Published in final edited form as:
J R Stat Soc Series B Stat Methodol. 2008 October 14; 71(1): 143–158. doi:10.1111/j.1467-9868.2008.00678.x.

Bayesian model selection using test statistics

Jianhua Hu and **Valen E. Johnson**
University of Texas M. D. Anderson Cancer Center, Houston, USA

Summary

Existing Bayesian model selection procedures require the specification of prior distributions on the parameters appearing in every model in the selection set. In practice, this requirement limits the application of Bayesian model selection methodology. To overcome this limitation, we propose a new approach towards Bayesian model selection that uses classical test statistics to compute Bayes factors between possible models. In several test cases, our approach produces results that are similar to previously proposed Bayesian model selection and model averaging techniques in which prior distributions were carefully chosen. In addition to eliminating the requirement to specify complicated prior distributions, this method offers important computational and algorithmic advantages over existing simulation-based methods. Because it is easy to evaluate the operating characteristics of this procedure for a given sample size and specified number of covariates, our method facilitates the selection of hyperparameter values through prior-predictive simulation.

Keywords

Bayes factor; Coherency; False discovery rate; *F*-statistic; Likelihood ratio statistic; Model selection

1. Introduction

In many applied settings, constructing a regression model for a response variable requires the selection of explanatory variables to include in the regression function. In linear regression, this problem can be posed algebraically as the selection of a model of the form

$$y_k = \sum_{j=1}^{q} \theta_{i_j} X_{k,i_j} + \varepsilon_k, \tag{1}$$

where $\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_q}$ denotes a subset of p covariate vectors $\mathbf{X}_1, \ldots, \mathbf{X}_p$ to maximize a given criterion.

When p is small or moderate in size, it is usually possible to compare all 2^p models and to select the model that optimizes the chosen criterion. Model selection criteria that are frequently used include R^2, the Akaike information criterion AIC, C_p and the Bayes information criterion BIC. Unfortunately, for large p it is usually not feasible to compute the criterion function for all possible models.

When p is large, a common approach for overcoming computational problems that are associated with selecting from among a large number of potential models is to reduce adaptively

Address for correspondence: Jianhua Hu, Department of Biostatistics, Division of Quantitative Sciences, University of Texas M. D. Anderson Cancer Center, Unit 447, 1515 Holcombe Boulevard, Houston, TX 77030, USA. JHu@mdanderson.org.

the number of models considered. This is essentially the idea behind stepwise procedures like forward selection and backward elimination (see, for example, Miller (1990)).

Bayesian methods have recently played a prominent role in the development of model selection criteria and are based on comparisons of the marginal probabilities assigned to the data by each potential model. In the 'small *p*' setting, research in this direction includes Lempers (1971), Atkinson (1978), Pericchi (1984), Smith and Spiegelhalter (1980), Spiegelhalter and Smith (1982), Zellner (1984), Stewart (1987) and Mitchell and Beauchamp (1988).

The advent of Gibbs sampling and other Markov chain Monte Carlo (MCMC) techniques have made it possible to extend Bayesian model selection procedures to the 'large *p*' setting. An early contribution to this development was made by George and McCulloch (1993, 1997), who proposed a 'stochastic search variable selection' (SSVS) procedure to determine promising subsets of predictor variables. Their idea was to use indicator variables to identify subset choices according to posterior probabilities defined within the context of a hierarchical Bayesian mixture model.

A somewhat different approach was proposed by Madigan and Raftery (1994). Their method focused on Bayesian model averaging for prediction. In making predictions for future observations, Bayesian model averaging accounts for model uncertainty, which often represents a major component of prediction uncertainty (e.g. Leamer (1978), Hodges (1987), Raftery (1996) and Draper (1995)). To apply Bayesian model averaging in high dimensional settings, Raftery *et al.* (1997) extended this algorithm in two ways. First, they averaged over a reduced set of models (i.e. Occam's window). Second, they used an MCMC simulation approach (called MC^3) to sample from the space of possible models.

