Statistica Sinica **11**(2001), 1005-1029

AUTOMATIC BAYESIAN MODEL AVERAGING FOR LINEAR REGRESSION AND APPLICATIONS IN BAYESIAN CURVE FITTING

Faming Liang[†], Young K Truong[†] and Wing Hung Wong[‡]

[†] *The National University of Singapore and* [‡] *Harvard School of Public Health*

Abstract: With the development of MCMC methods, Bayesian methods play a more and more important role in model selection and statistical prediction. However, the sensitivity of the methods to prior distributions has caused much difficulty to users. In the context of multiple linear regression, we propose an automatic prior setting, in which there is no parameter to be specified by users. Under the prior setting, we show that sampling from the posterior distribution is approximately equivalent to sampling from a Boltzmann distribution defined on C_p values. The numerical results show that the Bayesian model averaging procedure resulted from the automatic prior settin provides a significant improvement in predictive performance over other two procedures proposed in the literature. The procedure is extended to the problem of Bayesian curve fitting with regression splines. Evolutionary Monte Carlo is used to sample from the posterior distributions.

Key words and phrases: Bayesian model averaging, curve fitting, evolutionary Monte Carlo, Mallows' C_p, Markov chain Monte Carlo.

1. Introduction

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

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{Y}$ is an n-vector of response, $\boldsymbol{X} = [\mathbf{1}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_k]$ is an $n \times (k+1)$ design matrix, $\boldsymbol{\epsilon} \sim N_n(0, \sigma^2 I)$, $\boldsymbol{\beta}$ and σ^2 are unknown. Model selection and inference for linear regression have been extensively discussed in statistics. They are mainly investigated in two approaches, the criterion-based approach and the fully Bayesian approach.

The criterion-based approach works by selecting the "best" model under som criterion and then to make inferences as if the selected model were true. The most famous criteria may include adjusted R^2, C_p (Mallows (1973)), AIC (Akaike (1973)), BIC (Schwarz (1978)), PRESS (Allen (1974)), default Bayes factor (O'Hagan (1995), Berger and Pericchi (1996)), and some predictive criteria (Geisser and Eddy (1979), San Martini and Spezzaferri (1984), Laud and Ibrahim (1995)). The determination for the "best" model usually requires a comparison of all possible 2^k models, and this is prohibitive when k is large. Raftery, Madigan and Hoeting (1997) stated that selection of a single model ignores the uncertainty of the model itself and, as a consequence, the uncertainty of quantities of interest can be underestimated. More discussions on the issue can be found in Draper (1995), Raftery (1996), Hoeting, Madigan, Raftery and Volinsky (1999), and the references therein.

The fully Bayesian approach is to make inferences from a posterior distribution defined on the model space. An overview for this approach is given in George (1999). For example, if a quantity Δ is of interest, the Bayesian estimate can be obtained by averaging the quantities under each model weighted by the corresponding posterior probabilities. That is,

$$E(\Delta|D) = \sum_{i=0}^{K} \Delta_i P(M_i|D), \tag{2}$$

where D denotes the data, M_i denotes a model for D, K denotes the number of all models under consideration, Δ_i is an estimate of Δ based on M_i,

$$P(M_i|D) = \frac{P(D|M_i)P(M_i)}{\sum_{j=1}^{K} P(D|M_j)P(M_j)}, \tag{3}$$

where $P(M_i)$ is the prior probability of M_i, $P(D|M_j)$ is the likelihood function of D given M_j, and

$$P(D|M_j) = \int P(D|\vartheta_j, M_j)P(\vartheta_j|M_j)d\vartheta_j, \tag{4}$$

where ϑ_j is a parameter vector associated with M_j. One advantage of this approach is that it accounts for the model uncertainty beyond the single model selection. Also, it enables optimal prediction if the predictive performance is measured in the logarithmic scoring rule (Madigan and Raftery (1994)). However, it suffers from several difficulties in practical applications.

The first difficulty is that this approach requires a daunting specification for prior probabilities over a large class of models under consideration and prior specification for the specific parameters associated with each model. Furthermore, this approach is known to be rather sensitive to the prior specifications (Kass and Raftery (1995), George (1999)).

The second difficulty is that the sums over all possible models in equation (2) will be impractical when K is large. One approach to get around this difficulty is Occam's window method (Madigan and Raftery (1994)), which approximates the summation by averaging over a reduced set of models. The other approach is Markov chain Monte Carlo, which works by simulating a Markov chain from the posterior distribution $P(M|D)$. Let $M_0, \ldots, M_t, \ldots$ denote a series of models sampled from the Markov chain. Under suitable regularity conditions, the average

$$\hat{\Delta} = \frac{1}{m} \sum_{t=1}^{m} \Delta_t \tag{5}$$

converges with probability 1 to $E(\Delta|D)$ as $m \to \infty$ (Kass and Raftery (1995)). Related papers include Mitchell and Beauchamp (1988), George and McCulloch (1993, 1997), Gelfand, Dey and Chang (1992), Gelfand (1995), Madigan and Raftery (1994), Madigan and York (1995), Carlin and Chib (1995), Phillips and Smith (1995), Geweke (1996), Raftery, Madigan and Hoeting (1997), Clyde (1999), Hoeting, Madigan, Raftery and Volinsky (1999), and Fernadez, Ley and Stell (1999).

Finally, we would like to mention that computing the likelihood function for a given model is a more complex problem when its analytical form is not available. In that case, reversible jump MCMC (Green (1995)) may serve as a good tool for the Markov chain to move jointly over the model space and the parameter space. An alternative approach is proposed by George and McCulloch (1993), who get around the problem of dimension jumping by assuming a continuous distribution concentrated around zero for these coefficients and use the Gibbs sampler to sample from the parameter and model space.

In this paper, we propose a Bayesian model averaging procedure for multiple linear regression under an appropriate prior setting. The procedure is automatic and, in most cases, it can be applied without any user-specified parameter. Under the setting, we show that sampling from the posterior distribution over the model space is approximately equivalent to sampling from a Boltzmann distribution specified by C_p values. The procedure is compared with those proposed in the literature through numerical examples. The results show that it provides a significant improvement over them in predictive performance. Evolutionary Monte Carlo (Liang and Wong (2000)) is used to sample from the posterior distribution.

The remaining part of this article is organized as follows. In Section 2, we describe the automatic prior setting, under which the connections between the Bayesian approach and the C_p-criterion are studied. In Section 3, we describe the computational implementation for the new Bayesian approach. In Section 4, we present some numerical examples. In Section 5, we apply the new Bayesian approach to the problem of curve fitting with least squares splines.

2. Automatic Bayesian Model Averaging

2.1. Automatic Bayesian approach

Each model that we consider is of the form

$$\boldsymbol{Y} = \boldsymbol{X}_p \boldsymbol{\beta}_p + \boldsymbol{\epsilon}, \tag{6}$$

where $\boldsymbol{Y}$ is an n-vector of response, $p \leq k < n$, $\boldsymbol{X}_p = [\mathbf{1}, \boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_p^*]$ is an $n \times (p+1)$ design matrix, $\{\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_p^*\}$ is a subset selected from all potential predictors $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$, $\boldsymbol{\beta}_p = (\beta_{p0}, \beta_{p1}, \ldots, \beta_{pp})'$, $\boldsymbol{\epsilon} \sim N_n(0, \sigma^2 I)$, and $\boldsymbol{\beta}_p$ and σ^2 are considered to be unknown. When $p = k$, the model is called a full model. Throughout this article, we assume that the intercept term is included in the subset models, and any subset model $\boldsymbol{X}_p$ is of full column rank.

We first re-parameterize the model as follows.

$$\boldsymbol{Y} = \boldsymbol{Z}_p \gamma_p + \boldsymbol{\epsilon}, \tag{7}$$

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

Let $\xi = (\xi_1, \ldots, \xi_k)$ represent a model, where each element is a binary variable indicating the inclusion of the corresponding predictor. Let $\xi^{(p)}$ denote a model with p predictors. The set of free parameters of model $\xi^{(p)}$ is $\theta = (\xi_1, \ldots, \xi_k, \gamma_{p0}, \gamma_{p1}, \ldots, \gamma_{pp}, \sigma^2) = (\xi^{(p)}, \boldsymbol{\gamma}_p, \sigma^2)$.

The likelihood function of the model is

$$L_p(\boldsymbol{Y}|\boldsymbol{X}, \xi^{(p)}, \boldsymbol{\gamma}_p, \sigma^2) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\{-\frac{\|\boldsymbol{Y} - \boldsymbol{Z}_p \boldsymbol{\gamma}_p\|^2}{2\sigma^2}\}. \tag{8}$$

The prior distributions for θ are specified as follows. We assume that all potential predictors are linearly independent, and each has a prior probability μ to be included in the regression. Thus the prior probability of model $\xi^{(p)}$ is

$$p(\xi^{(p)}) = \mu^p (1-\mu)^{k-p}, \tag{9}$$

where μ is a hyperparameter to be determined later. We further assume that $\boldsymbol{\gamma}_p$ and σ^2 are *a priori* independent, and that they are subject to the following prior distributions:

$$P(\boldsymbol{\gamma}_p|\xi^{(p)}) = \frac{1}{(\sqrt{2\pi}\tau_p)^{p+1}} \exp\{\frac{-1}{2\tau_p^2} \sum_{i=0}^{p} \gamma_{pi}^2\}, \tag{10}$$

$$P(\sigma^2|\xi^{(p)}) = \begin{cases} \frac{1}{2\log(\tau_p^2)} \frac{1}{\sigma^2} & \text{if } \frac{1}{\tau_p^2} < \sigma^2 < \tau_p^2, \\ 0 & \text{otherwise,} \end{cases} \tag{11}$$

where τ_p is a hyperparameter to be determined later.

