



ELSEVIER

Statistics & Probability Letters 57 (2002) 53–63

**STATISTICS &
PROBABILITY
LETTERS**

www.elsevier.com/locate/stapro

Some connections between Bayesian and non-Bayesian methods for regression model selection

Faming Liang*

Department of Statistics and Applied Probability, The National University of Singapore, 3 Science Drive 2, Singapore 117543, Singapore

Received March 2001; received in revised form August 2001

Abstract

In this article, we study the connections between Bayesian methods and non-Bayesian methods for variable selection in multiple linear regression. We show that each of the non-Bayesian criteria, FPE_z , AIC, C_p and adjusted R^2 , has its Bayesian correspondence under an appropriate prior setting. The theoretical results are illustrated by numerical simulations. © 2002 Elsevier Science B.V. All rights reserved.

Keywords: Bayes factor; FPE_z criterion; Kullback–Leibler distance; MAP; Variable selection

1. Introduction

Consider a linear regression with a fixed number of potential predictors $\mathbf{x}_1, \dots, \mathbf{x}_k$,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (1)$$

where \mathbf{Y} is an n -vector of response, $\mathbf{X} = [\mathbf{1}, \mathbf{x}_1, \dots, \mathbf{x}_k]$ is an $n \times (k + 1)$ design matrix, $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_k)$ is a $(k + 1)$ -vector of regression coefficients, $\boldsymbol{\varepsilon} \sim \mathbf{N}_n(\mathbf{0}, \sigma^2 \mathbf{I})$, and $\boldsymbol{\beta}$ and σ^2 are unknown. The problem of interest to us is to find a subset model of the form

$$\mathbf{Y} = \mathbf{X}_p \boldsymbol{\beta}_p + \boldsymbol{\varepsilon}, \quad (2)$$

which is “best” under some criterion, where $0 \leq p \leq k$, $\mathbf{X}_p = [\mathbf{1}, \mathbf{x}_1^*, \dots, \mathbf{x}_p^*]$, $\mathbf{x}_1^*, \dots, \mathbf{x}_p^*$ are the selected predictors, and $\boldsymbol{\beta}_p = [\beta_0^*, \beta_1^*, \dots, \beta_p^*]$ is the vector of regression coefficients of the subset model. When $p = k$, the model is called the full model; when $p = 0$, the model is called the null model. Throughout this article, we assume that for any subset model the intercept term is always

* Fax: +65-872-3919.

E-mail address: stal_fm@nus.edu.sg (F. Liang).

included in and the design matrix X_p is of full column rank, and the true model is one of the subset models and it consists of k^* predictors with $0 \leq k^* \leq k$.

During the last three decades, numerous methods have been developed for this problem from both the Bayesian and non-Bayesian perspectives. The non-Bayesian methods are usually criterion-based. They work by selecting a “best” model under some criterion and then to make inferences as if the selected model were true. The most famous criteria may include $\text{Adj } \mathbb{R}^2$, FPE (Akaike, 1969), C_p (Mallows, 1973, 1995), AIC (Akaike, 1973; Sugiura, 1978; Hurvich and Tsai, 1989), PRESS (Allen, 1974; Stone, 1974), ϕ -criterion (Hannan and Quinn, 1979), GIC (Nishii, 1984) and others (Shao, 1993, 1997; Zhang, 1993; Foster and George, 1994; Zheng and Loh, 1995). The model determination usually requires a comparison for all possible 2^k models, and this is prohibitive when k is large. To reduce the computational amount required for a large value of k , some heuristic methods have been proposed by restricting the model space to a smaller number of potential subsets, e.g., branch and bound methods (Furnival and Wilson, 1974), stepwise procedures (Efroymson, 1966) and their variants. For details, see Miller (1990) and the references therein.

Bayesian methods include MAP (maximum a posteriori), Bayes factor (Jeffreys, 1961; Kass and Raftery, 1995) and predictive criteria-based methods (San Martini and Spezzaferri, 1984; Laud and Ibrahim, 1995). An overview is given in George (1999). The MAP method is to select the model with the maximum posterior probability in the model space. One famous example is BIC (Schwarz, 1978; Kass and Wasserman, 1995; Raftery, 1996; Pauler, 1998). Recently, other MAP examples are proposed based on different prior settings (George and McCulloch 1993, 1997; Phillips and Smith 1995; Geweke 1996), and Markov chain Monte Carlo (MCMC) methods are used to search for MAP models. The ergodicity of the Markov chains ensures that the MAP models will be found almost surely as the running time tends to infinity (Tierney, 1994).