The Bayesian approach to model selection has also been extended to generalized linear models —specifically probit regression models (Lee *et al.*, 2003). Using latent variable methodology, the method of Lee *et al.* (2003) casts variable selection for probit regression models into a framework which is similar to that previously developed for linear models.

The basic strategy of these MCMC algorithms is to traverse the space of possible models according to posterior model probabilities. In SSVS-type approaches, paths through the model space are selected by updating binary variables that indicate whether or not explanatory variables are included in a particular model. The values of these indicator variables are determined probabilistically according to prior distributions and the values of fully parametric Bayes factors between the models that they represent. Our innovation is to base these transitions on approximations of Bayes factors based on test statistics (e.g. Johnson (2005, 2008)). For brevity, we shall refer to these approximations as test-based Bayes factors (TBFs) for the remainder of this paper.

TBFs have the potential for greatly simplifying the process of Bayesian model selection, particularly in linear and generalized linear models. Through their use, the specification of prior distributions on nuisance parameters can often be avoided. Because this approach is based on the values of test statistics (which can generally be obtained through maximization procedures), it is computationally faster than model selection procedures that rely on numerical integration to obtain marginal densities of data.

Our approach also facilitates the specification of prior hyperparameters through the implementation of prior-predictive simulation studies. Such studies allow us to specify hyperparameters that are based on either the sampling properties of the induced model selection algorithm or the evaluation of decision theoretic criteria.

Finally, we note that Schwarz's BIC (Schwarz, 1978) has close connections to the TBF based on the likelihood ratio statistic (LRS). As a result, several theoretical properties of BIC also extend to our method (e.g. Nishii (1984)).

2. Model selection based on test statistics

The Bayes factor between two models represents the ratio of the marginal densities of the data obtained under each model and, together with prior model probabilities, determines the posterior odds. The basic idea behind our model selection procedure is to approximate the Bayes factor between two nested models by using the TBF that is obtained by modelling the distribution of a test statistic defined from the original data. TBFs based on the LRs (LRTBFs) have particularly favourable properties, which we illustrate below as the basic motivation behind our procedure.

Suppose that $\mathbf{y}_1, \ldots, \mathbf{y}_n$ represent n independent and identically distributed (IID) realizations of a q-dimensional random vector having density function $f(\mathbf{y}|\boldsymbol{\theta})$ defined with respect to a σ-finite measure μ, and that $\{\mathbf{y}_i\}$ assume values in a sample space S that does not depend on the p-dimensional parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Consider the test of the null model

$$H_1: \theta = (\theta_1^0, \theta_2), \tag{2}$$

where $\theta_1^0 = (\theta_1^0, \ldots, \theta_d^0)$ is specified and $\boldsymbol{\theta}_2 = (\theta_{d+1}, \ldots, \theta_p)$ is unconstrained, against the sequence of alternative models

$$H_{2n}: \theta^n = (\theta_1^n, \theta_2) \tag{3}$$

where

$$\theta_i^n = \theta_i^0 + \delta_{i,n}/n^{1/2}, \quad \text{with } \lim_{n\to\infty}(\delta_{i,n}) = \delta_i, i = 1, \ldots, d. \tag{4}$$

Let λ_n denote the ratio of the likelihood function evaluated at the constrained maximum likelihood estimate (obtained by fixing θ_1^0) to the likelihood function evaluated at the unconstrained maximum likelihood estimate. Under regularity conditions that are specified in Appendix A, the distribution of the LRS, say $z_n \equiv -2\log(\lambda_n)$, converges to a χ^2-distribution on d degrees of freedom when the null model is true. Under the same regularity assumptions, Davidson and Lever (1970) showed that the distribution of the LRS under the sequence of alternative models specified above converges to a non-central χ^2-distribution with non-centrality parameter

$$\Delta = \delta' \bar{\mathbf{C}}_{1,1} \delta \tag{5}$$

and d degrees of freedom. In equation (5), $\bar{\mathbf{C}}_{1,1}$ denotes the upper $d \times d$ submatrix of the inverse of the matrix $\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}^{-1}$ is the information matrix and $\boldsymbol{\delta} = \{\delta_i\}$.

