Some Bayesian predictive approaches to model selection

Nitai Mukhopadhyay[a,*], Jayanta K. Ghosh[b,c], James O. Berger[d]

[a]*Eli Lilly and Co., Indianapolis, IN 46285, USA*
[b]*Department of Statistics, Purdue University, West Lafayette, IN 47907, USA*
[c]*Indian Statistical Institute, 203 BT Road, Calcutta 700 035, India*
[d]*Institute of Statistics and Decision Sciences, Duke University, Durham, NC 27708, USA*

Abstract

A variety of pseudo-Bayes factors have been proposed, based on using part of the data to update an improper prior, and using the remainder of the data to compute the Bayes factor. A number of these approaches are of a bootstrap or cross-validation nature, with some type of average being taken over the data used for updating. Asymptotic characteristics of a number of these pseudo-Bayes factors are discussed, and it is shown how many behave quite differently from ordinary Bayes factors. It is also shown that arguments of predictive optimality, based on simply inserting the empirical distribution in place of the 'true predictive distribution', can be misleading; the particular example of this that is studied is the argument given in Bernardo and Smith [1994. Bayesian Theory. Wiley, Chichester] to the effect that the geometric intrinsic Bayes factor has an optimal predictive property.

Keywords: Bayes factor; Training sample

1. Introduction

Bayes rules for model selection in a prediction problem with squared error loss and their asymptotic properties are studied in Barbieri and Berger (2004), Mukhopadhyay (2000) and

Berger et al. (2003). Prediction may be interpreted in the same sense as Akaike, that is, one tries to predict a future replication of the given data. Possible other approaches to prediction may involve use of other loss functions, for example the Kullback–Leibler logarithmic loss function. Also, one may not want to predict a full replicate of the entire given data. A general formulation under squared error is given in Barbieri and Berger (2004). If one has n i.i.d. observations under models of fixed dimensions, and wants to predict the next observation with a logarithmic loss, then Rissanen (1986) shows that the optimal Bayesian model is asymptotically (as $n \to \infty$) the same as that obtained by maximizing BIC.

An interesting but somewhat different approach, based on cross validation, is due to Aitkin (1991), Geisser (1975) and Gelfand and Dey (1994). We are given a data set, $X = (x_1, \ldots, x_n)$. The data are subdivided into K sets S_j, $j = 1, \ldots, K$. Let X_{S_j} be the x_i's in S_j and $\widetilde{X}_{-S_j}$ be the x_i's not in S_j. If S is a singleton $\{x_j\}$, one writes x_j and X_{-j} for X_S and X_{-S}. The rule is to choose the model that maximizes the predictive probability

$$\prod_{j=1}^{K} q_M(X_{S_j}|X_{-S_j}), \tag{1}$$

where q_M is the posterior predictive density for model M defined as

$$q_M(X_{S_j}|X_{-S_j}) = \int_{\Theta_M} f_M(X_{S_j}|\underset{\sim}{\theta}_M)\pi_M(\underset{\sim}{\theta}_M|X_{-S_j})\,\mathrm{d}\underset{\sim}{\theta}_M,$$

where f_M is the likelihood and π_M is the prior under model M.

In this paper, we study some aspects of model selection based on predictive probability. In the first part, namely Sections 2 and 3, we rigorously explore a suggestion of Bernardo and Smith (1994) (also appearing in the discussion of O'Hagan (1995)) concerning replacement of the 'true predictive' by an empirical estimate; we show that the error arising from this replacement can be too large for the asymptotic approximation of Bernardo and Smith (1994) to be valid. In the second part of the paper, namely Sections 4 and 5, asymptotic characteristics of a number of cross-validatory Bayes factors are discussed, and it is shown how many behave quite differently from ordinary Bayes factors; in particular, they can even be inconsistent under the simple null model, M_0.