A defect on the philosophy of the MAP method is that the posterior distribution is sensitive to the prior distribution imposed on the model space by users. Bayes factor gets around this difficulty by dividing the prior odds from the posterior odds, and then is to compare the marginal probabilities of models. It is defined as

$$B_{10} = \frac{(P(M_1|\mathbf{Y})/P(M_0|\mathbf{Y}))}{(P(M_1)/P(M_0))} = \frac{P(\mathbf{Y}|M_1)}{P(\mathbf{Y}|M_0)},$$

where M_1 and M_0 denote two models under comparison. When $B_{10} > 1$, model M_1 is supported, otherwise, model M_0 is supported. However, the Bayes factor is far from perfection. When improper priors are imposed on some model-specific parameters, the Bayes factor can only be determined up to a constant, and in this case it does not make sense for the model comparison any more. A variety of approximate Bayesian factors have been proposed to overcome the difficulty. Geisser and Eddy (1979) and Gelfand et al. (1992) proposed to use a cross-validation predictive distribution to replace the marginal distribution, and the replacement yields the pseudo-Bayes factor B'_{10} ,

$$B'_{10} = \prod_{i=1}^n \frac{P(y_i|\mathbf{Y}_{(i)}, M_1)}{P(y_i|\mathbf{Y}_{(i)}, M_0)},$$

where $\mathbf{Y}_{(i)}$ denotes the data set with the i th case omitted. The other approximators and related references can be found in Spiegelhalter and Smith (1982), Perichi (1984), Aitkin (1991), Gelfand and Dey (1994), O'Hagan (1995), Berger and Pericchi (1996) and Moreno, Bertolino, and Racugno (1998).

Given the plethora of model selection criteria, a need exists for research which unifies existing criteria under common themes. This article contributes towards fulfilling this need by exploring some connections between Bayesian and non-Bayesian methods for variable selection in a multiple linear regression. Under an appropriate prior setting, we show that the MAP methods correspond to the FPE_α criteria, and that the pseudo-Bayes factor corresponds to the $\text{Adj } \mathbb{R}^2$ criterion and they both minimize the Kullback–Leibler distance between the predictive likelihoods of the true and candidate models.

This article is organized as follows. In Section 2 we study the connections between FPE_α criteria and MAP methods. In Section 3 we study the connections between the pseudo-Bayes factor and the $\text{Adj } \mathbb{R}^2$ criterion. In Section 4 we present some simulation results to confirm the theoretical results of this article.

2. FPE_α criteria and MAP models

The original FPE criterion is proposed by Akaike (1969) to minimize the final prediction error (FPE). For a particular subset model M_p , FPE is defined as

$$E(z_i - \hat{z}_i)^2 = \sigma^2 \left(1 + \frac{p'}{n} \right),$$

where $p' = p + 1$ is the total number of explanatory variables included in the subset model (including the intercept term), z_i denotes a new observation independent of the ones used for model determination, and \hat{z}_i denotes the prediction value of z_i . Akaike's derivation estimated σ^2 with $\hat{\sigma}_p^2 = \text{RSS}_p / (n - p')$, and this substitution yields

$$\text{FPE}(M_p) = \text{RSS}_p \frac{n + p'}{n - p'}$$

by ignoring a constant factor $1/n$ or equivalently

$$\text{FPE}(M_p) = \text{RSS}_p + 2p' \hat{\sigma}_p^2,$$

where $\text{RSS}_p = \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}_p(\mathbf{X}'_p\mathbf{X}_p)^{-1}\mathbf{X}'_p\mathbf{Y}$ is the residual sum of squares of model M_p . Later, Shibata (1984) suggested that $\hat{\sigma}_p^2$ to be replaced by $\hat{\sigma}_k^2 = \text{RSS}_k / (n - k - 1)$, and proposed a generalized form for the FPE criteria,

$$\text{FPE}_\alpha(M_p) = \text{RSS}_p + \alpha p' \hat{\sigma}_k^2, \tag{3}$$

where α is a penalty coefficient chosen by users. As is known, many criteria can be represented in the form of FPE_α with different values of α . For example, $\alpha = 2$ yields C_p and, approximately, AIC and PRESS; $\alpha = 2 \log k$ yields the risk inflation criterion (Foster and George, 1994); $\alpha = c \log \log n$ yields ϕ -criterion (Hannan and Quinn, 1979); $\alpha = \log n$ yields BIC; and if α is a function of n and $\lim_{n \rightarrow \infty} \alpha/n = 0$, it yields the GIC criterion (Nishii, 1984).