Assuming that $\boldsymbol{\delta}$ has a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $cn\boldsymbol{\Sigma}_{1,1}$ ($\boldsymbol{\Sigma}_{1,1}$ is the upper $d \times d$ submatrix of $\boldsymbol{\Sigma}$), Johnson (2008) showed that the marginal distribution of z_n under the sequence of alternative models specified in expression (3)–(4) is approximated by a gamma distribution with shape parameter $d/2$ and scale parameter $1/2(cn + 1)$ (denoted below by $g\{\cdot|d/2, 1/2(cn + 1)\}$).

For this sequence of alternative models, the TBF in favour of the alternative model for a fixed value of n can be obtained by marginalizing over the non-centrality parameter, resulting in

$$\mathrm{TBF}(H_{2n}|H_1)=\frac{g\{z_n|d/2, 1/2(cn+1)\}}{g(z_n|d/2, 1/2)}=(cn+1)^{-d/2}\exp\left\{\frac{cnz_n}{2(cn+1)}\right\}. \quad (6)$$

The factor cn appearing in the covariance matrix of the non-centrality parameter implies that the difference between values of $\boldsymbol{\theta}$ that are drawn under the full and reduced models is $O_p(1)$, which ensures that the LRTBF is consistent (Johnson, 2008).

Similar results have been obtained in non-IID settings. For example, Taniguchi (1991) demonstrated that the LRS has the distribution of a non-central χ^2-distribution against local alternatives for general classes of stochastic processes and applied these results to multivariate analyses, time series analyses and non-linear regression problems. Of particular interest here are the results of Cordeiro *et al.* (1994), who showed that the asymptotic distribution of the LRS under local alternative models in generalized linear models is also a non-central χ^2-distribution, and that the relationship between the associated non-centrality parameter and information matrix is similar to the relationship that was cited above for the IID case. More recently, Banerjee (2005) showed that the LRS has a non-central χ^2-distribution in certain classes of regular semiparametric models. In general, if the distribution of the LRS under the null hypothesis is approximately χ^2, and if its distribution under the alternative hypothesis can be approximated by a non-central χ^2-distribution, then a TBF of the form (6) can be obtained by assuming that the non-centrality parameter under the alternative model has a rescaled χ^2-distribution. Marginalizing over the distribution of the non-centrality parameter then leads to equation (6) (Johnson, 2008).

For non-local alternatives, the LRS grows exponentially fast with increasing sample size when the alternative hypothesis is true (Bahadur, 1965), which means that the LRTBF is also consistent in this setting.

2.1. A Markov chain Monte Carlo model selection algorithm

Returning to the problem of model selection, we now illustrate how TBFs can be used to select between models that are indexed by the p-dimensional parameter $\boldsymbol{\theta}$.

To fix the notation, assume that a component of $\boldsymbol{\theta}$, say θ_h, can be excluded from a model if its value is 0, and denote a model by $\mathbf{j} = \{j_1, \ldots, j_k\}$ ($1 \leq j_1 < \ldots < j_k \leq p$) if and only if $\theta_{j_1} \neq 0, \ldots, \theta_{j_k} \neq 0$ and all other elements of $\boldsymbol{\theta}$ are 0. Let $\mathcal{J}$ denote the set of 2^p possible models that can be defined from the p components of $\boldsymbol{\theta}$. In the context of linear models (1), model $\mathbf{j}$ corresponds to the regression model that includes all covariates X_i for which $i \in \mathbf{j}$. We denote the null model by $\varnothing$ and write $\mathbf{k} \subseteq \mathbf{j}$ to indicate that model $\mathbf{j}$ contains all components of $\boldsymbol{\theta}$ that are present in model $\mathbf{k}$.