Our basic paradigm consists of two models, M_1 and M_2, for predicting the future observation y, with predictive densities $q_i(y|X) = q(y|X, M_i), i = 1, 2$. The true model q_A assumes an exchangeable distribution for $(X_1, \widetilde{X_2}, \ldots, X_n, \widetilde{Y})$. A Bayes factor proposed in Berger and Pericchi (1995), the geometric intrinsic Bayes factor (GIBF), is given as

$$\mathrm{GIBF}_{21} = \left[\prod_{i=1}^{n} \frac{q_2(X_{-i}|x_i)}{q_1(X_{-i}|x_i)}\right]^{1/n}.$$

This was originally introduced as a device for developing Bayes factors with improper priors; such priors cannot be used directly because they involve an arbitrary multiplicative constant. Later, in Berger and Pericchi (1997), the GIBF is shown to correspond to an actual Bayes factor with respect to what is called a proper 'intrinsic prior', under some conditions. In the following section we discuss the predictive interpretation of the GIBF given by Bernardo and Smith (1994).

The range of Bayes factors studied in Berger and Pericchi (1996) and Gelfand and Dey (1994) cover the two extremes (minimal and maximal) of the size of the training sample. In view of the predictive approach of these Bayes factors, as explained in the preceding paragraph, they are referred to as predictive Bayes factors. In Section 4 we give a comparative study of these Bayes factors.

2. The predictive approach of Bernardo and Smith

We start with the motivation of a criterion due to Gelfand and Dey (1994) which takes the predictive density, as in Eq. (1), by conditioning on a training sample of size $n-1$. The logarithmic utility function arises from comparing the Kullback–Leibler divergence of the two models from the true model q_A as follows. M_2 is chosen if

$$\int \log\left(\frac{q_A(y|\underset{\sim}{X})}{q_1(y|\underset{\sim}{X})}\right) q_A(y|\underset{\sim}{X})\,dy - \int \log\left(\frac{q_A(y|\underset{\sim}{X})}{q_2(y|\underset{\sim}{X})}\right) q_A(y|\underset{\sim}{X})\,dy > 0$$

$$\text{i.e.,}\quad \int \log\left(\frac{q_2(y|\underset{\sim}{X})}{q_1(y|\underset{\sim}{X})}\right) q_A(y|\underset{\sim}{X})\,dy > 0. \tag{2}$$

But evaluation of the above is not possible, since q_A is not specified. In the case of exchangeable x_i's, a Monte Carlo approximation of the LHS of Eq. (2), suggested in Bernardo and Smith (1994), is

$$\frac{1}{n}\sum_{j=1}^{n} \log\left[\frac{q_2(x_j|X_{-j})}{q_1(x_j|X_{-j})}\right] \equiv \frac{1}{n}\sum_{j=1}^{n} \log \mathrm{BF}_{21}[x_j, X_{-j}]. \tag{3}$$

Thus the selection criterion chooses M_2 if

$$\prod_{j=1}^{n} \mathrm{BF}_{21}[x_j, X_{-j}] > 1, \tag{4}$$

which is one of the criterion proposed by Gelfand and Dey (1994). In the sequel, we denote the LHS of (4) as $\mathrm{BF}_{21}^{\mathrm{GD}}$.

The transition from (2) to (3) is not easily justified for two reasons. In (2), X has dimension n, while in (3) the prediction is based on $(n-1)$ dimensional x's. Secondly, in Eq. (2), $\underset{\sim}{X}$ is held fixed, whereas in (3) one seems to use the empirical distribution of the x's.

For these reasons, it seems more reasonable to replace X by X_S where X_S denotes a subset of size s of the whole data set X and to replace (2) by an expectation with respect to the empirical distribution function of X_S. Then the criterion would become: choose M_2 if

$$\frac{1}{\binom{n}{s}}\sum_{X_S} \int \log\left(\frac{q_2(X_{-S}|X_S)}{q_1(X_{-S}|X_S)}\right) q_A(X_{-S}|X_S)\,dX_{-S} > 0.$$

A Monte Carlo approximation to this expression yields the GIBF with training sample of size s. The criterion becomes: choose M_2 if

$$\prod_{s \text{ subsets}} \mathrm{BF}_{21}(X_{-S}, X_S) > 1.$$

If one starts with a training sample of size $s \ll n$, it is not reasonable to talk of predicting a large data set of dimension $(n-s)$ based on a small segment of length s. (For s close to n, the prediction motivation is more reasonable.) Thus, for $s \ll n$, Bernardo and Smith (1994) refer to the criterion as a measure of "fidelity to the data".