To study the connections between FPE_α criteria and MAP methods, we consider the following prior setting for model M_p . First, the model is reparameterized as

$$\mathbf{Y} = \mathbf{Z}_p \boldsymbol{\gamma}_p + \boldsymbol{\varepsilon}, \tag{4}$$

where a QR decomposition is performed on \mathbf{X}_p and $\mathbf{X}_p = \mathbf{Z}_p \mathbf{R}_p$. Thus, \mathbf{Z}_p is an $n \times (p+1)$ matrix with orthonormal columns, \mathbf{R}_p is upper triangular, and $\boldsymbol{\gamma}_p = \mathbf{R}_p \boldsymbol{\beta}_p$. The likelihood function of the model is

$$L_p(\mathbf{Y}|\mathbf{X}, M_p, \boldsymbol{\gamma}_p, \sigma^2) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp \left\{ -\frac{1}{2\sigma^2} (\mathbf{Y} - \mathbf{Z}_p \boldsymbol{\gamma}_p)' (\mathbf{Y} - \mathbf{Z}_p \boldsymbol{\gamma}_p) \right\}. \quad (5)$$

We assume that all k predictors are linearly independent, and each has a prior probability μ to be included in the model. Thus, the prior probability imposed on the model M_p is

$$P(M_p) = \mu^p (1 - \mu)^{k-p}, \quad (6)$$

where μ is a hyperparameter to be determined later. We further assume that $\boldsymbol{\gamma}_p$ and σ^2 are a priori independent, and they are subject to the Jeffreys non-informative priors,

$$P(\boldsymbol{\gamma}_p | M_p) \propto 1, \quad (7)$$

$$P(\sigma^2) \propto \frac{1}{\sigma^2}. \quad (8)$$

Multiplying (5)–(8), we get the posterior distribution (up to a multiplicative constant),

$$\begin{aligned} P(M_p, \boldsymbol{\gamma}_p, \sigma^2 | \mathbf{Y}) &\propto P(\mathbf{Y} | \mathbf{X}, \boldsymbol{\gamma}_p, \sigma^2, M_p) P(\boldsymbol{\gamma}_p | M_p) P(\sigma^2) P(M_p) \\ &= \mu^p (1 - \mu)^{k-p} \frac{1}{(\sqrt{2\pi})^n} \frac{1}{(\sigma^2)^{(n/2+1)}} \exp \left\{ -\frac{1}{2\sigma^2} \|\hat{\boldsymbol{\varepsilon}}\|^2 \right\} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=0}^p (\gamma_{pi} - \hat{\gamma}_{pi})^2 \right\}, \end{aligned} \quad (9)$$

where $\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \mathbf{X}_p \hat{\boldsymbol{\beta}}_p = \mathbf{Y} - \mathbf{Z}_p \hat{\boldsymbol{\gamma}}_p$ is the residual vector. In the derivation we used that

$$\|\mathbf{Y} - \mathbf{X}_p \boldsymbol{\beta}_p\|^2 = \|\mathbf{Y} - \mathbf{Z}_p \boldsymbol{\gamma}_p\|^2 = \|\hat{\boldsymbol{\varepsilon}}\|^2 + \|\boldsymbol{\gamma}_p - \hat{\boldsymbol{\gamma}}_p\|^2.$$

Integrating out $\boldsymbol{\gamma}_p$ and σ^2 from (9) and then taking a logarithm, we get the log-posterior of model M_p (up to an additive constant),

$$\begin{aligned} \log P(M_p | \mathbf{Y}) &= p \log \left(\frac{\mu}{1 - \mu} \right) - \frac{n - p - 1}{2} \log(\pi) - \frac{n - p - 1}{2} \log(\text{RSS}_p) \\ &\quad + \log \Gamma \left(\frac{n - p - 1}{2} \right). \end{aligned} \quad (10)$$

This posterior distribution only depends on one tunable parameter μ , the value of which reflects our prior knowledge on the number of predictors of the regression. If we think more predictors should be included in the regression, μ can be set to a large value, otherwise, it should be set to a small value. A particular form of μ is of special interest,

$$\mu_\alpha = \frac{1}{1 + \sqrt{2\pi} \hat{\sigma}_k \exp(\alpha/2 + 1/(2(n-1)))},$$

where $\hat{\sigma}_k = \sqrt{\text{RSS}_k / (n - k - 1)}$, α is a value specified by users. The following theorem shows the correspondence between FPE_α criteria and MAP methods for some specific values of α .

Theorem 2.1. Under the prior setting (6)–(8), when $\mu = \mu_x$ and $n \gg p$, we have

$$\log P(M_p | \mathbf{Y}) \approx \text{constant} - \text{FPE}_x(p) / [2\hat{\sigma}_k^2] \tag{11}$$

for the models with $\varepsilon_p \approx 0$, where $\varepsilon_p = (\hat{\sigma}_p^2 - \hat{\sigma}_k^2) / \hat{\sigma}_k^2$.