If $\Pi(\mathbf{j}|\pi)$ denotes the prior probability that is assigned to model $\mathbf{j}$ for a given value of a hyperparameter π, standard results from Bayesian theory imply that the posterior probability of model $\mathbf{j}$ can be expressed as

$$\Pr(\mathbf{j}|\pi, \boldsymbol{Y}) = \mathrm{BF}(\mathbf{j}|\varnothing)\Pi(\mathbf{j}|\pi) / \sum_{\mathbf{i}\in\mathcal{J}} \mathrm{BF}(\mathbf{i}|\varnothing)\Pi(\mathbf{i}|\pi). \tag{7}$$

Here, BF($\mathbf{i}|\varnothing$) denotes the Bayes factors between model $\mathbf{i}$ and the null model, which is calculated from the complete data $\mathbf{Y} = \{\mathbf{y}_h\}$. Approximating BF($\mathbf{i}|\varnothing$) by TBF($\mathbf{i}|\varnothing$) for all $\mathbf{i} \in \mathcal{J}$ leads to an approximation of the posterior probability of model $\mathbf{j}$ given by

$$\Pr\{\mathbf{j}|\pi, \mathbf{z}(\boldsymbol{Y})\} \approx \mathrm{TBF}(\mathbf{j}|\varnothing)\Pi(\mathbf{j}|\pi) / \sum_{\mathbf{i}\in\mathcal{J}} \mathrm{TBF}(\mathbf{i}|\varnothing)\Pi(\mathbf{i}|\pi), \tag{8}$$

where $\mathbf{z}(\mathbf{Y})$ denotes the vector of test statistics computed from $\mathbf{Y}$ that are used to compute the TBFs. The vector $\mathbf{z}(\mathbf{Y})$ can be based on any test statistic, provided that the distributions of the resulting statistics are approximately known under each model $\mathbf{j} \in \mathcal{J}$. For example, the form of the LRTBF that is provided in equation (6) is for the case in which model $\mathbf{j}$ corresponds to hypothesis H_{2n} and model $\varnothing$ corresponds to H_1. If $0 \leq \mathrm{TBF}(\mathbf{i}|\varnothing) < \infty$ for all models $\mathbf{i} \in \mathcal{J}$, this approximation defines a probability distribution on the model space. The validity of such approximations for choosing between two nested models was explored by Johnson (2005, 2008) and is explored in model selection settings in the examples that follow.

Although in general the choice of Π is arbitrary, in this paper we assume that the prior probability that is assigned to model $\mathbf{i}$ can be expressed as

$$\Pi(\mathbf{i}|\pi) = \begin{cases} w\pi^{|\mathbf{i}|}(1-\pi)^{p-|\mathbf{i}|} & \text{if } |\mathbf{i}| < n, \\ 0 & \text{otherwise}, \end{cases} \tag{9}$$

where $|\mathbf{i}|$ denotes the number of parameters that are included in model $\mathbf{i}$ and w is a constant of proportionality. To reflect uncertainty in the unknown value of π, we assume that the prior for π, say $g(\pi)$, is a beta$\{\delta p, (1-\delta)p\}$ distribution for some $0 < \delta < 1$. The precision parameter of this prior density is chosen to be proportional to p so that the prior retains influence as p becomes large; this influence is important for discouraging large models in high dimensional settings. The selection of the hyperparameter δ is discussed below.

We now focus on the problem of estimating posterior model probabilities based on expressions (8) and (9) when p is large. As in the case of fully parametric Bayes factors, it is necessary to use a simulation procedure to obtain estimates of model probabilities in this setting.

To define such an algorithm, let γ_k, $k = 1, \ldots, p$, denote binary variables that indicate whether θ_k is included in a model ($\gamma_k = 1$) or not ($\gamma_k = 0$). Note that there is a one-to-one correspondence between the vectors $\boldsymbol{\gamma} = \{\gamma_k\}$ and models $\mathbf{j}$, i.e. $\gamma_k = 1$ if and only if $k \in \mathbf{j}$.