Multiplying (8), (9), (10) and (11), we have the following posterior distribution (up to a multiplicative constant),

$$\begin{aligned} & P(\xi^{(p)}, \boldsymbol{\gamma}_p, \sigma^2|\boldsymbol{Y}) \\ & \propto P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\gamma}_p, \sigma^2, \xi^{(p)})P(\boldsymbol{\gamma}_p|\xi^{(p)})P(\sigma^2|\xi^{(p)})P(\xi^{(p)}) \\ & = \mu^p(1-\mu)^{k-p}(2\pi)^{-(n+p+1)/2}(\sigma^2)^{-(n/2+1)}[2\log(\tau_p^2)]^{-1}(\tau_p^2)^{-(p+1)/2} \\ & \quad \exp\{-\frac{RSS_p}{2\sigma^2}\}\exp\{-\frac{1}{2\sigma^2}\sum_{i=0}^{p}(\gamma_{pi}-\hat{\gamma}_{pi})^2 - \frac{1}{2\tau_p^2}\sum_{i=0}^{p}\gamma_{pi}^2\}, \end{aligned} \tag{12}$$

where $RSS_p = \|\boldsymbol{Y} - \boldsymbol{X}_p\hat{\boldsymbol{\beta}}_p\|^2 = \|\boldsymbol{Y} - \boldsymbol{Z}_p\hat{\gamma}_p\|^2$ is the regression sum of squares, $\|\boldsymbol{Y} - \boldsymbol{Z}_p\boldsymbol{\gamma}_p\|^2 = RSS_p + \sum_{i=0}^{p}(\gamma_{pi} - \hat{\gamma}_{pi})^2$.

Integrating out $\boldsymbol{\gamma}_p$ from (12), we have

$$\begin{aligned} & P(\xi^{(p)}, \sigma^2|\boldsymbol{Y}) \\ & \propto \mu^p(1-\mu)^{k-p}(2\pi)^{-n/2}(\sigma^2)^{-(n/2+1)}[2\log(\tau_p^2)]^{-1}(\tau_p^2)^{-(p+1)/2}(\sigma^{-2}+\tau_p^{-2})^{-(p+1)/2} \\ & \quad \exp\{-\frac{1}{2\sigma^2}RSS_p - \frac{1}{2\sigma^2}\frac{\tau_p^{-2}}{\sigma^{-2}+\tau_p^{-2}}\sum_{i=0}^{p}\hat{\gamma}_{pi}^2\}. \end{aligned} \tag{13}$$

Taking the advantage of the flexibility of the prior setting, we assume that τ_p^2's are restricted so that

$$\log(\tau_p^2)(\tau_p^2)^{(p+1)/2} = \log(\tau_0^2)(\tau_0^2)^{1/2}. \tag{14}$$

Thus, we have

$$\begin{aligned} \frac{P(\xi^{(p)}|\boldsymbol{Y})}{P(\xi^{(0)}|\boldsymbol{Y})} = (\frac{\mu}{1-\mu})^p \int_{1/\tau_p^2}^{\tau_p^2} \frac{1}{(\sigma^2)^{n/2+1}} \frac{1}{(\sigma^{-2}+\tau_p^{-2})^{(p+1)/2}} \\ \exp\Big\{-\frac{RSS_p}{2\sigma^2} - \frac{1}{2\sigma^2}\frac{\tau_p^{-2}}{\sigma^{-2}+\tau_p^{-2}}\sum_{i=0}^{p}\hat{\gamma}_{pi}^2\Big\}d\sigma^2, \end{aligned} \tag{15}$$

where $\xi^{(0)}$ denotes the null model, the uniqueness of which allows us to regard $P(\xi^{(0)}|\boldsymbol{Y})$ as a constant in the derivation.

Let $\tau_0^2 \longrightarrow \infty$, the above ratio converges to

$$\lim_{\tau_0^2\to\infty}\frac{P(\xi^{(p)}|\boldsymbol{Y})}{P(\xi^{(0)}|\boldsymbol{Y})} = (\frac{\mu}{1-\mu})^p\Gamma(\frac{n-p-1}{2})(RSS_p/2)^{-(n-p-1)/2}. \tag{16}$$

Taking a logarithm, we have the log-posterior (up to an additive constant),

$$\begin{aligned} \log P(\xi^{(p)}|\boldsymbol{Y}) = p\log(\frac{\mu}{1-\mu}) + \frac{n-p-1}{2}\log 2 - \frac{n-p-1}{2}\log(RSS_p) \\ + \log\Gamma(\frac{n-p-1}{2}). \end{aligned} \tag{17}$$

In (17), the only parameter to be determined is μ.

By choosing μ appropriately, in the below theorem we show that the model selection procedure of maximizing the posterier (17) is approximately equivalent to minimizing Mallows' C_p (Mallows (1973)). For the subset model (6). We recall that Mallows' C_p is defined as $C_p = \frac{RSS_p}{\hat{\sigma}_k^2} + 2p' - n$, where RSS_p is the residual sum of squares from the p-variable subset model being considered, $\hat{\sigma}_k^2$ is estimated from the full model by $RSS_k/(n-k-1)$, $p' = p+1$ is the total number of predictors of the model including the intercept.

Theorem 2.1. *Under the prior setting* (9), (10) *and* (11), *when* $\mu = \mu_r$, *sampling from the posterior distribution* (17) *is approximately equivalent to sampling from the Boltzmann distribution*

$$f(\xi^{(p)}) \propto \exp\{-C_p(\xi^{(p)})/2\}, \tag{18}$$

where $C_p(\xi^{(p)})$ *denotes the* C_p *value of model* $\xi^{(p)}$. *Here* μ_r *is called a reference value of* μ *and is defined as* $\mu_r = 1/[1+\hat{\sigma}_k \exp(1+1/(2(n-1)))]$, *where* $\hat{\sigma}_k$ *is an estimator of* σ *from the full model with* $\hat{\sigma}_k = \sqrt{RSS_k/(n-k-1)}$.

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

$$\log \Gamma(\frac{n-p-1}{2}) \approx -\frac{n-p-1}{2} + \frac{n-p-2}{2} \log(\frac{n-p-1}{2}) + \frac{1}{2}\log(2\pi). \tag{19}$$

Substituting (19) into (20), we have

$$\log P(\xi^{(p)}|\boldsymbol{Y}) \approx p\log(\frac{\mu}{1-\mu}) + \frac{1}{2}\log(2\pi) - \frac{n-p-1}{2} - \frac{1}{2}\log(\frac{n-p-1}{2}) \\ -\frac{n-p-1}{2}\log(\hat{\sigma}_p^2), \tag{20}$$

where $\hat{\sigma}_p^2 = RSS_p/(n-p-1)$, and

$$\log P(\xi^{(p)}|\boldsymbol{Y}) - \log P(\xi^{(k)}|\boldsymbol{Y}) \approx \text{constant} + p\log(\frac{\mu}{1-\mu}) + \frac{p}{2} - \frac{1}{2}\log(1-\frac{p}{n-1}) \\ +\frac{p}{2}\log(\hat{\sigma}_k^2) - \frac{n-p-1}{2}\log(\hat{\sigma}_p^2/\hat{\sigma}_k^2), \tag{21}$$

where $\xi^{(k)}$ denotes the full model, and $\hat{\sigma}_k^2 = RSS_k/(n-k-1)$. The uniqueness of the full model allows us to regard $\log P(\xi^{(k)}|\boldsymbol{Y})$ as a constant in the derivation.

Writing $\log(\hat{\sigma}_p^2/\hat{\sigma}_k^2) = \log[1+(\hat{\sigma}_p^2-\hat{\sigma}_k^2)/\hat{\sigma}_k^2]$, when $(\hat{\sigma}_p^2-\hat{\sigma}_k^2)/\hat{\sigma}_k{}^2 \approx 0$, we have the following approximation to (21):

$$\log P(\xi^{(p)}|\boldsymbol{Y}) - \log P(\xi^{(k)}|\boldsymbol{Y}) \approx \text{constant} + p\log(\frac{\mu}{1-\mu}) + \frac{p}{2} + \frac{p}{2(n-1)} + \frac{p}{2}\log(\hat{\sigma}_k^2) \\ -\frac{n-p-1}{2}(\hat{\sigma}_p^2/\hat{\sigma}_k^2 - 1) \approx \text{constant} - \frac{C_p}{2} + (I), \tag{22}$$

where $(I) = p[1 + 0.5/(n-1) + \log(\mu/(1-\mu)) + \log(\hat{\sigma}_k)]$. In the derivation of (22), we assumed $n \gg p$ and made use of the approximation $\log(1-p/(n-1)) \approx -p/(n-1)$. Let $\mu = \mu_r$, we have $(I) = 0$. The proof is completed.