Proof. By the Stirling approximation, when $n \gg p$ we have

$$\log \left(\Gamma \left(\frac{n-p-1}{2} \right) \right) \approx -\frac{n-p-1}{2} + \frac{n-p-2}{2} \log \left(\frac{n-p-1}{2} \right) + \frac{1}{2} \log(2\pi). \tag{12}$$

Substituting (12) into (10), we have

$$\begin{aligned} \log P(M_p | \mathbf{Y}) &\approx p \log \left(\frac{\mu}{1-\mu} \right) - \frac{n-p-2}{2} \log(2\pi) - \frac{n-p-1}{2} \\ &\quad - \frac{1}{2} \log \left(\frac{n-p-1}{2} \right) - \frac{n-p-1}{2} \log \hat{\sigma}_p^2 \\ &\approx \text{constant} + p \log \left(\frac{\mu}{1-\mu} \right) + \frac{p}{2} [1 + \log(2\pi)] - \frac{1}{2} \log \left(1 - \frac{p}{n-1} \right) \\ &\quad + \frac{p}{2} \log \hat{\sigma}_k^2 - \frac{n-p-1}{2} \log(\hat{\sigma}_p^2 / \hat{\sigma}_k^2). \end{aligned} \tag{13}$$

The uniqueness of the full model allows us to regard $\hat{\sigma}_k^2$ as a constant in the derivation. With the Taylor expansion, we have $\log(\hat{\sigma}_p^2 / \hat{\sigma}_k^2) = \log(1 + \varepsilon_p) \approx \varepsilon_p$ for $\varepsilon_p \approx 0$, and $\log(1 - p/(n-1)) \approx -p/(n-1)$. Hence,

$$\begin{aligned} \log P(M_p | \mathbf{Y}) &\approx \text{constant} + p \log \left(\frac{\mu}{1-\mu} \right) + \frac{p}{2} [1 + \log(2\pi)] + \frac{p}{2(n-1)} \\ &\quad + \frac{p}{2} \log \hat{\sigma}_k^2 - \frac{n-p-1}{2} (\hat{\sigma}_p^2 / \hat{\sigma}_k^2 - 1) \\ &= \text{constant} - \frac{1}{2\hat{\sigma}_k^2} [\text{RSS}_p + (I)], \end{aligned} \tag{14}$$

where

$$(I) = -p' \hat{\sigma}_k^2 \left[2 \log \left(\frac{\mu}{1-\mu} \right) + \frac{1}{(n-1)} + \log(2\pi) + \log \hat{\sigma}_k^2 \right].$$

When $\mu = \mu_x$, $(I) = \alpha p' \hat{\sigma}_k^2$. The proof is completed. \square

In Theorem 2.1, the condition $\varepsilon_p \approx 0$ can be satisfied by many models, including the true model, all overfitting models, and a part of underfitting models. For the true and overfitting models, $\hat{\sigma}_p^2$ and $\hat{\sigma}_k^2$ are both consistent estimators of σ^2 . For the underfitting models, $\hat{\sigma}_p^2$ is biased upward (Rencher, 2000, p. 157). If $\hat{\sigma}_p^2$ is far from $\hat{\sigma}_k^2$, $-\text{FPE}_x(p) / [2\hat{\sigma}_k^2]$ provides an under-approximator for $\log P(M_p | \mathbf{Y})$, fortunately, the models of this kind usually have a very small value of total masses

in the posterior $P(M_p|\mathbf{Y})$. Hence, approximately we have

$$P(M_p|\mathbf{Y}) \propto \exp\{-\text{FPE}_x(p)/[2\hat{\sigma}_k^2]\}. \quad (15)$$

This is confirmed by our numerical results in Section 4. A similar relationship between MAP and the C_p criterion was obtained by Liang, Truong, and Wong (2001), under a quite different prior setting.

3. Predictive information, pseudo-Bayes factor and $\text{Adj } \mathbb{R}^2$

Under the Bayesian framework the predictive likelihood of a candidate model M can be written as

$$\mathbf{f} = \prod_{i=1}^m f(z_i|\mathbf{Y}, M), \quad (16)$$

where \mathbf{Y} denotes the n observations used for the model determination, z_1, \dots, z_m denote the m new observations independent of \mathbf{Y} . Similarly, the predictive likelihood of the true model M_* can be written as

$$\mathbf{f}_* = \prod_{i=1}^m f(z_i|\mathbf{Y}, M_*). \quad (17)$$

A useful measure for the discrepancy between \mathbf{f} and \mathbf{f}_* is the Kullback–Leibler distance (up to a constant)