In the next section, we examine the accuracy of the above approximation and the extent to which it clarifies the derivation or significance of the GIBF.

3. Calculations with a general likelihood

Consider the GIBF with a training sample of size $s = 1$. We are concerned with the accuracy with which (3) approximates (2) for, say, M_1. As explained in the previous section, we begin by replacing (2) with a further expectation with respect to the empirical distribution function of x. Denote $y_{(k)} = (y_1, \ldots, y_k)$. Then, following the suggestion of Bernardo and Smith (1994),

$$\int \log\, q_1(y_{(n-1)}|x) q_A(y_{(n-1)}|x)\, \mathrm{d}y_{(n-1)} \approx \frac{1}{n}\sum_{i=1}^{n} \log\, q_1(x_{-i}|x_i), \tag{5}$$

where $\approx$ simply means that the approximation is good in some sense. Here $y_{(n-1)}$ denotes a future observation of size $n-1$, independent of the data, and x is the training sample, assumed to yield a proper posterior; specifically, suppose $q_1 = \int f_\theta(x)\pi(\mathrm{d}\theta)$, where $\pi(\theta)$ is the prior and $\int f_\theta(x)\pi(\mathrm{d}\theta) < \infty$. Making a Laplace approximation to the RHS of Eq. (5) (which is easy to justify rigorously) yields

$$\begin{aligned}
\mathrm{RHS} &= \frac{1}{n}\sum_{i=1}^{n} \log\left\{\frac{\int \prod_{j=1}^{n} f_\theta(x_j)\pi(\mathrm{d}\theta)}{\int f_\theta(x_i)\pi(\mathrm{d}\theta)}\right\} \\
&= \frac{1}{2}\log(2\pi) + \sum_{j=1}^{n} \log f_{\hat\theta}(x_j) - \frac{1}{2}\log\left(-\sum_{j=1}^{n} \frac{\partial^2 \log f_{\hat\theta}(x_j)}{\partial\theta^2}\right) \\
&\quad - \frac{1}{n}\sum_{i=1}^{n} \log \int f_\theta(x_i)\pi(\mathrm{d}\theta) + \log\, \pi(\hat\theta) + o_p(1),
\end{aligned} \tag{6}$$

where $\hat\theta$ is the MLE and θ_0 is the true value.

Assume that $q_A = f_{\theta_0}$. Then, the LHS of Eq. (5) is $E_{\theta_0}(\log\, q_1(y_{(n-1)}|x))$. A similar Laplace approximation of $\log\, q_1(y_{(n-1)}|x)$ yields

$$\begin{aligned}
\log\, q_1(y_{(n-1)}|x) &= \frac{1}{2}\log(2\pi) + \sum_{j=1}^{n-1} \log f_{\hat\theta}(y_j) + \log f_{\hat\theta}(x) - \log \int f_\theta(x)\pi(\mathrm{d}\theta) \\
&\quad - \frac{1}{2}\log\left(-\sum_{j=1}^{n-1} \frac{\partial^2 \log f_{\hat\theta}(y_j)}{\partial\theta^2} - \frac{\partial^2 \log f_{\hat\theta}(x)}{\partial\theta^2}\right) + \log\, \pi(\hat\theta) + o_p(1).
\end{aligned} \tag{7}$$

The derivations of (6) and (7) are done by simple Laplace integration under the assumption of boundedness of the third derivative of the log likelihood and boundedness of π'.

The LHS of Eq. (5) is the expectation of Eq. (7) for fixed x. We now integrate x w.r.t. f_{θ_0}. From the structure of (6) and (7), the LHS and RHS of (5) have the same expectation, with a difference of order o(1) if the term on LHS of (5) is integrated out as mentioned above. But, although their expectations agree up to o(1), the LHS and RHS of Eq. (5) do not agree up to $o_p(1)$. It is easy to see that the difference between RHS and LHS of (5) can be written as a sum of n (or $n-1$) i.i.d. random variables and their mean. This difference is $O_p(\sqrt{n})$ by the central limit theorem. So the difference between the target quantity, namely LHS of (5) and the approximation, namely RHS of (5), is far from being negligible.