On the basis of the correspondence between $\boldsymbol{\gamma}$ and the model space, a Gibbs sampling procedure for obtaining an approximate posterior sample from the model space can be defined as follows.

- *Step 1*: choose an initial value for the vector $\boldsymbol{\gamma}$ satisfying $\Sigma\gamma_k < n$ and set the iteration number $i = 1$. Set $\pi_0 = \Sigma\gamma_k/n$.
- *Step 2*: for $k = 1, \ldots, p$, perform a Gibbs update of γ_k on the basis of approximation (8) as follows. Let $\mathbf{j}_0$ denote the model corresponding to $\gamma_k = 0$ (i.e. $\{\gamma_1, \ldots, \gamma_{k-1}, 0, \gamma_{k+1}, \ldots, \gamma_p\}$), and let $\mathbf{j}_1$ denote the model corresponding to $\gamma_k = 1$ (i.e. $\{\gamma_1, \ldots, \gamma_{k-1},$

1, γ_{k+1}, …, γ_p}). Then, from approximation (8), the conditional probability of model $\mathbf{j}_0$, given γ_h, $h \neq k$, and π_{i-1}, is

$$r = \frac{\Pr\{\mathbf{j}_0|\pi_{i-1}, \mathbf{z}(\boldsymbol{Y})\}}{\Pr\{\mathbf{j}_0|\pi_{i-1}, \mathbf{z}(\boldsymbol{Y})\} + \Pr\{\mathbf{j}_1|\pi_{i-1}, \mathbf{z}(\boldsymbol{Y})\}} = \frac{\mathrm{TBF}(\mathbf{j}_0|\varnothing)\Pi(\mathbf{j}_0|\pi_{i-1})}{\mathrm{TBF}(\mathbf{j}_0|\varnothing)\Pi(\mathbf{j}_0|\pi_{i-1}) + \mathrm{TBF}(\mathbf{j}_1|\varnothing)\Pi(\mathbf{j}_1|\pi_{i-1})}. \quad (10)$$

Similarly, the conditional probability of model $\mathbf{j}_1$ is $1 - r$. To update γ_k, set $\gamma_k = 0$ with probability r; otherwise set $\gamma_k = 1$.

- *Step 3*: record model $\mathbf{m}_i$, the model corresponding to the current value of $\boldsymbol{\gamma}$.
- *Step 4*: sample π_i from its full conditional distribution, which is proportional to

$$\Pi(\mathbf{m}_i|\pi_i)g(\pi_i) \propto \pi_i^{\upsilon}(1-\pi_i)^{p-\upsilon}\pi_i^{\delta p-1}(1-\pi_i)^{p-\delta p-1},$$

a beta distribution with parameters ($\delta p + \upsilon$, $2p - \delta p - \upsilon$), where $\upsilon = |\mathbf{m}_i| = \Sigma_h \gamma_h$.

- *Step 5*: increment i and return to step 2.

After a burn-in period of I updates, the chain of models $\mathbf{m}_{I+1}$, $\mathbf{m}_{I+2}$, … represents an ergodic Markov chain that can be used to perform approximate inference regarding the posterior probability of various model configurations.

The ergodicity of the Markov chain $\mathbf{m}_{I+1}$, $\mathbf{m}_{I+2}$, … follows from the fact that it is possible to step between any two models in this MCMC scheme in two iterations. Convergence properties of this Gibbs sampler are investigated further in Section 5.

2.2. Properties of the likelihood ratio test-based Bayes factor for model selection

In principle, any test statistic can be used to define TBFs between nested models. Indeed, *F*-statistics have essentially been used this way for linear models (Liang *et al.*, 2005). We briefly consider this possibility in the example of Section 3.1 but note that LRTBFs have several favourable properties that are not shared by TBFs which are based on other test statistics.