$$D(M, M_*) = -\frac{2}{m} E_* \log(\mathbf{f}), \quad (18)$$

where E_* denotes taking expectation with respect to \mathbf{f}_* . The $D(M, M_*)$ is called the predictive information in this article. It can be estimated through a cross-validation statistic,

$$\hat{D}(M, M_*) = -\frac{2}{n} \sum_{i=1}^n \log f(y_i|\mathbf{Y}_{(i)}, M), \quad (19)$$

where $\mathbf{Y}_{(i)}$ denotes an $(n-1)$ -vector of observations with y_i omitted. Comparing with B'_{10} , it is easy to see the equivalence between the pseudo-Bayes factor and minimizing $D(M, M_*)$ for variable selection in a multiple linear regression. To compute $\hat{D}(M, M_*)$, we have the following theorem.

Theorem 3.1. *Under the prior setting (6)–(8), minimizing the predictive information $D(M, M_*)$ is approximately equivalent to minimizing the studentized residual sum of squares (SRSS), which for a particular model M_p is*

$$\text{SRSS}(M_p) = \sum_{i=1}^n d_i^2,$$

where $d_i = e_i/[\hat{\sigma}_k \sqrt{1 - h_{ii}}]$, e_i is the OLS residual for the i th case, h_{ii} is the i th diagonal element of the hat matrix $\mathbf{H}_p = \mathbf{X}_p(\mathbf{X}'_p \mathbf{X}_p)^{-1} \mathbf{X}'_p$, and $\hat{\sigma}_k = \sqrt{\text{RSS}_k/(n-k-1)}$.

Proof. With the identity

$$P(y_i | \mathbf{Y}_{(i)}, M) = \frac{P(M | \mathbf{Y})}{P(M | \mathbf{Y}_{(i)})} \frac{P(\mathbf{Y})}{P(\mathbf{Y}_{(i)})},$$

we have

$$\hat{D}(M, M_*) = -\frac{2}{n} \sum_{i=1}^n [\log P(M | \mathbf{Y}) - \log P(M | \mathbf{Y}_{(i)})] - \frac{2}{n} \sum_{i=1}^n [\log P(\mathbf{Y}) - \log P(\mathbf{Y}_{(i)})].$$

Hence, minimizing $\hat{D}(M, M_*)$ is equivalent to minimizing

$$\hat{D}'(M, M_*) = -2 \sum_{i=1}^n [\log P(M | \mathbf{Y}) - \log P(M | \mathbf{Y}_{(i)})].$$

Similar to (13), we have

$$\begin{aligned} \log P(M | \mathbf{Y}) &\approx \text{constant} + p \log\left(\frac{\mu}{1-\mu}\right) + \frac{p}{2}[1 + \log(2\pi)] - \frac{1}{2} \log\left(1 - \frac{p}{n-1}\right) \\ &\quad + \frac{p}{2} \log \hat{\sigma}_0^2 - \frac{n-p-1}{2} \log(\hat{\sigma}_p^2 / \hat{\sigma}_0^2) \\ &\approx \text{constant} + p \log\left(\frac{\mu}{1-\mu}\right) + \frac{p}{2}[1 + \log(2\pi)] + \frac{p}{2(n-1)} \\ &\quad + \frac{p}{2} \log \hat{\sigma}_0^2 - \frac{n-p-1}{2} (\hat{\sigma}_p^2 / \hat{\sigma}_0^2 - 1) \\ &= \text{constant} - \frac{1}{2\hat{\sigma}_0^2} [\text{RSS}_p + (II)], \end{aligned} \tag{20}$$

where $\hat{\sigma}_0^2 = \sum_{i=1}^n (y_i - \bar{y})^2 / (n-1)$ can be regarded as a constant in the derivation, and

$$(II) = -p\hat{\sigma}_0^2 \left[2 \log\left(\frac{\mu}{1-\mu}\right) + \frac{1}{(n-1)} + \log(2\pi) + \log \hat{\sigma}_0^2 \right].$$

Let $\mu = [1 + \sqrt{2\pi}\hat{\sigma}_0 \exp(1/(2(n-1)))]^{-1}$, we have $(II) = 0$. In the derivation of (20), we used the Taylor expansion, $\log(\hat{\sigma}_p^2 / \hat{\sigma}_0^2) \approx 1 + (\hat{\sigma}_p^2 - \hat{\sigma}_0^2) / \hat{\sigma}_0^2$ by ignoring the higher order terms ($|(\hat{\sigma}_p^2 - \hat{\sigma}_0^2) / \hat{\sigma}_0^2| \leq 1, \forall M_p$). Hence, with the special value of μ , we have