If we set $\mu = \mu_r$, which is computed from the data, the resulting posterior will have no parameter to be specified by users. So the prior setting (9), (10) and (11) is called an automatic prior setting, and the corresponding Bayesian model averaging procedure is called automatic Bayesian model averaging (ABMA). We realize that a simulation from the posterior distribution $P(\xi^{(p)}|\boldsymbol{Y})$ is only related with the relative magnitudes of the posterior probabilities of models, instead of the absolute magnitudes of them. Hence, to keep the simulation invariant to scale changes on the response variable, we only need to keep $P(\xi^{(p)}|\boldsymbol{Y})/P(\xi^{(k)}|\boldsymbol{Y})$ invariant to the scale changes for each model $\xi^{(p)}$. It is easy to see from equation (21) that the particular choice of μ_r counters the possible scale changes on the variance $\hat{\sigma}_k^2$, and thus keeps the simulation invariant to scale changes on the response variable. Clearly the simulation is also invariant to scale changes on the predictors. In fact, any choice of μ of the form $1/(1+\lambda\hat{\sigma}_k)$ will keep the simulation invariant to scale changes on the response variable and predictors, where λ is a penalty coefficient independent of p and $\hat{\sigma}_k$.

In addition to Bayesian model averaging, this theorem has a lot of implications for model selection. The minimum C_p model can be searched for by a MCMC simulation from the posterior distribution and the followed selection for the highest frequency model. The ergodicity of the Markov chain ensures that the minimum C_p model will be found almost surely as the running time tends to infinity (Tierney (1994)). The MCMC simulation also provides a pool of candidate models, from which we can select the models with $C_p \approx p+1$, as suggested by Mallows (1973), or the models with $C_p \leq p+1$, as suggested by Hocking (1976) and Mallows (1995).

2.2. Predictive performance

Since forecasting is a primary purpose of statistical data analysis (David (1984)), the predictive performance is a main assessment for a statistical approach or procedure. In this article, the predictive performance of ABMA is assessed using the three criteria.

The first criterion is the logarithmic score (LOGS) (Good (1952), Hoeting, Madigan, Raftery and Volinsky (1999)). It is defined as

$$-\frac{1}{|D^T|}\sum_{d\in D^T}\log\{\sum_{M\in\Omega}P(d|M,D)P(M|D)\},$$

where D^T denotes the test data set, and $|D^T|$ denotes the number of observations in D^T. The smaller the logarithmic score, the better the predictive performance.

Hoeting, Madigan, Raftery and Volinsky (1999) argued that the logarithmic score is a combined measure of the predictive bias (a systematic tendency to predict on the low or high side) and the lack of calibration (a systematic tendency to over- or understate predictive accuracy).

The second criterion is the mean squared prediction error (MSPE), which is also the most frequently used criterion in various circumstances. It is defined as

$$\frac{1}{|D^T|}\sum_{d\in D^T}\sum_{M\in\Omega}[\hat{d}(M,D)-d]^2 P(M|D),$$

where $\hat{d}(M,D)$ denotes the prediction value for the future observation d given the training data D and model M.

The third criterion is the mean absolute prediction error (MAPE). It is defined as

$$\frac{1}{|D^T|}\sum_{d\in D^T}\sum_{M\in\Omega}|\hat{d}(M,D)-d|P(M|D).$$

3. Computational Implementation

In this article, we use the evolutionary Monte Carlo (EMC) algorithm (Liang and Wong (2000)) to sample from the posterior distribution derived in Section 2. A new mutation operator is developed based on the specific structure of the problem as follows.

Suppose that we want to sample from a distribution $f(\xi) \propto \exp(-H(\xi))$, where $H(\cdot)$ denotes an energy function of ξ and it corresponds to -log-posterior in a simulation from a posterior distribution. In EMC, a sequence of distributions $f_1(\xi)$, $\ldots$, $f_N(\xi)$ are first constructed: $f_i(\xi) \propto \exp\{-H(\xi)/t_i\}$, $i = 1,\ldots,N$, where t_i is called a temperature of $f_i(\cdot)$. The temperature sequence $\boldsymbol{t} = (t_1,\ldots,t_N)$ forms a ladder with $t_1 > \cdots > t_N \equiv 1$. Issues related to the choice of the temperature ladder can be found in Liang and Wong (2000) and the references therein. Let ξ^i denote a sample from $f_i(\cdot)$, it is called an individual or a chromosome in genetic algorithms (Holland (1975), Goldberg (1989)). The N individuals $\xi^1,\ldots,\xi^N$ form a population denoted by $\boldsymbol{z} = \{\xi^1,\ldots,\xi^N\}$, N is called the population size. We assume that the individuals of the same population are mutually independent, and the Boltzmann distribution of the population is

$$f(\boldsymbol{z}) \propto \exp\{-\sum_{i=1}^{N} H(\xi^i)/t_i\}. \tag{23}$$

The population is updated by mutation, crossover and exchange operators (described below).

3.1. Mutation

In mutation a chromosome, say ξ^m, is chosen at random from the current population $\boldsymbol{z}$. Then ξ^m is mutated to a new chromosome $\xi^{m'}$ by some type of moves (described below). A new population $\boldsymbol{z}' = \{\xi^1, \ldots, \xi^{m'}, \ldots, \xi^N\}$, is accepted with probability min(1, r_m) according to the Metropolis-Hastings rule (Metropolis et al. (1953), Hastings (1970)), where

$$r_m = \frac{f(\boldsymbol{z}')}{f(\boldsymbol{z})}\frac{T(\boldsymbol{z}|\boldsymbol{z}')}{T(\boldsymbol{z}'|\boldsymbol{z})} = \exp\{-(H(\xi^{m'}) - H(\xi^m))/t_m\}\frac{T(\boldsymbol{z}|\boldsymbol{z}')}{T(\boldsymbol{z}'|\boldsymbol{z})}, \tag{24}$$

and $T(\cdot|\cdot)$ denotes the transition probability between two populations. Otherwise, $\boldsymbol{z}$ is unchanged.

In this article, the mutation operator incorporates the moves of reversible jump MCMC (Green (1995)): the "birth", "death" and "simultaneous" moves. Let S denote the set of predictors of the current model, S^c the complementary set of S, and $p = \|S\|$ the number of predictors in S. In the "birth" step, a predictor is uniformly chosen from S^c and is proposed to be added to the model; in the "death" step, a predictor is uniformly chosen from S and is proposed to be deleted from the model; "simultaneous" move means that the "birth" and "death" steps are performed simultaneously (in this step a predictor, say $\boldsymbol{x}_c$, is uniformly chosen from S and another, say $\boldsymbol{x}_c^*$, is uniformly chosen from S^c, it is proposed to replace $\boldsymbol{x}_c$ by $\boldsymbol{x}_c^*$). Let $P(p, \text{birth})$, $P(p, \text{death})$ and $P(p, \text{simultaneous})$ denote the proposal probabilities of the three types of moves for a model with p predictors, respectively. In our examples, we set $P(p, \text{birth}) = P(p, \text{death}) = P(p, \text{simultaneous}) = 1/3$ for $1 < p < k$, and $P(k, \text{death}) = P(0, \text{birth}) = 1$. The ratios of the transition probabilities are as follows. For the "birth" step, we have

$$\frac{T(\boldsymbol{z}|\boldsymbol{z}')}{T(\boldsymbol{z}'|\boldsymbol{z})} = \frac{P(p+1, \text{death})}{P(p, \text{birth})}\frac{k-p}{p+1}.$$

For the "death" step, we have

$$\frac{T(\boldsymbol{z}|\boldsymbol{z}')}{T(\boldsymbol{z}'|\boldsymbol{z})} = \frac{P(p-1, \text{birth})}{P(p, \text{death})}\frac{p}{k-p+1}.$$

For the "simultaneous" step, we have $T(\boldsymbol{z}|\boldsymbol{z}')/T(\boldsymbol{z}'|\boldsymbol{z}) = 1$, since $T(\boldsymbol{z}|\boldsymbol{z}') = T(\boldsymbol{z}'|\boldsymbol{z}) = 1/[p(k-p)]$.

3.2. Crossover

In crossover, different offspring are produced by a recombination of parental chromosomes selected from the current population according to a so called roulette wheel selection procedure. In this procedure, one chromosome is selected with probability proportional to its Boltzmann probability

$$p(\xi^i) \propto \exp\{-H(\xi^i)/t_s\}, \tag{25}$$

where t_s is called the selection temperature. Here we set $t_s \equiv t_N$, although this is not necessary. A second chromosome is uniformly selected from the remainder of the population so that

$$P(\xi^i, \xi^j|\boldsymbol{z}) = \frac{1}{(N-1)C(X)}[\exp\{-H(\xi^i)/t_s\} + \exp\{-H(\xi^j)/t_s\}], \tag{26}$$

where $C(X) = \sum_{i=1}^N \exp\{-H(\xi^i)/t_s\}$.

Now ξ^i and ξ^j are parental chromosomes, new offspring $\xi^{i'}$ and $\xi^{j'}$ are generated as follows. First, an integer c is drawn uniformly on $\{1, 2, \ldots, k\}$, then $\xi^{i'}$ and $\xi^{j'}$ are constructed by swapping the genes to the right of the crossover point between the two parental chromosomes. The following diagram illustrates the 1-point crossover operator,

$$\begin{array}{ccc} (\xi_1^i, \ldots, \xi_k^i) & & (\xi_1^i, \ldots, \xi_c^i, \xi_{c+1}^j, \ldots, \xi_k^j) \\ & \Longrightarrow & \\ (\xi_1^j, \ldots, \xi_k^j) & & (\xi_1^j, \ldots, \xi_c^j, \xi_{c+1}^i, \ldots, \xi_k^i), \end{array}$$

where c is called a crossover point. If there are k crossover points, the operator is called the k-point crossover. Only the 1-point crossover operator is used here.