The LRTBF has a coherency property which is not inherited by TBFs that are based on other test statistics. Specifically, if three nested models satisfy $\mathbf{k}_1 \subseteq \mathbf{k}_2 \subseteq \mathbf{k}_3$, then the TBFs between the models are coherent, i.e.

$$\mathrm{TBF}(\mathbf{k}_3|\mathbf{k}_2)\mathrm{TBF}(\mathbf{k}_2|\mathbf{k}_1) = \mathrm{TBF}(\mathbf{k}_3|\mathbf{k}_1). \quad (11)$$

This coherency property of the LRTBF implies that equation (10) can be re-expressed as

$$\frac{\Pi(\mathbf{j}_0|\pi_{i-1})}{\Pi(\mathbf{j}_0|\pi_{i-1}) + \mathrm{TBF}(\mathbf{j}_1|\mathbf{j}_0)\Pi(\mathbf{j}_1|\pi_{i-1})} = \frac{1-\pi_{i-1}}{1-\pi_{i-1}+\pi_{i-1}\mathrm{TBF}(\mathbf{j}_1|\mathbf{j}_0)}. \quad (12)$$

We note that this coherency property does not hold for TBFs that are based on, for example, the F-statistic (e.g. Liang *et al.* (2005)). This means that marginal posterior model probabilities that are estimated by using other test statistics are not invariant to the choice of the baseline model.

Because the LRS can be computed in most regular parametric models, the LRTBF gains the advantage of applicability to a broad class of model selection problems, with the only requirement that the distribution of the LRS be approximately non-central χ^2 when the larger of two tested models is true.

Finally, a favourable asymptotic property of the model selection procedure that is based on the LRTBF can be inferred from results that were elaborated by Nishii (1984). For linear models of the form (1), suppose that the error terms ε_k are IID $N(0, \sigma^2)$ random variables, and let $\mathbf{X}_n = \{X_{ij}\}, j = 1, \ldots, p, i = 1, \ldots, n$, denote the design matrix. Assume further that $\mathbf{X}_n'\mathbf{X}_n$ is positive definite for all $n > n_0$, and that $M = \lim_{n\to\infty}(n^{-1}\mathbf{X}_n'\mathbf{X}_n)$ exists and is positive definite. Let $\mathbf{j}_t$ denote the 'true' (i.e. data-generating) model. Define $J_1 = \{\mathbf{j} \in \mathcal{J} | \mathbf{j}_t \not\subseteq \mathbf{j}\}$ and $J_2 = \{\mathbf{j} \in \mathcal{J} | \mathbf{j}_t \subseteq \mathbf{j}\}$. Then the LRTBF between an arbitrary model $\mathbf{j} \in \mathcal{J}$ and $\varnothing$ satisfies

$$\log\{\mathrm{TBF}(\mathbf{j}|\varnothing)\} = -\frac{d}{2}\log(cn+1) + \frac{cnz_n}{2(cn+1)}, \tag{13}$$

where $d = |\mathbf{j}|$ and z_n is the LRS. Regarded as a model selection criterion, condition (13) represents a special case of Nishii's generalized information criterion. Letting $\hat{\mathbf{j}}$ denote the model that is obtained by maximizing expression (13) over the set $\mathcal{J}$, Nishii obtained the following result.

Theorem 1 (Nishii, 1984)—Assume that the conditions that were stated in the previous paragraph obtain and define $p_n(\mathbf{j}) = \Pr(\hat{\mathbf{j}} = \mathbf{j})$ for $\mathbf{j} \in \mathcal{J}$. Then the following conditions describe $p_n(\mathbf{j})$.

- **a.** If $\mathbf{j} \in J_1$, then $p_n(\mathbf{j}) = o(n^{-h})$ for any positive constant h.
- **b.** If $\mathbf{j} \in J_2 - \{\mathbf{j}_t\}$, then $p_n(\mathbf{j}) = o(1)$.

Thus, only the true model $\mathbf{j}_t$ has a non-negligible probability of being the model that is selected by condition (13) as $n \to \infty$.