$$\hat{D}'(M, M_*) \approx \sum_{i=1}^n \left[\frac{\text{RSS}_p}{\hat{\sigma}_0^2} - \frac{\text{RSS}_{p(i)}}{\hat{\sigma}_{0(i)}^2} \right], \tag{21}$$

where $\hat{\sigma}_{0(i)}^2$ denotes the estimate of σ^2 from the null model with the i th case omitted. When n is large, we have $\hat{\sigma}_{0(i)}^2 \approx \hat{\sigma}_0^2$ for $i = 1, \dots, n$. Thus

$$\hat{D}'(M, M_*) \approx \sum_{i=1}^n \frac{\text{RSS}_p - \text{RSS}_{p(i)}}{\hat{\sigma}_0^2},$$

or equivalently,

$$\hat{D}''(M, M_*) \approx \sum_{i=1}^n \frac{\text{RSS}_p - \text{RSS}_{p^{(i)}}}{\hat{\sigma}_k^2},$$

where $\text{RSS}_{p^{(i)}}$ denotes the residual sum of squares of model M_p with the i th case omitted. An algebraically equivalent expression for $\text{RSS}_p - \text{RSS}_{p^{(i)}}$ is

$$\text{RSS}_p - \text{RSS}_{p^{(i)}} = \frac{e_i^2}{1 - h_{ii}}.$$

An estimated variance of e_i is

$$s^2\{e_i\} = \hat{\sigma}_k^2(1 - h_{ii}). \quad (22)$$

Hence, $\hat{D}''(M, M_*) \approx \sum_{i=1}^n d_i^2$, where d_i is the studentized residual for the i th observation. \square

The conclusion of Theorem 3.1 is independent of the prior setting (6). This is easily known from the form of $\hat{D}(M, M_*)$. In the proof, a specific value of μ is chosen only to facilitate the derivation. This theorem shows that $D(M, M_*)$ can be measured approximately from a single regression run without requiring n separate runs, each time omitting one of the n cases. Note that $\text{trace}(\mathbf{H}_p) = \sum_{i=0}^n h_{ii} = p'$. When $n \gg k$, we have

$$\begin{aligned} \log\{\text{SRSS}(M_p)\} &\approx \log\{\text{RSS}_p/(1 - p'/n)\} - \log(\hat{\sigma}_k^2) \\ &= \log(\hat{\sigma}_p^2) - \log(\hat{\sigma}_k^2) + \log n. \end{aligned} \quad (23)$$

To study the connections between the pseudo-Bayes factor and the $\text{Adj } \mathbb{R}^2$ criterion, we first recall the $\text{Adj } \mathbb{R}^2$ statistic.

Definition 3.1. Adjusted \mathbb{R}^2 statistic.

$$\text{Adj } \mathbb{R}^2 = 1 - \frac{\hat{\sigma}_p^2}{\text{MS}(\text{total})},$$

where $\hat{\sigma}_p^2 = \text{RSS}_p/(n - p')$ and $\text{MS}(\text{total}) = \sum_{i=1}^n (y_i - \bar{y})^2/(n - 1)$.

By comparing (23) and $\text{Adj } \mathbb{R}^2$, it is clear that minimizing SRSS is approximately equivalent to maximizing $\text{Adj } \mathbb{R}^2$.

Summarizing this section, we have the following conclusion: under the Jeffreys non-informative priors, the pseudo-Bayes factor corresponds to the $\text{Adj } \mathbb{R}^2$ criterion, and they both minimize the predictive Kullback–Leibler discrepancy between the predictive likelihoods of the true and candidate models.

4. Numerical examples

These data come from Draper and Smith (1981). They consist of 9 predictors and 25 observations, which are taken at intervals from a steam plant at a large industrial concern. These data have been used by many authors to illustrate the regression model selection (Miller, 1990).

