp\left\{-\frac{1}{\alpha^\beta}\sum_{i=1}^{n}x_i^\beta\right\} \quad ,$$

where χ^{-2}_{2n} denotes the "inverse chi-square distribution", $\overline{\log x} = \frac{1}{n}\sum_{i=1}^{n}\log x_i$ and NIG denotes the normal inverse gamma distribution.

Next, we show and comment the results, when different methods are applied:

(1) Method by Berger and Pericchi. They obtained that the model M_1 is preferred to M_2, and the model M_2 is preferred to M_3.

(2) Method by Gutiérrez-Peña and Walker. They obtained that the models M_1 and M_3 are preferred to M_2.

First of all, we remark that the results are different for the two methods shown above.

(3) Method in this article. We apply our method to these data (notice that the third posterior distribution is not in closed form, and so, a Markov chain Monte Carlo (MCMC) method is needed for simulations). The following discrepancies are obtained:

$$D_1(\mathbf{x}) = 6.5$$
$$D_2(\mathbf{x}) = 2.8$$
$$D_3(\mathbf{x}) = 11.8$$

Therefore, with this method, the model M_2 is preferred to M_1 and M_3. This result is also different from the results obtained with the other two methods. The calibration of the discrepancies, by using the algorithm proposed in Section 5, throws light upon these results:

- It is obtained (by simulation) that the discrepancy between the real data and the model M_1, $D_1(\mathbf{x}) = 6.5$, is between percentiles 94 and 95. Therefore M_1 is a bad (although not very bad) model for these real data.

- It is obtained (by simulation) that the discrepancy between the real data and the model M_2, $D_2(\mathbf{x}) = 2.8$, is between percentiles 74 and 75. Therefore M_2 is a reasonable (although not especially good) model for these real data.
- It is obtained (by simulation) that the discrepancy between the real data and the model M_3, $D_3(\mathbf{x}) = 11.8$, is between percentiles 90 and 91. Therefore M_3 is a bad (although not very bad) model for these real data.

Main conclusions

- The best model according to our method is M_2 (the lognormal model) and is a reasonable model for these real data, because the calibration shows that this discrepancy is between percentiles 74 and 75.
- The discrepancies for the models M_1 and M_3 are larger than the discrepancy for the model M_2. It is important to notice that their calibrations are bad, but not very bad. Possibly, this is the reason why they were chosen when other methods are used.

Acknowledgements

This research has been partially supported by grant MTM2010-17366 (the first author) and under grant MTM2008-03282 (the second author) from the Spanish ministry of Science and Innovation. We thank the referees for useful suggestions that improved the article.

References

Aitkin, M. (1991). Posterior Bayes factors (with discussion). *Journal of the Royal Statistical Society B*, 53, 11–142.

Berger, J. O., Bernardo, J. M. and Sun, D. (2009). The formal definition of reference priors. *Annals of Statistics*, 37, 905–938.

Berger, J. O. and Pericchi, L. R. (1996). The intrinsic Bayes factors for model selection and prediction. *Journal of the American Statistical Association*, 91, 109–122.

Bernardo, J. M. (2005). Intrinsic credible regions: an objective Bayesian approach to interval estimation (with discussion). *Test*, 14, 317–384.

Bernardo, J. M. and Rueda, R. (2002). Bayesian hypothesis testing: a reference approach. *International Statistical Review*, 70, 351–372.

Carota, C. Parmigiani, G. and Polson, N. G. (1996). Diagnostic measures of model criticism. *Journal of the American Statistical Association*, 91, 753–762.

De la Horra, J. (2008). Bayesian model selection: Measuring the χ^2 discrepancy with the uniform distribution. *Communications in Statistics-Theory and Methods*, 37, 1412–1424.

De la Horra, J. and Rodríguez-Bernal, M. T. (2005). Bayesian model selection: a predictive approach with losses based on distances L^1 and L^2. *Statistics & Probability Letters*, 71, 257–265.

De la Horra, J. and Rodríguez-Bernal, M. T. (2006). Prior density selection as a particular case of Bayesian model selection: a predictive approach. *Communications in Statistics-Theory and Methods*, 35, 1387–1396.

Geisser, S. and Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, 74, 153–160.

Gelfand, A. E. (1995). Model determination using sampling-based methods. In: Gilks, W., Richardson, S., Spiegelhalter, D. (Eds.), *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London, 145–161.

Gelfand, A. E., Dey, D. K. and Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. In: Bernardo, J. M., Berger, J. O., Dawid, A. P., Smith, A. F. M. (Eds.), *Bayesian Statistics 4*. Oxford University Press, Oxford, 147–167.

Gelfand, A. E. and Ghosh, S. (1998). Model choice: a minimum posterior predictive loss approach. *Biometrika*, 85, 1–11.

Gutiérrez-Peña, E. and Walker, S. G. (2001). A Bayesian predictive approach to model selection. *Journal of Statistical Planning and Inference*, 93, 259–276.

Laud, P. W. and Ibrahim, J. G. (1995). Predictive model selection. *Journal of the Royal Statistical Society B*, 57, 247–262.

McCulloch, R. E. (1989). Local model influence. *Journal of the American Statistical Association*, 84, 473–478.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison (with discussion). *Journal of the Royal Statistical Society B*, 57, 99–138.

Proschan, F. (1963). Theoretical explanation of observed decreasing failure rate. *Technometrics*, 5, 375–383.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

San Martini, A. and Spezzaferri, F. (1984). A predictive model selection criterion. *Journal of the Royal Statistical Society B*, 46, 296–303.

Soofi, E. S., Ebrahimi, N. and Habibullah, M. (1995). Information distinguishability with application to the analysis of failure data. *Journal of the American Statistical Association*, 90, 657–668.

Spiegelhalter, D. J. and Smith, A. F. M. (1982). Bayes factors for linear and log-linear models with vague prior information. *Journal of the Royal Statistical Society B*, 44, 377–387.

Trottini, M. and Spezzaferri, F. (2002). A generalized predictive criterion for model selection. *Canadian Journal of Statistics*, 30, 79–96.