A new population is constructed by replacing the parental chromosomes with the new "offspring", and it is accepted with probability min(1,r_c) according to the Metropolis-Hastings rule, with

$$r_c = \frac{f(\boldsymbol{z}')}{f(\boldsymbol{z})}\frac{T(\boldsymbol{z}|\boldsymbol{z}')}{T(\boldsymbol{z}'|\boldsymbol{z})} = \exp\{-(H(\xi^{i'}) - H(\xi^i))/t_i - (H(\xi^{j'}) - H(\xi^j))/t_j\}\frac{T(\boldsymbol{z}|\boldsymbol{z}')}{T(\boldsymbol{z}'|\boldsymbol{z})}, \tag{27}$$

where $T(\boldsymbol{z}'|\boldsymbol{z}) = P(\xi^i, \xi^j|\boldsymbol{z})\ P(\xi^{i'}, \xi^{j'}|\ \xi^i, \xi^j)$, $P(\xi^i, \xi^j|\boldsymbol{z})$ denotes the selection probability of (ξ^i, ξ^j) from the population $\boldsymbol{z}$, $P(\xi^{i'}, \xi^{j'} \mid \xi^i, \xi^j)$ denotes the generating probability of $(\xi^{i'}, \xi^{j'})$ from the parental chromosomes (ξ^i, ξ^j). Since the k-point crossover operator is symmetric in the sense $P(\xi^{i'},\ \xi^{j'}|\xi^i, \xi^j) = P(\xi^i,\ \xi^j|\xi^{i'},\ \xi^{j'})$, the ratio of the transition probabilities in (27) is reduced to the ratio of selection probabilities.

3.3. Exchange

Given the current population $\boldsymbol{z}$ and the attached temperature ladder $\boldsymbol{t}$, we try to make an exchange between ξ^i and ξ^j without changing the t's, i.e., initially we have $(\boldsymbol{z}, \boldsymbol{t}) = (\xi^1, t_1, \ldots, \xi^i, t_i,\ \ldots,\ \xi^j, t_j, \ldots, \xi^N, t_N)$ and we want to change it to $(\boldsymbol{z}', \boldsymbol{t}) = (\xi^1, t_1, \ldots, \xi^j, t_i, \ldots,\ \xi^i, t_j, \ldots,\ \xi^N, t_N)$. The new population is accepted with probability min(1,r_e) according to the Metropolis rule, where

$$r_e = \frac{f(\boldsymbol{z}')}{f(\boldsymbol{z})} \frac{T(\boldsymbol{z}|\boldsymbol{z}')}{T(\boldsymbol{z}'|\boldsymbol{z})} = \exp\{(H(\xi^i) - H(\xi^j))(\frac{1}{t_i} - \frac{1}{t_j})\} \frac{T(\boldsymbol{z}|\boldsymbol{z}')}{T(\boldsymbol{z}'|\boldsymbol{z})}. \tag{28}$$

Typically, the exchange is only performed on two states with neighboring temperatures, i.e., $|i-j| = 1$. If $p(\xi^i)$ is the probability that ξ^i is chosen to exchange with the other state, and $w_{i,j}$ is the probability that ξ^j is chosen to exchange with ξ^i, we have $T(\boldsymbol{z}'|\boldsymbol{z}) = p(\xi^i)w_{i,j} + p(\xi^j)w_{j,i}$. Thus $T(\boldsymbol{z}'|\boldsymbol{z}) = T(\boldsymbol{z}|\boldsymbol{z}')$, and these factors cancel in (28).

3.4. Algorithm

With the operators described above, one iteration of EMC consists of the following two steps.

- Apply the mutation or the crossover operator to the population with probability q and $1-q$ respectively, q is the mutation rate.
- Try to exchange ξ^i with ξ^j for $N-1$ pairs (i,j) with i being sampled uniformly on $\{1, \ldots, N\}$ and $j = i \pm 1$ with probability $w_{i,j}$, where $w_{i,i+1} = w_{i-1,i} = 0.5$ and $w_{1,2} = w_{N,N-1} = 1$.

EMC differs from other MCMC algorithms in two respects. First, the algorithm incorporates the learning ability of genetic algorithms (Holland (1975), Goldberg (1989)) by evolving with crossover operators. The crossover operator tends to preserve the good genes of the population and they work as a guideline for further iterations. Second, the algorithm incorporates the exploring ability of simulated annealing (Kirkpatrick, Gelatt and Vecchi (1983)) by simulating a sequence of distributions along a temperature ladder. Simulation at a high temperature provides more chances for the system to escape from local minima. In addition, a large dimension jumping in the model space is allowed in the crossover operator.

The structure of EMC is also very flexible. If $q = 1$, i.e., only the mutation operator is performed, EMC reduces to parallel tempering (Geyer (1991), Hukushima and Nemoto (1996)). If $q = 1$ and $N = 1$, EMC reduces to the usual single-chain MCMC algorithm. The efficient operators developed in MCMC, e.g., the Gibbs sampler (Geman and Geman (1984)) and reversible jumps (Green (1995)), can be incorporated as a mutation operator by EMC.

4. Numerical Examples

4.1. Crime data

It is thought that criminal activities are outcomes of rational economic decision processes, and the probability of punishment acts as a deterrent for them.

Ehrlich (1973) developed this argument theoretically and tested it empirically using aggregate data from 47 U.S. states in 1960. Later, errors in the data set were corrected by Vandaele (1978). The corrected data set has been used by Raftery, Madigan and Hoeting (1997) and Fernández, Ley and Steel (2001) as an illustrative example for their Bayesian model averaging procedures. For convenience, hereafter, the latter two procedures will be referred to as RMH and FLS, respectively. We also use the corrected data set as an illustrative example and compare with the RMH and FLS procedures.

Consider a linear regression model as in (6), where the response variable, y, group observations on the crime rate, and the 15 potential predictors are given in Table 1. As in RMH and FLS, we transform all variables to logarithms except for one dummy variable, the indicator variable for southern states.

ABMA was applied to the data with $\mu = \mu_r$. EMC was used to sample from the posterior distribution with the following parameter setting: the population size $N = 20$, the temperature ladder $\boldsymbol{t} = \{t_1, \ldots, t_N\}$ is equally spaced between 5 and 1, and the mutation rate $q = 0.5$. Figure 1(a) is a barplot which shows the true Boltzmann distribution of C_p as defined in (18). Figure 1(b) is the histogram of C_p of the models sampled in one run. The sample size is 50,000. The similarity of these two plots shows empirically the approximate equivalence between sampling from the Boltzmann distribution of C_p and sampling from the posterior distribution with $\mu = \mu_r$. The result of Theorem 2.1 is confirmed.

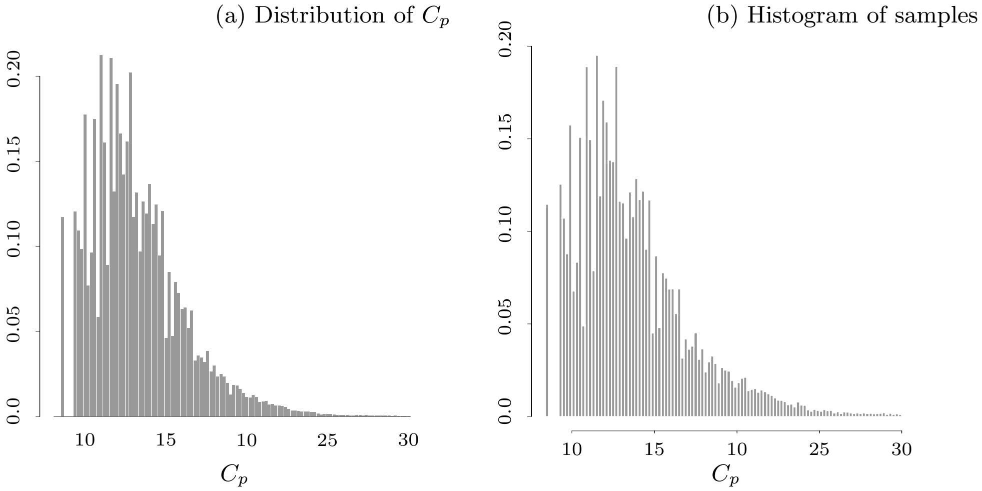

Figure 1. A comparison of the true Boltzmann distribution defined on C_p (a) and estimated by ABMA (b) for the crime data.

Let $P(\beta_i \neq 0|D)$ denote the posterior probability that predictor $\boldsymbol{x}_i$ is included in the true model. Table 1 compares the estimates of $P(\beta_i \neq 0|D)$'s obtained by ABMA, RMH and FLS. For comparison, the $1-p$-values obtained by an ordinary least squares (OLS) regression on a full model are also given in the table.

In RMH, the authors imposed an uniform prior distribution on the model space, $P(M_i) \propto 2^{-k}, \quad i = 1, \ldots, 2^k$. For parameters $\boldsymbol{\beta}_p$ and σ^2 associated with a model with p predictors, they assumed that $\boldsymbol{\beta}_p \sim N_{p+1}(\boldsymbol{\psi}, \sigma^2 V)$, $\nu\lambda/\sigma^2 \sim \chi^2_\nu$, where ν, λ, $\boldsymbol{\psi}$ and V are hyperparameters to be specified *a priori*. Typically, they set $\boldsymbol{\psi} = (\hat{\beta}_0, 0, 0, \ldots, 0)$, where $\hat{\beta}_0$ is the OLS estimate of β_0; $V = \text{diag}[s_Y^2, \phi^2 s_1^{-2}, \ldots, \phi^2 s_p^{-2}]$, where s_Y^2 is the sample variance of Y, s_i^2 is the sample variance of $\boldsymbol{x}_i$ for $i = 1, \ldots, p$, and $s_i^2 = n(\boldsymbol{x}_i'\boldsymbol{x}_i)^{-1}$; $\nu = 2.58$, $\lambda = 0.28$ and $\phi = 2.85$. The posterior distribution is sampled using the Markov chain Monte Carlo model composition (MC^3) method (Madigan and York (1995)). The program is available at `http://lib.stat.cmu.edu/S/bma`.