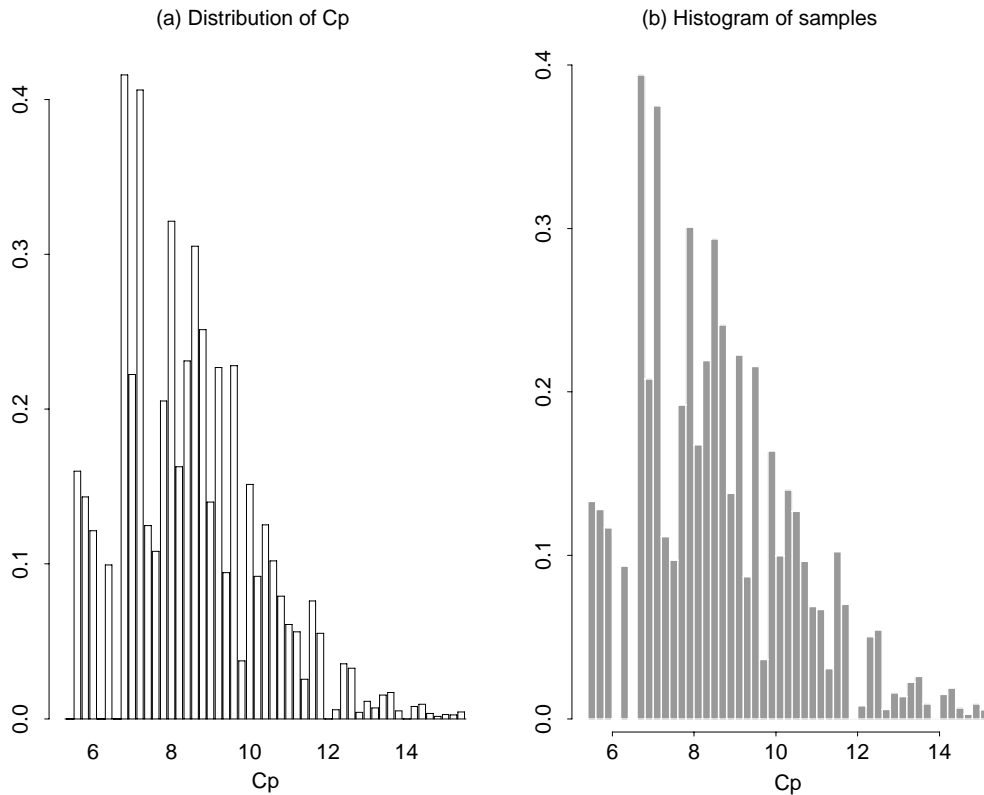


Fig. 1. A comparison of the true Boltzmann distribution defined on C_p values (a) and estimated using posterior samples (b) for the steam plant data.

The posterior distribution (10) was simulated using evolutionary Monte Carlo (EMC) (Liang and Wong, 2000). The Metropolis–Hastings algorithm (Metropolis et al., 1953; Hastings, 1970) or reversible jump MCMC (Green, 1995) may also serve the same purpose for this example. In the simulation we set $\mu = \mu_2$. Fig. 1(b) is the histogram of the C_p values of the models sampled in one run of EMC. The sample size is 20000. For comparison, Fig. 1(a) shows the Boltzmann distribution $\Pr(M) \propto \exp\{-C_p(M)/2\}$. The high similarity of the two plots shows that C_p provides a good approximation to the log-posterior when $\mu = \mu_2$. The result of Theorem 2.1 is confirmed.

References

- Aitkin, M., 1991. Posterior Bayes factors. *J. Roy. Statist. Soc. B* 53, 111–142.
- Akaike, H., 1969. Fitting autoregressive models for prediction. *Ann. Inst. Statist. Math.* 21, 243–247.
- Akaike, H., 1973. Information theory and the extension of the maximum likelihood principle. In: Petrov, B.N., Czaki, F. (Eds.), *Proceedings of the International Symposium on Information Theory*. Akademia Kiadoó, Budapest, pp. 267–281.
- Allen, D.M., 1974. The relationship between variable selection and data augmentation and a method for prediction. *Technometrics* 16, 125–127.
- Berger, J.O., Pericchi, L.R., 1996. The intrinsic Bayes factor for model selection and prediction. *J. Amer. Statist. Assoc.* 91, 109–122.