The log(GIBF) may be thought of as a refinement to BIC, as log(GIBF) differs from BIC by $O_p(1)$. So, any attempt to prove a stronger desirable property possessed by log(GIBF) but not BIC, must display log(GIBF) as an approximation to some desirable target up to $o_p(1)$. The target (2) fails this test. It is interesting that the expectation of log(GIBF) matches the expectation of the log of the target quantity (LHS of 5) up to O(1), but this does not result in an implementable procedure in general.

If one adds a null model and calculates the Bayes factor, the conclusion does not change. The next section demonstrates an example showing this.

3.1. Example

A simple example will demonstrate that the phenomenon also holds for $s = n-1$. Suppose $x_1, \ldots, x_n \overset{\text{iid}}{\sim} N(\theta, 1)$ and that we are comparing the models $M_1: \theta = 0$ and $M_2: \theta \in \Re$, $\pi_2(\theta) = 1$ for all θ. Easy algebra shows that

$$\sum \log m_1(x_i|x_{-i}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum x_i^2,$$

$$\sum \log m_2(x_i|x_{-i}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2(n-1)}\sum (x_i - \bar{x})^2 - \frac{n}{2}\log\left(1 + \frac{1}{n-1}\right).$$

So the Bayes factor of Gelfand and Dey (1994) is given by

$$\log \mathrm{BF}_{21}^{\mathrm{GD}} = \frac{n}{2}\bar{x}^2 - \frac{1}{2(n-1)}\sum (x_i - \bar{x})^2 - \frac{n}{2}\log\left(1 + \frac{1}{n-1}\right). \tag{8}$$

Suppose $y = (y_1, \ldots, y_n)$ are independent future observations, $\underset{\sim}{x} = (x_1, X_2, \ldots, x_{n-1})$ are all independent with distribution given by the above models. Suppose q_A is the true unknown density $N(\theta, 1)$, θ unknown. Then, using $\pi_2(\theta|\underset{\sim}{x}) = N(\bar{x}, 1/(n-1))$ as prior, we have as target,

$$E_{q_A}\left(\log \frac{q_2(y \mid \underset{\sim}{x})}{q_1(y \mid \underset{\sim}{x})}\right) = nE_{q_A}\left(\log \frac{\sqrt{n-1}\,e^{-(n-1)(y_1-\bar{x})^2/2n}}{\sqrt{n}\,e^{-y_1^2/2}}\right)$$

$$= nE_{q_A}\left(\frac{1}{2}\left\{y_1^2 - \frac{n-1}{n}(y_1 - \bar{x})^2\right\} - \frac{1}{2}\log\left(\frac{n}{n-1}\right)\right). \tag{9}$$

Thus the difference between the approximation (8) and the target quantity (9) is given by

$$\begin{aligned}\Delta &= \frac{1}{2}\left\{\frac{1}{(n-1)}\sum(x_i-\bar{x})^2 - n\bar{x}^2 + nE_{q_A}\left\{y_1^2 - \frac{n-1}{n}(y_1-\bar{x})^2\right\}\right\}\\ &= \frac{1}{2}\left\{\frac{1}{(n-1)}\sum(x_i-\bar{x})^2 - n\bar{x}^2 + nE_{q_A}(y_1^2) - (n-1)E_{q_A}(y_1-\bar{x})^2\right\}\\ &= \frac{1}{2}\left\{\left(1+\mathrm{O_p}\left(\frac{1}{\sqrt{n}}\right)\right) - n\left(\theta+\frac{z}{\sqrt{n-1}}\right)^2 + n(\theta^2+1) - (n-1)\frac{n}{n-1}\right\}\\ &= \frac{1}{2}\left\{1 - \frac{nz^2}{n-1} - \frac{2n\theta z}{\sqrt{n-1}}\right\} + \mathrm{O_p}\left(\frac{1}{\sqrt{n}}\right)\\ &= \mathrm{O_p}(\sqrt{n}).\end{aligned}$$

Hence the apparent motivation or justification given by Bernardo and Smith (1994) does not hold. The above argument will go through for any fixed s but, for large s close to n, replacing (2) by an expectation with respect to the empirical distribution of X_S cannot be justified.