In FLS, the authors imposed an uniform prior distribution on the model space as in RMH, $P(M_i) \propto 2^{-k}, i = 1, \ldots, 2^k$. For $\boldsymbol{\beta}_p$ and σ^2, they assume that $P(\sigma) \propto \sigma^{-1}$, $P(\beta_0) \propto 1$, $P(\beta_1, \ldots \beta_p) \sim N(0, \sigma^2(g_0 X'X)^{-1})$, where X is the design matrix but excluding the intercept column, and g_0 is a hyperparameter. Typically, for the crime data, they set $g_0 = 1/k^2$, which is also the value suggested by the Risk Inflation Criterion (RIC) of Foster and George (1994). The posterior distribution is sampled by using MC^3. The program is available at `http://www.research.att.com/~volinsky/bma.html`.

Table 1. The estimated $P(\beta_i \neq 0|D)$'s for the crime data: OLS: $1-p$-values obtained by an ordinary least squares regression on a full model; ABMA: the estimates obtained by ABMA, where we set $\mu = \mu_r$; RMH: the estimates obtained by RMH; FLS: the estimates obtained by FLS.

No.	Predictor	OLS	ABMA	RMH	FLS
1	Percentage of males age 14-24	0.996	0.935	0.79	0.758
2	Indicator variable for southern state	0.358	0.378	0.17	0.141
3	Mean years of schooling	0.999	0.992	0.98	0.955
4	Police expenditure in 1960	0.700	0.701	0.72	0.658
5	Police expenditure in 1959	0.041	0.497	0.50	0.381
6	Labor force participation rate	0.611	0.329	0.06	0.075
7	Number of males per 1,000 females	0.794	0.345	0.07	0.085
8	State population	0.870	0.511	0.23	0.222
9	Number of nonwhites per 1,000 people	0.976	0.835	0.62	0.507
10	Unemployment rate of urban males age 14-24	0.309	0.388	0.11	0.106
11	Unemployment rate of urban males age 35-39	0.940	0.766	0.45	0.451
12	Wealth	0.894	0.542	0.30	0.175
13	Income inequality	1.000	0.999	1.00	0.998
14	Probability of imprisonment	0.996	0.961	0.83	0.789
15	Average time served in state prisons	0.867	0.551	0.22	0.180

Table 1 shows that ABMA performs quite differently from RMH and FLS in estimating $P(\beta_i \neq 0|D)$'s for some predictors, although not for others. ABMA performs more like OLS than RMH and FLS. Actually, the ABMA estimates can be regarded as a compromise between the OLS estimates and RMH or FLS estimates. Note that the latter two estimates are very similar since they have used similar prior distributions for the regression coefficients, both related to the g-prior of Zellner (1986). If we regard OLS as the Bayesian estimates with the noninformative prior, we can conclude that the priors used in RMH and FLS are more informative than the automatic prior setting used by ABMA. One question then is which of these settings is better? This question is partially answered in the following sections by comparing the predictive performances of these procedures on a variety of examples.

Table 2 shows the ten highest posterior models of ABMA. For comparison, their C_p values are also given in the table. The best ten models have covered the minimum C_p, AIC, BIC and PRESS models and the maximum adjR^2 model. However, they have little connection with the best ten models of RMH and FLS. Only the second model in Table 2 appeared in the best ten models of RMH and FLS. Note that RMH and FLS perform similarly in this example, eight models are shared among their ten best.

Table 2. The ten highest posterior models resulting from the automatic prior setting with $\mu = \mu_r$ for the crime data. The probability is expressed as a percentage. The underlined value presents the minimum value of the corresponding criterion statistic. a: the minimum BIC model, it is also the only model (among the ten models) which appears in the best ten models of RMH and FLS. In RMH and FLS, it is ranked 5 and 9, respectively. b: the minimum AIC and the maximum adjR^2 model. c: the minimum PRESS model.

No.	Prob.(%)	C_p	Included predictors												
1	1.066	$\underline{8.504}$	1	3	4				9		11	12	13	14	15
2^a	1.026	8.547	1	3	4				9		11		13	14	15
3	0.734	9.268	1	3	4				9	10	11	12	13	14	15
4	0.731	9.236	1	3	4				9	10	11		13	14	15
5	0.695	9.334	1	3	4			8	9		11	12	13	14	
6^b	0.688	9.458	1	3	4		7	8	9		11	12	13	14	15
7	0.671	9.403	1	3		5			9		11	12	13	14	15
8	0.623	9.581	1	3	4			8	9		11	12	13	14	15
9^c	0.605	9.605	1	3	4			8	9		11		13	14	
10	0.587	9.697	1	3	4		7	8	9		11	12	13	14	

To assess the predictive performance of ABMA, we split the data into two parts. The first 35 observations (about 75% of the total observations) are used for model building, the training data, the remaining 12 observations (about 25% of the total observations) are used for model testing, the test data.

First we run EMC on the training data with the same EMC setting as in the last section. Note here the μ_r value has changed, it is re-computed from the first 35 observations. We run EMC ten times independently. Each run generated 10000 samples after the first 5000 burn-in steps. The LOGS, MSPE and MAPE were computed on the test data. The results are summarized in Table 3. For comparison, RMH and FLS were also run with the parameter settings used by the authors. In FLS, each run produced 10^6 samples after the first 2.5×10^4 burn-in steps. In RMH, each run produced 3×10^4 samples. Since RMH is coded in S-PLUS and the running speed is very slow, it was only run one time. The results are reliable enough for comparison. For completeness, corresponding results of the minimum C_p criterion are also given in the table. The table shows that ABMA is superior to the other procedures in LOGS and MSPE. In MAPE, the predictive performances of all four procedures are nearly the same, it is less sensitive than the other two criteria. Note that although ABMA is intimately related to the C_p criterion, its predictive performance outperforms the minimum C_p model as expected, since it has accounted for the uncertainty of the model.

Table 3. The comparison of the predictive performance of ABMA, RMH, FLS, and the minimum C_p criterion for the crime data. The number in the parenthesis denotes the standard deviation of the preceding number.

Method	LOGS	MSPE	MAPE
ABMA	$0.258(1 \times 10^{-3})$	$0.094(2 \times 10^{-4})$	$0.241(2 \times 10^{-4})$
FLS	$0.524(5 \times 10^{-5})$	$0.111(2 \times 10^{-6})$	$0.240(3 \times 10^{-6})$
RMH	0.487	0.111	0.242
Mallows C_p	0.553	0.099	0.242

4.2. A simulated example

The following example is modified from one in George and McCulloch (1993). Generate $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_5$ i.i.d. $\sim N_{50}(0,1)$ and $\boldsymbol{\epsilon} \sim N_{50}(0, 2.5^2)$, set $\boldsymbol{x}_1 = 0.5\boldsymbol{z}_1 + \boldsymbol{z}_4$, $\boldsymbol{x}_2 = 0.15\boldsymbol{z}_2 + \boldsymbol{z}_5$, $\boldsymbol{x}_i = \boldsymbol{z}_i$ for $i = 3, 4, 5$, and $\boldsymbol{Y} = \boldsymbol{x}_4 + 1.2\boldsymbol{x}_5 + \boldsymbol{\epsilon}$. Let $\boldsymbol{Y} = [\boldsymbol{1}, \boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_5]\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\beta} = (0, 0, 0, 0, 1, 1.2)$. Twenty data sets were generated independently, the average value of the correlation coefficients was 0.89 between $\boldsymbol{x}_1$ and $\boldsymbol{x}_4$ and 0.99 between $\boldsymbol{x}_2$ and $\boldsymbol{x}_5$. This example illustrates the performance of ABMA in the presence of strong collinearity or model uncertainty. For the problems with little model uncertainty, RMH has shown that the predictive performance is not significantly improved by model averaging.

In this example, the first ten data sets were used for training, and the second ten data sets were used for testing. To get a more extensive test, the models sampled/selected for one training set were tested on all ten test sets. Thus, totally 100 values were computed for each predictive assessment statistic. These values can be regarded approximately as *i.i.d.* observations.

For each training set EMC was run for 6000 iterations with the same setting as in the last example. The first 1000 iterations were discarded for the burn-in process. Since the variations of the estimated predictive assessment statistics are very small in different runs for this example, EMC was only run one time for each training set. The results are summarized in Table 4 and Figure 2. For comparison, RMH was also run one time for each training set, where each run consists of 3000 iterations. The results show that the predictive performance of ABMA is significantly better than that of RMH and the minimum C_p criterion.

Table 4. A comparison of the predictive performance of ABMA, the minimum C_p criterion and RMH for the simulated data, where "Mean" denotes the average difference of the two methods, "SD" denotes the standard deviation of the "Mean" value.