- Draper, N.R., Smith, H., 1981. *Applied Regression Analysis*, 2nd Edition. Wiley, New York.
- Efroymson, M.A., 1966. Multiple regression analysis. In: Ralston, A., Wilf, H.S. (Eds.), *Mathematical Methods for Digital Computers*. Wiley, New York, pp. 191–203.
- Foster, D.P., George, E.I., 1994. The risk inflation criterion for multiple regression. *Ann. Statist.* 22, 1947–1975.
- Furnival, G.M., Wilson, R.W., 1974. Regression by leaps and bounds. *Technometrics* 16, 499–511.
- Geisser, S., Eddy, W., 1979. A predictive approach to model selection. *J. Amer. Statist. Assoc.* 74, 153–160.
- Gelfand, A., Dey, D., 1994. Bayesian model choice: asymptotic and exact calculations. *J. Roy. Statist. Soc. B* 56, 501–514.
- Gelfand, A.E., Dey, D.K., 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*, Vol. 4. Oxford University Press, Oxford, pp. 147–167.
- George, E.I., 1999. Bayesian model selection. In: Kotz, S., Read, C., Banks, D. (Eds.), *Encyclopedia of Statistical Sciences Update*, Vol. 3. Wiley, New York, pp. 39–46.
- George, E.I., McCulloch, R.E., 1993. Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* 88, 881–889.
- George, E.I., McCulloch, R.E., 1997. Approaches for Bayesian variable selection. *Statist. Sinica* 7, 339–373.
- Geweke, J., 1996. Variable selection and model comparison in regression. In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (Eds.), *Bayesian Statistics*, Vol. 5. Oxford University Press, Oxford, pp. 609–620.
- Green, P.J., 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 82, 711–732.
- Hannan, E.J., Quinn, B.G., 1979. The determination of the order of an autoregression. *J. Roy. Statist. Soc. B* 41, 190–195.
- Hastings, W.K., 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57, 97–109.
- Hurvich, C.M., Tsai, C.L., 1989. Regression and time series model selection in small samples. *Biometrika* 76, 297–307.
- Jeffreys, H., 1961. *Theory of Probability*, 3rd Edition. Oxford University Press, London.
- Kass, R.E., Raftery, A.E., 1995. Bayes factors. *J. Amer. Statist. Assoc.* 90, 773–795.
- Kass, R.E., Wasserman, L., 1995. A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J. Amer. Statist. Assoc.* 90, 928–934.
- Laud, P.W., Ibrahim, J.G., 1995. Predictive model selection. *J. Roy. Statist. Soc. B* 57, 247–262.
- Liang, F., Wong, W.H., 2000. Evolutionary Monte Carlo: applications to C_p model sampling and change point problem. *Statist. Sinica* 10, 317–342.
- Liang, F., Truong, Y.K., Wong, W.H., 2001. Automatic Bayesian model averaging for linear regression and applications in Bayesian curve fitting. *Statist. Sinica* 11, 1005–1029.
- Mallows, C.L., 1973. Some comments on C_p . *Technometrics* 15, 661–676.
- Mallows, C.L., 1995. More comments on C_p . *Technometrics* 37, 362–372.
- Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E., 1953. Equation of state calculations by fast computing machines. *J. Chem. Phys.* 21, 1087–1091.
- Miller, A.J., 1990. *Subset Selection in Regression*. Chapman and Hall, New York.
- Moreno, E., Bertolino, F., Racugno, W., 1998. An intrinsic limiting procedure for model selection and hypotheses testing. *J. Amer. Statist. Assoc.* 93, 1451–1460.
- Nishii, R., 1984. Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist.* 12, 758–765.
- O'Hagan, A., 1995. Fractional Bayes factor for model comparison (with discussion). *J. Roy. Statist. Soc. B* 57, 99–138.
- Pauler, D., 1998. The Schwarz criterion and related methods for the normal linear model. *Biometrika* 85, 13–27.
- Perichi, L., 1984. An alternative to the standard Bayesian procedure for discrimination between normal linear models. *Biometrika* 71, 576–586.
- Phillips, D.B., Smith, A.F.M., 1995. Bayesian model comparison via jump diffusions. In: Gilks, W.R., Richardson, S., Spiegelhalter, D.J. (Eds.), *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London, pp. 215–239.
- Raftery, A.E., 1996. Approximate Bayes factors and accounting for model uncertainty in generalized linear models. *Biometrika* 83, 251–266.
- Rencher, A.C., 2000. *Linear Models in Statistics*. Wiley, New York.
- San Martini, A., Spezzaferrri, F., 1984. A predictive model selection criterion. *J. Roy. Statist. Soc. B* 46, 296–303.
- Schwarz, G., 1978. Estimating the dimension of a model. *Ann. Statist.* 6, 461–464.
- Shao, J., 1993. Linear model selection by cross-validation. *J. Amer. Statist. Assoc.* 88, 486–494.
- Shao, J., 1997. An asymptotic theory for linear model selection. *Statist. Sinica* 7, 221–264.

- Shibata, R., 1984. Approximate efficiency of a selection procedure for the number of regression variables. *Biometrika* 71, 43–49.
- Spiegelhalter, D.J., Smith, A.F.M., 1982. Bayes factors for linear and log-linear models with vague prior information. *J. Roy. Statist. Soc. B* 44, 377–387.
- Stone, M., 1974. Cross-validation choice and assessment of statistical predictions. *J. Roy. Statist. Soc. B* 36, 111–147.
- Sugiura, N., 1978. Further analysis of the data by Akaike's information criterion and the finite corrections. *Commun. Statist. A* 7, 13–26.
- Tierney, L., 1994. Markov chains for exploring posterior distributions (with discussion). *Ann. Statist.* 22, 1701–1762.
- Zhang, P., 1993. Model selection via multifold cross-validation. *Ann. Statist.* 21, 299–313.
- Zheng, X., Loh, W.Y., 1995. Consistent variable selection in linear models. *J. Amer. Statist. Assoc.* 90, 151–156.