4. Comparative study of different Bayes factors

How do these heuristic BF's compare with each other? There seem to be at least two different ways of doing this. The first method, due to Berger and Pericchi (1996), is to compare the intrinsic priors, π, corresponding to each BF under consideration; an intrinsic prior is a prior distribution such that the Bayes factor with respect to that prior differs from the ad hoc Bayes factor by $\mathrm{o_p}(1)$. The second method is to study consistency of the BF's by exhibiting their logarithms as penalized likelihoods and then compare the penalties as well as their impact on consistency. We will follow the second route here, for a special case in which exact algebraic expressions can be derived.

Suppose $x_i \sim N(\theta,1)$ and $\theta \sim \pi(\theta)=1$, $i=1,\ldots,n$. In this section we denote it as model M. Denoting the subset of size s, used to update the prior, by X_S and the remaining data set by X_{-S}, we have

$$m(X_{-S}|X_S) = \frac{\sqrt{s}}{(2\pi)^{(n-s)/2}\sqrt{n}}\exp\left\{-\frac{1}{2}\sum_{x_i\in X_{-S}}(x_i-\bar{x}_{-S})^2 - \frac{1}{2}\frac{s(n-s)}{n}(\bar{x}_{-S}-\bar{x}_S)^2\right\}. \tag{10}$$

Thus the geometric average of $m(X_{-S}|X_S)$, denoted by $m^G(\underset{\sim}{X})$, is given by

$$\begin{aligned}\log m^G(\underset{\sim}{X}) &= \frac{1}{\binom{n}{s}}\sum_{X_S}\log m(X_{-S}|X_S)\\ &= \left(-\frac{1}{2}\left(\frac{n-s}{n-1}\right)\sum_{i=1}^{n}(x_i-\bar{x})^2 - \frac{n-s}{2}\log(2\pi)\right) - \frac{1}{2}\log\frac{n}{s},\end{aligned} \tag{11}$$

where the first term is the maximized likelihood and the second term, namely $\log(n/s)$, can be interpreted as a penalty. The size of the updating sample, X_S, can be varied from 1 to $n-1$. The divisor $\binom{n}{s}$ is not used by Gelfand and Dey (1994) in their definition of the predictive Bayes factor.

One may think of the predictive Bayes factor as a proxy to an unknown product likelihood; the lack of the divisor keeps the orders of $\mathrm{BF}^{\mathrm{GD}}$ and Bayes factors based on a subjective or noninformative prior comparable.

It is interesting to compare the penalties arising from $s = 1$ and $s = n - 1$. For $s = 1$, i.e., in the case of the GIBF,

$$m(x_{-i}|x_i) = \frac{1}{(2\pi)^{(n-1)/2}\sqrt{n}} \exp\left\{ -\frac{1}{2}\sum_{j\neq i}(x_j - \bar{x}_{-i})^2 - \frac{n-1}{2n}(\bar{x}_{-i} - x_i)^2 \right\}$$

and the geometric average of the above is given by

$$\log m^G(\underset{\sim}{x}) = -\frac{1}{2}\left\{ \sum (x_i - \bar{x})^2 + (n-1)\log(2\pi) + \log n \right\}.$$

The corresponding Bayes factor of M, against the simpler model with $\theta = 0$, is given by

$$\begin{aligned} \log \mathrm{BF}^{\mathrm{G}}_{s=1} &= -\frac{1}{2}\sum (x_i - \bar{x})^2 + \log n + \frac{n-1}{2n}\sum x_i^2 \\ &= \frac{1}{2} n\bar{x}^2 - \frac{1}{2n}\sum x_i^2 + \log n. \end{aligned}$$

The other extreme, used by Gelfand and Dey (1994), is to update the prior with $s = n - 1$ observations. The marginal density of $\underset{\sim}{X}$, under model M, denoted m^{GD}, simplifies to

$$\begin{aligned} \log m^{\mathrm{GD}}(\underset{\sim}{x}) &= \sum \log m^{\mathrm{GD}}(x_i|x_{-i}) \\ &= -\frac{1}{2}\left\{ \frac{n}{n-1}\sum (x_i - \bar{x})^2 + n\log(2\pi) + n\log\left(\frac{n}{n-1}\right) \right\}, \end{aligned} \tag{12}$$

leading to the Bayes factor for M against the simpler model with $\theta = 0$,

$$\log \mathrm{BF}^{\mathrm{GD}}_{s=n-1} = \frac{n^2}{2(n-1)}\bar{x}^2 - \frac{1}{2(n-1)}\sum x_i^2 + n\log\left(\frac{n}{n-1}\right).$$