Criterion	ABMA-C_p		ABMA-RMH	
	Mean($\times$100)	SD($\times$100)	Mean($\times$100)	SD($\times$100)
LOGS	-0.92	0.28	-0.98	0.17
MSPE	-16.63	4.19	-12.97	2.62
MAPE	-2.40	0.73	-1.71	0.49

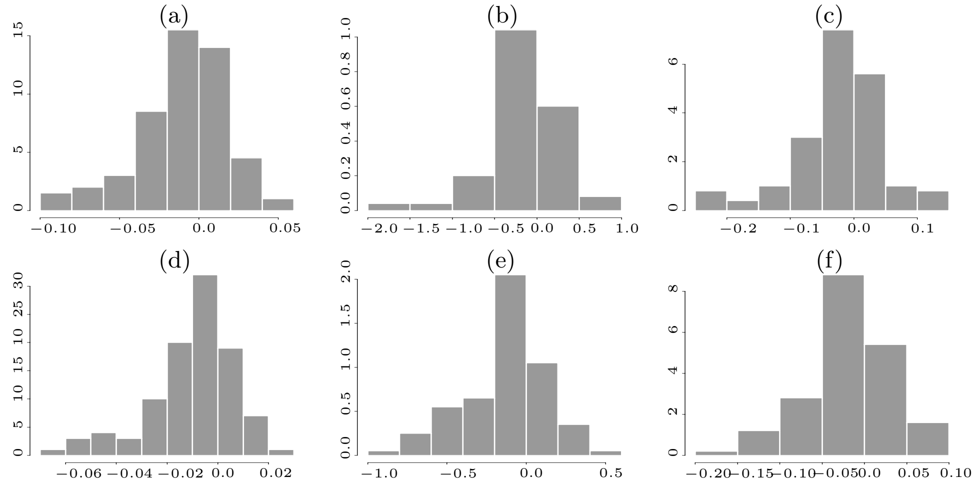

Figure 2. A comparison of the predictive performance of ABMA, the minimum C_p criterion and RMH for the simulated data. The histograms (a)-(c) shows the difference between ABMA and the minimum C_p criterion: (a) LOGS; (b) MSPE; (c) MAPE. The histograms (d)-(f) shows the difference between ABMA and RMH: (d) LOGS; (e) MSPE; (f) MAPE.

4.3. Sensitivity analysis

To assess the influence of the value of μ on the model selection and predictive performance, we consider the following example. Generate $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{15} \sim N_{200}(0,1)$, and $\boldsymbol{\epsilon} \sim N_{200}(0, 2.5^2)$. Set $\boldsymbol{x}_1 = 0.1\boldsymbol{z}_1 + \boldsymbol{z}_6$, $\boldsymbol{x}_2 = 0.2\boldsymbol{z}_2 + \boldsymbol{z}_7$, $\boldsymbol{x}_3 = 0.3\boldsymbol{z}_3 + \boldsymbol{z}_8$, $\boldsymbol{x}_4 = 0.4\boldsymbol{z}_4 + \boldsymbol{z}_9$, $\boldsymbol{x}_5 = 0.5\boldsymbol{z}_5 + \boldsymbol{z}_{10}$, $\boldsymbol{x}_i = \boldsymbol{z}_i$ for $i = 6, \ldots, 15$, and $\boldsymbol{y} = 0.5(\boldsymbol{x}_6 + \boldsymbol{x}_7 + \boldsymbol{x}_8 + \boldsymbol{x}_9 + \boldsymbol{x}_{10}) + \boldsymbol{\epsilon}$. The first 100 observations are used as the training data and the second 100 observations are used as the test data. An OLS regression of $\boldsymbol{y}$ on all predictors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{15}$ produces a multiple R^2 of 0.22 and least squares estimates as in Table 5. No predictor is significant at the 0.1 level. The low R^2 value shows that the linear relationship between $\boldsymbol{y}$ and predictors is very weak. Thus, the value of μ will have more influence on the model selection and the resulting predictive performance.

Table 5. The ordinary least squares estimates for the test example. The last row shows the p-values of the corresponding predictors.

	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	β_{11}	β_{12}	β_{13}	β_{14}	β_{15}
β	0	0	0	0	0	0	0.5	0.5	0.5	0.5	0.5	0	0	0	0	0
$\hat{\beta}$	-0.19	-2.20	0.86	-0.83	-0.68	0.96	2.34	-0.06	1.22	0.34	-0.30	-0.01	0.03	-0.23	-0.30	-0.36
p	0.52	0.44	0.54	0.36	0.92	0.13	0.42	0.97	0.20	0.65	0.67	0.98	0.93	0.53	0.31	0.23

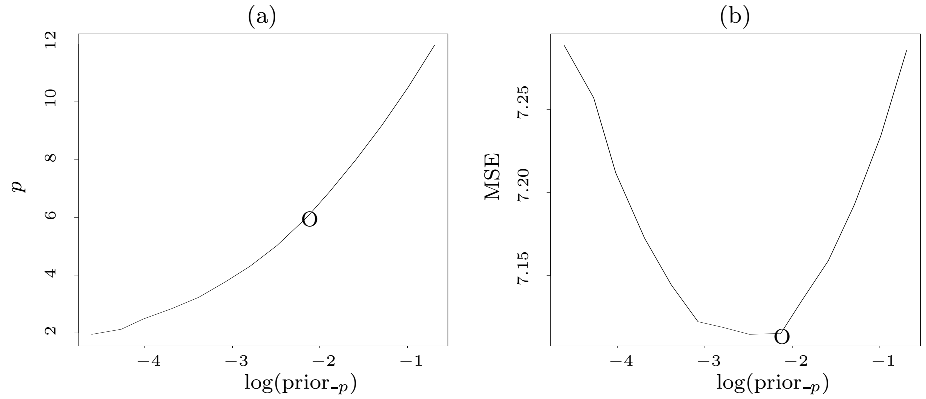

Figure 3. The assessment of the influence of μ on model selection for the test example. The circle corresponds to the reference value μ_r. The x-axis is plotted in the logarithm scale. (a) The Bayesian estimate of the number of predictors in the regression. (b) The Bayesian estimate of the mean squared error (MSE).

ABMA is applied to the training data with 15 different values of μ: 0.01, 0.014, 0.018, 0.025, 0.034, 0.046, 0.061, 0.083, 0.112, 0.1185, 0.151, 0.204, 0.275 0.372 and 0.5 where, aside from $\mu_r = 0.1185$, values are roughly equally spaced in [0.01,0.5] in logarithm. For each value of μ, EMC was run for 10000 iterations after the first 5000 burn-in steps with the same parameter setting as in the crime data example. Figure 3 assesses the influence of μ on model selection by showing the curves of the estimated number of predictors and the estimated mean squared error (MSE) versus $\log(\mu)$. The curves are obtained by averaging over 10000 samples for each value of μ. The estimated number of predictors increases as μ increases. However, the curve of the estimated MSE is U-shaped, and it attains its minimum around μ_r. Figure 4 assesses the influence of μ on the predictive performance by showing the curves of LOGS, MSPE, and MAPE versus $\log(\mu)$. The curves are obtained by averaging over 10000 samples for each value of μ. The three curves are all U-shaped, and they attain their minima around μ_r. The curves of LOGS and MSPE are similar, and are more sensitive to the values of μ than that of MAPE. Note that even for this special example, there is still a small neighborhood (in the logarithm scale) around μ_r, where ABMA has a quite stable performance. As shown in Figures 3 and 4, a value of μ slightly less than μ_r often results in a good predictive performance. These results give further assurance that the reference value μ_r is a reasonable choice of μ for Bayesian model averaging.

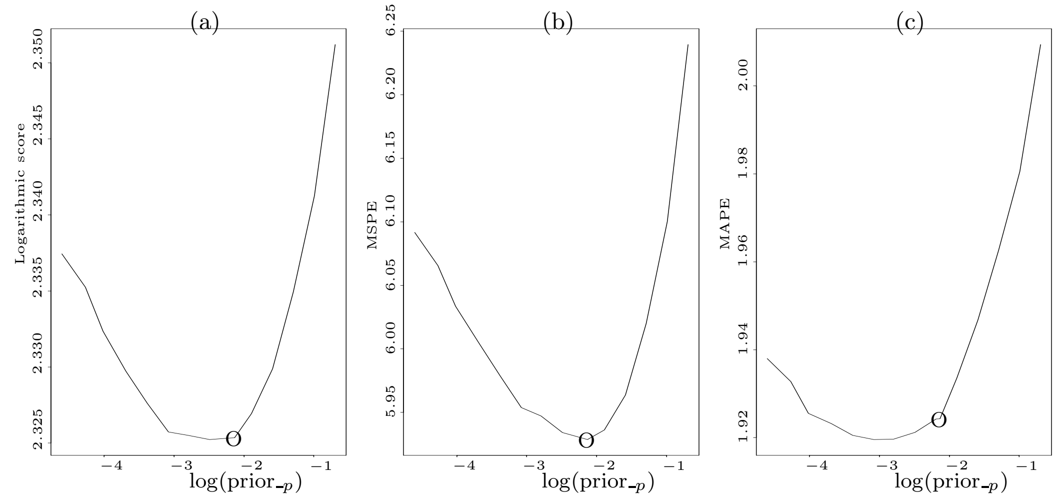

Figure 4. The assessment of the influence of μ on predictive performance for the test example. The circle corresponds to the reference value μ_r. The x-axis is plotted in the logarithm scale. (a) LOGS. (b) MSPE. (C) MAPE.

5. Bayesian Curve Fitting with Least Squares Splines

Suppose we are given a nonparametric regression model where observations (x_i, y_i), $i = 1, \ldots, n$, satisfy the equations

$$y_i = f(x_i) + \epsilon_i, \tag{29}$$

where $\epsilon_i \sim N(0,\sigma^2)$, $a \le x_1 \le \cdots \le x_n \le b$, and $f \in C_2^m[a,b]$. The least squares spline technique is to approximate the unknown function $f(\cdot)$ with a regression function of the form

$$S(x) = \sum_{j=1}^{m} \alpha_j x^{j-1} + \sum_{j=1}^{p} \beta_j (x - t_j)_+^{m-1}, \tag{30}$$

for $x \in [a,b]$, where $z_+ = \max(0,z)$, p is the number of knots, the t_i's ($i = 1,\ldots,p$) denote knot locations, and the α's and β's are regression coefficients.

The curve fitting problem has been considered by many authors. Related works include those on smoothing splines, kernel smoothers (Kohn and Ansley (1987), Kohn, Ansley and Tharm (1991), Hastie and Tibshirani (1990), Eubank (1999) and references therein), stepwise knot replacement (Friedman and Silverman (1989), Friedman (1991), Kooperberg and Stone (1992), Hansen, Kooperberg and Sardy (1998), Hansen and Kooperberg (2000)) and variable bandwidth kernel methods (Müller and Stadtmüller (1987), Fan and Gijbels (1995)). Recently, Denison, Mallick and Smith (1998) proposed a Bayesian approach implemented with a hybrid sampler. In this article, we provide a fully Bayesian approach for the problem.

We use a modified least squares spline,

$$S^*(x) = \sum_{l=1}^{m} \beta_{l,0} (x - t_0)^{l-1} + \sum_{j=1}^{p} \sum_{l=m_0}^{m} \beta_{l,j} (x - t_j)_+^{l-1}, \tag{31}$$

where $t_0 = x_1$. The modification reduces the continuity constraints of (30), and the resulting spline (31) is more flexible. Replacing the function $f(x)$ in (29) by $S^*(x)$, we have the following regression equation,

$$y_i = \sum_{l=1}^{m} \beta_{l,0} (x_i - t_0)^{l-1} + \sum_{j=1}^{p} \sum_{l=m_0}^{m} \beta_{l,j} (x_i - t_j)_+^{l-1} + \epsilon_i, \tag{32}$$

where $\epsilon_i \sim N(0,\sigma^2)$ for $i = 1,\ldots,n$. We assume that the possible knot locations are the n regular points on $[a,b]$.

In a matrix form, (32) can be written as,

$$\boldsymbol{Y} = \boldsymbol{X}_p \boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{33}$$

where $\boldsymbol{Y}$ is an n-vector of observation, $\boldsymbol{\epsilon} \sim N(0,\sigma^2 I)$, $\boldsymbol{X}_p = [\boldsymbol{1}, (\boldsymbol{x}-t_0)_+^1, \ldots, (\boldsymbol{x}-t_0)_+^{m-1}, (\boldsymbol{x}-t_1)_+^{m_0-1}, \ldots, (\boldsymbol{x}-t_p)_+^{m-1}]$ and $\boldsymbol{\beta} = (\beta_{1,0}, \ldots, \beta_{m,0}, \beta_{m_0,1}, \cdots, \beta_{m,p})$. Here $\boldsymbol{\beta}$ includes $m + (m - m_0 + 1)p$ individual parameters. Note that in this problem, the number of potential predictors can be larger than the number of observations and the full model is not well defined, Theorem 2.1 is not applicable. But we found that the "automatic" prior setting still provides a simple treatment for the problem. With the "automatic" prior setting, we have the following log-posterior (up to an additive constant),

$$\begin{aligned}\log P(\xi^{(p)}|\boldsymbol{Y}) = p\log(\frac{\mu}{1-\mu}) + \frac{n-wp-m}{2}\log 2 \\ -\frac{n-wp-m}{2}\log(\boldsymbol{Y}'\boldsymbol{Y} - \boldsymbol{Y}'\boldsymbol{X}_p(\boldsymbol{X}_p'\boldsymbol{X}_p)^{-1}\boldsymbol{X}_p'\boldsymbol{Y}) \\ +\log\Gamma(\frac{n-wp-m}{2}),\end{aligned} \tag{34}$$

where $w = (m - m_0 + 1)$ denotes the number of terms after adding one more knot to the regression. Note in (9), k equals to n, the number of possible knots. The Bayesian estimator of $f(x)$ is

$$\hat{f}(x) = \sum_{i=0}^{K} \hat{f}_i(x)P(M_i|\boldsymbol{Y}), \tag{35}$$

where K denotes the number of all models under consideration, $\hat{f}_i(x) = \boldsymbol{X}_i(\boldsymbol{X}_i' \boldsymbol{X}_i)^{-1}\boldsymbol{X}_i'\boldsymbol{Y}$, and $\boldsymbol{X}_i$ is the design matrix corresponding to model M_i. Given the samples $M_1, \ldots, M_t, \ldots$ sampled from the posterior distribution (34), under the ergodicity of the sampler, $f(x)$ can then be estimated according to (5).

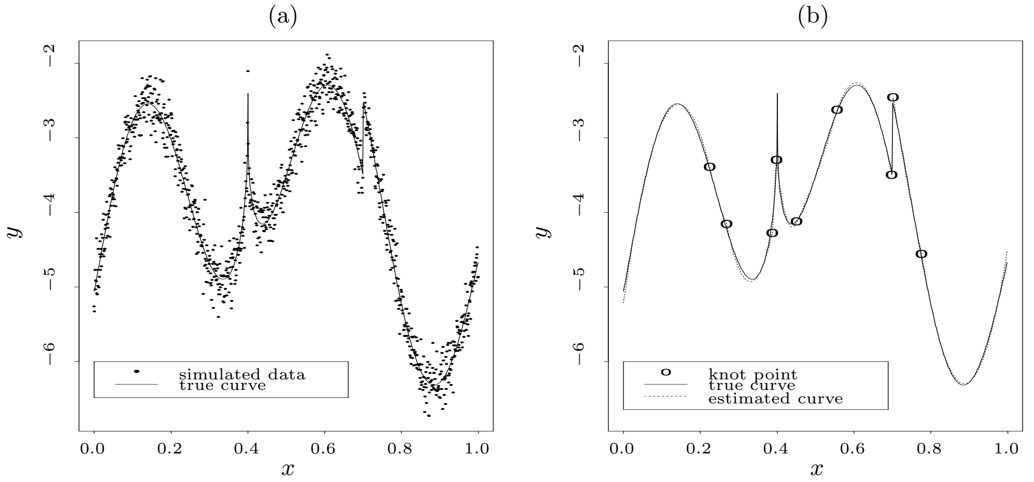

Figure 5. (a)The Simulated data and the true regression curve. (b) The MAP estimate of the knot points and the regression curve in one run of ABMA with $\mu = 0.01$.

This approach was tested on the following example. The regression function is given by $f(x) = 2\sin(4\pi x) - 6|x - 0.4|^{0.3} - 0.5\text{sign}(0.7 - x), x \in [0, 1]$. The data are equally spaced between 0 and 1 with sample size $n = 1000$, the SD of the Gaussian noise is $\sigma = 0.2$. Figure 5(a) plots the simulated data and the true regression curve. This function has a narrow spike at 0.4 and a jump at 0.7, and the approximation to it is a challenge. This example has been analyzed by several authors using regression spline estimation, see Wang (1995) and Koo (1997).

EMC was applied to simulate from the posterior distribution. We set $m_0 = 2$ and $m = 3$, and the resulting $\hat{f}(x)$ is a continuous piecewise quadratic polynomial. EMC was run for 1000 iterations, the first 500 iterations were discarded for the burn-in process. Figure 5(b) shows the maximum *a posteriori* (MAP) estimate of the knot points and the regression curve obtained in one run of EMC with $\mu = 0.01$. Figure 6(a) and (b) show two Bayesian estimates of the regression curve. They are obtained in two runs with $\mu = 0.01$ and $\mu = 0.015$ respectively. The respective CPU times were 7.5m and 10.2m on an Alpha-500 workstation. Figure 6(a) and (b) show that the underlying regression function has been well approximated by ABMA, including the spike and the jump.

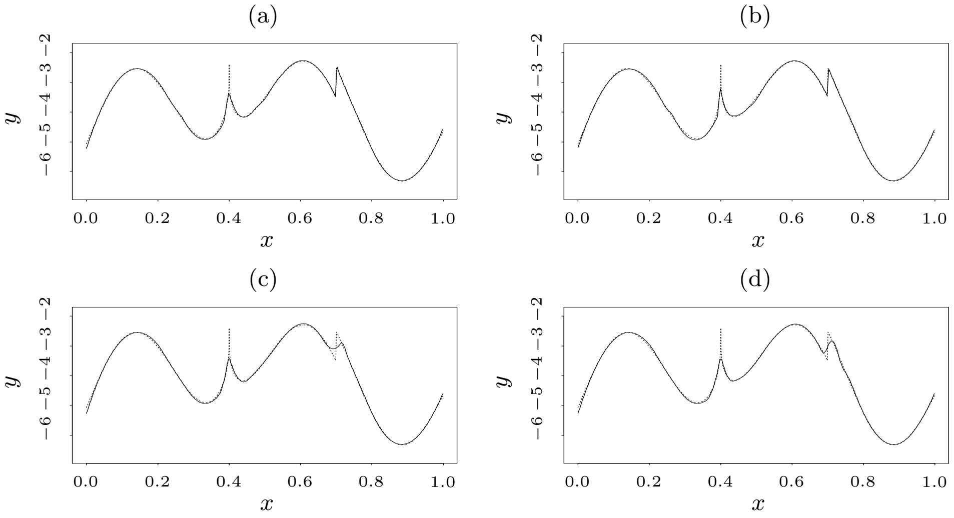

Figure 6. A comparison of ABMA and the hybrid sampler for the curve fitting example. The solid line is the estimated curve and dotted line is the true curve. (a) ABMA with $\mu = 0.01$. (b) ABMA with $\mu = 0.015$. (c) The hybrid sampler with $\lambda = 3$. (d) The hybrid sampler with $\lambda = 5$.