Thus the Bayes factor with m^{GD} acts like a penalized likelihood criterion with a penalty ≈ 1 for large n, whereas the marginal computed by the GIBF method gives a $\log n$ penalty. The constant penalty of m^{GD} will give inconsistency under the null. This can be shown as follows. In the above normal set up, the difference of the maximized loglikelihoods of the two models is given by $\Delta\mathscr{L} = \frac{1}{2}n\bar{x}^2$. For any constant penalty c, the probability of rejecting the null under the null is $P_{\mathrm{H}_0}(n\bar{x}^2 > 2c)$, which is positive since $\sqrt{n}\bar{x}$ has a fixed normal distribution under the null model.

Both methods are consistent under the simpler model M_1. For GIBF it follows from the consistency of BIC and the fact that GIBF differ from BIC by $\mathrm{O_p}(1)$. $\mathrm{BF}^{\mathrm{GD}}$ has lower penalty than GIBF and hence selects M_1 more often. Hence consistency of GIBF under M_1 implies that of the $\mathrm{BF}^{\mathrm{GD}}$.

In general, the constant penalty arising in the case of m^{GD} can be chosen to be any other constant simply by making the size of the updating sample proportional to n. From Eq. (10), an updating sample of size $s = [n\alpha]$ for $\alpha \in (0, 1)$ induces a penalty of magnitude $\log(1/\alpha)$.

Remark 4.1. The phenomenon of constant penalty also occurs in the posterior predictive density of Aitkin, 1991, giving a $\frac{1}{2}\log 2$ penalty.

We can get some further insight into inconsistency of some of the predictive Bayes factors as follows.

The predictive Bayes factor uses a fraction of the data to update the prior and uses the remaining part to compute the Bayes factor. So an adjustment is needed to bring them to the same scale as a BF with a prior π that uses the whole data to compute the Bayes factor. To see what adjustment is needed, consider the logarithms of the marginal of the simpler model M_1. A simple termwise comparison shows, for the marginal under M_1 with updating sample size s, that $\log m_1^G(\underset{\sim}{X}) = [(n-s)/n]\log(m_{1\pi})$, where $m_{1\pi}$ is the marginal of the simpler model under some prior π. This suggests that one should adjust the predictive Bayes factor by multiplying by $n/(n-s)$. That this adjustment makes sense, is clear from the following two special cases leading to GIBF and the $\mathrm{BF}^{\mathrm{GD}}$:

$$\frac{n}{n-1}\mathrm{BF}^{\mathrm{G}}_{s=1} \sim \mathrm{GIBF}'$$

$$n\,\mathrm{BF}^{\mathrm{G}}_{s=n-1} = \mathrm{BF}^{\mathrm{GD}}.$$

With this adjustment, one obtains the log of the predictive Bayes factor, $\mathrm{BF}^{\mathrm{GD}}$, as equal to $\frac{1}{2}n\bar{x}^2$, up to $\mathrm{O_p}(1)$. It is easy to show, using the last part of Section 5, that this can be approximated, up to $\mathrm{O_p}(1)$, by a Bayes factor with respect to a uniform prior supported on $\hat{\theta} \pm (c/\sqrt{n})$. As indicated earlier, this leads to inconsistency under the simpler model. When s is a constant, free of n, as for the GIBF, then this can similarly be viewed as an approximation to the Bayes factor when π is an uniform prior on $\hat{\theta} \pm c$. This prior is much less peaked than the prior associated with $s = [n\alpha]$ and, hence, more acceptable as a default prior. This also leads to consistency. To this extent, the GIBF seems more acceptable intuitively as a default Bayes method than the Bayes factors proposed in Gelfand and Dey (1994).