In this problem, the procedure is no longer automatic, since the value of μ needs to be specified by users. Here μ is treated as a hyperparameter and a value is directly assigned to it. The value can be improved by trial and error. Looking at Figure 5(a) and ignoring the true regression curve, we may feel that roughly 10 to 15 knots are needed to fit this data since ten local minima or maxima (including two boundary points) appear in the plot. Hence, 0.01 (10/1000) or 0.015 (15/1000) may serve as a good initial value of μ. This method usually gives a satisfactory result, as shown in Figures 5 and 6. The value of μ reflects our prior knowledge on the smoothness of the underlying function. If we believe it is very smooth, μ should be set to a small value, otherwise, it should be larger.

For comparison, we also applied the hybrid sampler (Denison, Mallick and Smith (1998)) to this example. The authors' program is downloadable from `http://www.ma.ic.ac.uk/∼dgtd`. In the hybrid sampler, one assumes that p has a prior truncated Poisson distribution with hyperparameter λ, σ^2 has a prior inverse Gamma and $\sigma^2 \sim IG(0.001, 0.001)$. One iteration of the hybrid sampler consists of the following two steps.

- Update the knot points, $t_1, \ldots, t_p$;
- Update σ^2.

The first step is accomplished by reversible jump MCMC (Green (1995)), the second step is accomplished by the Gibbs sampler (Geman and Geman (1984)). Given the knots $t_1, \ldots, t_p$, the coefficients $\boldsymbol{\beta}$ are estimated by standard least squares theory. We set $m_0 = 2$ and $m = 3$ as in EMC. The hybrid sampler was run for 3000 and 2000 iterations with $\lambda = 3$ and $\lambda = 5$ respectively. The computational times were 13.5m and 12.2m, respectively (on the same workstation as the runs of ABMA). In both runs, the first half of the iterations were discarded for the burn-in process, and the remaining iterations were recorded for inferences. Figure 6(c) and (d) shows the fitted curves from the two runs. Clearly the jump of the underlying function is less well approximated. We also ran the program with $\lambda = 1, 2$ and 10 and the same computational time. The results were similar.

This example shows that ABMA is superior to the hybrid sampler for the regression spline example, whatever in the computational time or the accuracy of fit. One reason is that, under the "automatic" prior setting, the nuisance parameters are integrated out from the full posterior and the resulting marginal posterior distribution is easier to sample from.

Acknowledgement

The authors would like to thank the referees for their suggestions and comments.

References

Akaike, H. (1973). Information theory and the extension of the maximum likelihood principle. In *Proc. Int. Symp. Information Theory* (Edited by B. N. Petrov and F. Czáki), 267-281. Akademia Kiadoó, Budapest.

Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics* **16**, 125-127.

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

Carlin, B. P. and Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *J. Roy. Statist. Soc. Ser. B* **57**, 473-484.

Clyde, M. (1999). Bayesian model averaging and model search strategies. In *Bayesian Statistics* **6** (Edited by J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), University Press, Oxford.

David, A. P. (1984). Statistical theory–the frequential approach. *J. Roy. Statist. Soc. Ser. A* **147**, 278-292.

Denison, D. G. T., Mallick, B. K. and Smith, A. F. M. (1998). Automatic Bayesian curve fitting. *J. Roy. Statist. Soc. Ser. B* **60**, 333-350.

Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion). *J. Roy. Statist. Soc. Ser. B* **57**, 45-97.

Ehrlich, L. (1973). Participation in illegitimate activities: a theoretical and empirical investigation. *J. Political Economy* **81**, 521-567.

Eubank, R. L. (1999). *Nonparametric Regression and Spline Smoothing.* 2nd edition. Marcel Dekker, New York.

Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *J. Roy. Statist. Soc. Ser. B* **57**, 371-394.

Fernández, G., Ley, E. and Steel, M. F. J. (2001). Benchmark priors for Bayesian model averaging. *J. Econometrics*, **100**, 381-427.

Foster, D. P. and George, E. I. (1994). The risk inflation criterion for multiple regression. *Ann. Statist.* **22**, 1947-1975.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *Ann. Statist.* **19**, 1-141.

Friedman, J. H. and Silverman, B. W. (1989). Flexible parsimonious smoothing and additive modeling (with discussion). *Technometrics* **31**, 3-21.

Geisser, S. and Eddy, W. (1979). A predictive approach to model selection. *J. Amer. Statist. Assoc.* **74**, 153-160.

Gelfand, A. E. (1995). Model determination using sampling-based methods. In *Markov Chain Monte Carlo in Practice* (Edited by W. R. Gilks, S. Richardson and D. J. Spiegelhalter), 145-161. Chapman and Hall, London.

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

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721-741.

George, E. I. (1999). Bayesian model selection. In *Encyclopedia of Statistical Science Update* **3** (Edited by S. Kotz, C. Read and D. Banks), 39-46. Wiley, New York.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88**, 881-889.

George, E. I. and 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 *Bayesian Statistics* **5** (Edited by J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), 609-620. Oxford Press.

Geyer, C. J. (1991). Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statistics: Proceedings of the* 23*rd Symposium on the Interface* (Edited by E. M. Keramigas), 153-163. Interface Foundation, Fairfax Station.

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization, & Machine Learning.* Addison Wesley.

Good, I. J. (1950). *Probability and the Weighting of Evidence.* Charles Griffin, London.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711-732.

Hansen, M. and Kooperberg, C. (2000). Spline adaptation in extended linear models. *Statist. Sci.*, tentatively accepted.

Hansen, M., Kooperberg, C. and Sardy, S. (1998). Triogram models. *J. Amer. Statist. Assoc.* **93**, 101-119.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models.* Chapman and Hall, London.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97-109.

Hocking, R. R. (1976). The analysis and selection of variables in linear regression. *Biometrics* **32**, 1-49.

Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. (1999). Bayesian model averaging: a tutorial (with discussion). *Statist. Sci.* **14**, 382-417.

Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems*, University of Michigan Press, Ann Arbor.

Hukushima K. and Nemoto, K. (1996). Exchange Monte Carlo method and application to spin glass simulations. *J. Phys. Soc. Jpn.* **65**, 1604-1608.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90**, 773-795.

Kirkpatrick, S., Gelatt, Jr., C. D. and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science* **220**, 671-680.

Kohn, R. and Ansley, C. (1987). A new algorithm for spline smoothing based on smoothing a stochastic process. *SIAM J. Sci. Statist. Comput.* **8**, 33-48.

Kohn, R., Ansley, C. and Tharm, D. (1991). The performance of cross-validation and maximum likelihood estimators of spline smoothing parameters. *J. Amer. Stat. Assoc.* **86**, 1042-1050.

Koo, J. Y. (1997). Spline estimation of discontinuous regression functions. *J. Comput. Graph. Statist.* **6**, 266-284.

Kooperberg, C. and Stone, C. (1992). Logspline density estimation for censored data. *J. Comput. Graph. Statist.* **1**, 301-328.

Laud, P. W. and Ibrahim, J. G. (1995). Predictive model selection. *J. Roy. Statist. Soc. Ser. B* **57**, 247-262.

Liang, F. and Wong, W. H. (2000). Evolutionary Monte Carlo: applications to C_p model sampling and change point problem. *Statist. Sinica* **10**, 317-342.

Madigan, D. and Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. Amer. Statist. Assoc.* **89**, 1535-1546.

Madigan, D. and York, J. (1995). Bayesian graphical models for discrete data. *Internat. Statist. Rev.* **63**, 215-232.

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. and Teller, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087-1091.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression (with discussion). *J. Amer. Statist. Assoc.* **83**, 1023-1036.

Müller, H. G. and Stadtmüller, U. (1987). Variable bandwidth kernel estimators of regression curves. *Ann. Statist.* **15**, 182-201.

O'Hagan, A. (1995). Fractional Bayes factor for model comparison (with discussion). *J. Roy. Statist. Soc. Ser. B* **57**, 99-138.

Phillips, D. B. and Smith, A. F. M. (1995). Bayesian model comparison via jump diffusions. In *Markov Chain Monte Carlo in Practice* (Edited by W. R. Gilks, S. Richardson and D. J. Spiegelhalter), 215-239. Chapman and Hall, London.

Raftery, A. E. (1996). Approximate Bayes factor and accounting for model uncertainty in generalized linear models. *Biometrika* **83**, 251-266.

Raftery, A. E., Madigan, D. and Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *J. Amer. Statist. Assoc.* **92**, 179-191.

San Martini, A. and Spezzaferri, F. (1984). A predictive model selection criterion. *J. Roy. Statist. Soc. Ser. B* **46**, 296-303.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-464.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Ann. Statist.* **22**, 1701-1762.

Vandaele, W. (1978). Participation in illegitimate activities: Ehrlich revisited. In *Dererrence and Incapacitation* (Edited by A. Blumstein, J. Cohen and D. Nagin), 270-335. National Academy of Science Press, Washington.

Wang, Y. (1995). Jump and sharp cusp detection by wavelets. *Biometrika* **82**, 385-397.

Zeller, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques: Essays in Honour of Bruno de Finetti* (Edited by P. K. Goel and A. Zellner), 233-243. North-Holland, Amsterdam.

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

E-mail: stalfm@nus.edu.sg

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

Email: statykn@nus.edu.sg

Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, MA02115, U.S.A.

Email: wwong@hsph.harvard.edu

(Received May 2000; accepted February 2001)