5. Calculations

Proof of Eq. (11). From Eq. (10),

$$\log m(X_{-S}|X_S) = -\frac{1}{2}\left\{\sum_{X_{-S}}(x_i - \bar{x}_{-S})^2 + \log\binom{n}{s} + (n-s)\log(2\pi)\right\} + \frac{s(n-s)}{n}(\bar{x}_{-S} - \bar{x}_S)^2.$$

Let

$$Q(\underset{\sim}{x}) = \frac{1}{\binom{n}{s}}\sum_{X_S}\left\{\sum_{X_{-S}}(x_i - \bar{x}_{-S})^2 + \frac{s(n-s)}{n}(\bar{x}_{-S} - \bar{x}_S)^2\right\}.$$

Then $Q(\underset{\sim}{x})$ is a quadratic form in $(x_1, \ldots, x_n)$ that is symmetric with respect to permutations of $\underset{\sim}{x}$. This implies that

$$Q(\underset{\sim}{x}) = a\sum_{i=1}^{n} x_i^2 + b\sum_{i\neq j} x_i x_j.$$

Also, $Q(\underset{\sim}{x}) = 0$ for $x_1 = x_2 = \cdots = x_n = d$(say) implies

$$nad^2 + bn(n-1)d^2 = 0 \Rightarrow b = -\frac{a}{n-1}.$$

Therefore,

$$Q(\underset{\sim}{x}) = a\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n-1}\sum_{i\neq j} x_i x_j\right) = a\frac{n}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{13}$$

Now,

$$\begin{aligned}
Q(\underset{\sim}{X}) &= \frac{1}{\binom{n}{s}}\sum_{X_S}\left\{\sum_{X_{-S}} x_i^2 - (n-s)\bar{x}_{-S}^2 + \frac{s(n-s)}{n}\bar{x}_{-S}^2 + \frac{s(n-s)}{n}\bar{x}_S^2 - \frac{2s(n-s)}{n}\bar{x}_S\bar{x}_{-S}\right\}\\
&= \frac{1}{\binom{n}{s}}\sum_{X_S}\left\{\sum_{X_{-S}} x_i^2 - \frac{(n-s)^2}{n}\bar{x}_{-S}^2 + \frac{s(n-s)}{n}\bar{x}_S^2 - \frac{2s(n-s)}{n}\bar{x}_S\bar{x}_{-S}\right\}\\
&= \frac{1}{\binom{n}{s}}\sum_{X_S}\left\{\sum_{X_{-S}} x_i^2 - \frac{1}{n}\left(\sum_{X_{-S}} x_i\right)^2 + \frac{n-s}{ns}\left(\sum_{X_S} x_i\right)^2 - \frac{2s(n-s)}{n}\bar{x}_S\bar{x}_{-S}\right\}\\
&= \frac{1}{\binom{n}{s}}\sum_{X_S}\left\{\left(1-\frac{1}{n}\right)\sum_{X_{-S}} x_i^2 + \frac{n-s}{ns}\sum_{X_S} x_i^2 + \text{cross products}\right\}\\
&= \frac{1}{\binom{n}{s}}\sum_{i=1}^{n} x_i^2\left(\frac{n-1}{n}\binom{n-1}{s} + \frac{n-s}{ns}\binom{n-1}{s-1}\right) + \text{cross products}\\
&= \frac{1}{\binom{n}{s}}\sum_{i=1}^{n} x_i^2\binom{n-1}{s} + \text{cross products}\\
&= \frac{\binom{n-1}{s}}{\binom{n}{s}}\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n-1}\sum_{i\neq j} x_i x_j\right) \quad \text{[by (13)]}\\
&= \frac{n-s}{n}\frac{n}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\\
&= \frac{n-s}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.
\end{aligned}$$

This implies Eq. (11). □

5.1. *Explanation of criteria with constant penalty as a Bayes factor*

Suppose

$$x_i \sim f_\theta(\cdot), \quad i = 1, \ldots, n \quad \text{and} \quad \pi(\theta) = \text{Uniform}\left(\hat{\theta} - \frac{c}{\sqrt{n}}, \hat{\theta} + \frac{c}{\sqrt{n}}\right).$$

Also denote

$$a = -\frac{1}{n}\sum \frac{\partial^2 \log f_\theta(x_i)}{\partial \theta^2}\bigg|_{\hat{\theta}}.$$

Then the log marginal of $\underset{\sim}{x}$ is

$$\begin{aligned}
\log m(\underset{\sim}{x}) &= \log \int \prod_{i=1}^{n} f_\theta(x_i)\pi(\theta)\,\mathrm{d}\theta \\
&\approx \log \int \exp\left\{\sum_{i=1}^{n} \log f_{\hat{\theta}}(x_i) + \frac{(\theta-\hat{\theta})^2}{2}\sum \frac{\partial^2 \log f_\theta(x_i)}{\partial \theta^2}\bigg|_{\hat{\theta}}\right\}\pi(\theta)\,\mathrm{d}\theta \\
&= \sum_{i=1}^{n} \log f_{\hat{\theta}}(x_i) + \log\left[\frac{\sqrt{n}}{2c}\int_{\hat{\theta}-(c/\sqrt{n})}^{\hat{\theta}+(c/\sqrt{n})} \mathrm{e}^{-((\theta-\hat{\theta})^2/2)na}\,\mathrm{d}\theta\right] \\
&= \sum_{i=1}^{n} \log f_{\hat{\theta}}(x_i) + \log\left[\frac{1}{2c\sqrt{a}}\int_{-c\sqrt{a}}^{c\sqrt{a}} \mathrm{e}^{-(\theta^2/2)}\,\mathrm{d}\theta\right] \\
&= \sum_{i=1}^{n} \log f_{\hat{\theta}}(x_i) + \log\left[\sqrt{2\pi}\frac{\Phi(c\sqrt{a})-\Phi(-c\sqrt{a})}{2c\sqrt{a}}\right]. \qquad (14)
\end{aligned}$$

The penalty

$$-\log\left[\sqrt{2\pi}\frac{\Phi(c\sqrt{a})-\Phi(-c\sqrt{a})}{2c\sqrt{a}}\right] \to \begin{cases} 0 & \text{as } c \to 0, \\ \infty & \text{as } c \to \infty. \end{cases}$$

Thus, by suitably adjusting the value of c, we can obtain any constant penalty. The highly concentrated nature of this prior makes it somewhat undesirable from a Bayesian point of view. In the above, if we take $c = c_n$ and $c_n \to \infty$, then the penalty is proportional to $\log c_n$. Different choice of c_n would produce different values of the penalty.

Acknowledgements

We thank our referee and Dr. M.J. Bayarri for reviewing and providing suggestions on the first draft of the paper. This research was supported by the US National Science Foundation, under grants DMS-9802261 and DMS-0103265.

References

Aitkin, M., 1991. Posterior Bayes factors (disc: P128–142). J. Roy. Statist. Soc. Ser. B 53, 111–128.

Barbieri, M., Berger, J., 2004. Optimal predictive model selection. Ann. Statist. 32 (3), 870–897.

Berger, J., Ghosh, J., Mukhopadhyay, N., 2003. Approximations and consistency of Bayes factors as model dimension grows. J. Statist. Plann. Inference 112, 241–258.

Berger, J., Pericchi, L., 1995. The intrinsic Bayes factor for linear models. In: Bernardo, J.M. (Ed.), Bayesian Statistics, vol. V. London, Oxford University Press, pp. 23–42.

Berger, J., Pericchi, L., 1996. The intrinsic Bayes factor for model selection and prediction. J. Amer. Statist. Assoc. 91 (433), 109–122.

Berger, J., Pericchi, L., 1997. On the justification of default and intrinsic Bayes factors. In: Lee, J.C. (Ed.), Modeling and Prediction. Springer, New York, pp. 276–293.

Bernardo, J., Smith, A., 1994. Bayesian Theory. Wiley, Chichester.

Geisser, S., 1975. The predictive sample reuse method with applications. J. Amer. Statist. Assoc. 70, 320–328.

Gelfand, A., Dey, D., 1994. Bayesian model choice: asymptotics and exact calculations. J. Roy. Statist. Soc. Ser. B 56, 501–514.

Mukhopadhyay, N.D., 2000. Bayesian model selection for high dimensional models with prediction error loss and 0–1 loss. Ph.D. Thesis, Purdue University.

O'Hagan, A., 1995. Fractional Bayes factors for model comparisons. J. Roy. Statist. Soc. Ser. B 57 (1), 99–138.

Rissanen, J., 1986. Stochastic complexity and modeling. Ann. Statist. 14, 1080–1100.