2.3. Simulation-based methods for specifying hyperparameters

In standard Bayesian model selection procedures, it is sometimes difficult to interpret the meaning of model hyperparameters. For example, the marginal probability that a variable is included in the regression function is often a complicated function of several hyperparameters.

An advantage of basing model selection procedures on TBFs is that it is simple to simulate the operating characteristics of the selection procedure as the underlying model hyperparameters are varied. For example, the LRTBF selection procedure depends on two hyperparameters: c and δ. To evaluate the effects of (c, δ) on either the false discovery rate (FDR) or false positive rate of the LRTBF model selection procedure, the MCMC algorithm that was defined in Section 2.1 can be run over a range of these hyperparameters, *without using actual data*, i.e., instead of calculating test statistics from data, the values of the test statistics from either the null or alternative models are simulated. A prior belief that a proportion ρ of regression parameters in a model is non-zero means that χ_1^2 random variables are simulated when updating one of the $p(1 - \rho)$ variables assumed to have a regression coefficient equal to 0. When updating the indicator variables corresponding to non-zero regression coefficients, independent $g\{1/2, 1/2(cn + 1)\}$ random variables are simulated instead. Using this procedure, the frequentist operating characteristics of a model selection algorithm can be evaluated without having to simulate data explicitly from the given experimental or observational design.

3. Examples

3.1. Linear model variable selection

To explore the properties of our method for approximating posterior model probabilities, we applied the algorithm from the previous section to a well-known data set that was initially presented by Vandaele (1978) and to which Raftery *et al.* (1997) later applied Bayesian model selection procedures. The data set contained information that had been collected from 47 states on recorded crime rates from 1959 to 1960 in the USA. 15 demographic and socio-economic variables, which are listed in Table 1, were considered potential predictors of the crime rate.

From these 15 candidate predictors, we have $2^{15} = 32768$ potential models. To facilitate comparisons with other model selection procedures, we transformed the crime rates to the logarithmic scale before our analysis.

For illustration, we assumed that the logarithm of the crime rate was related to some subset of the explanatory variables according to model (1), for which we assumed the error terms ε_k to be IID $N(0, \sigma^2)$ random variables. We applied our model selection algorithm by using TBFs based on both the LRS and F-statistics. In the LRTBF model selection procedure, we used the prior simulation procedure from Section 2.3, with observed values of $n = 47$ and $p = 15$ and an assumed value of $c = 2$, to choose a hyperparameter value for the model inclusion probability of δ.

We arbitrarily chose a value of $c = 2$, which made the prior variance of the regression parameter twice that implied by the information matrix. In practice, we have found that values of c in the range (2, 6) yield high posterior probability models that have both favourable predictive properties in cross-validation experiments and favourable operating characteristics. Fig. 1 illustrates the results from this procedure as values of δ were varied between 0 and 0.5 and the proportion of covariates that were assumed to have non-zero regression coefficients was varied in $\rho \in \{1/3, 7/15, 3/5\}$.

Points in the scatter plot represent the empirically observed FDR for each value of δ. From Fig. 1, it follows that, if we were to assume *a priori* that fewer than a third of the covariates were substantively related to the crime rate, then a choice of $\delta = 0.5$ for $c = 2$ would result in an average FDR of less than 22%. At the same parameter values, the probability of obtaining four or fewer false positive results was similarly estimated to be less than 0.95. As these operating characteristics seemed reasonable, we used those parameter values in the LRTBF model selection algorithm that is described below.

Although a TBF that is based on the F -statistic does not have the favourable coherency property of the LRTBF, for pairs of nested linear models there is a close connection between the TBF that is based on the F -statistic and fully parametric Bayes factors (Johnson, 2008). For this reason, we also examined the numerical properties of selection procedures that were based on a TBF derived from the F -statistic.

To define a TBF that is based on the F -statistic between nested linear models, let $\mathbf{X}_c$ denote the design matrix containing covariates in the larger of the two nested models, and assume that the prior distribution on $\boldsymbol{\theta}$ under the larger model is a normal distribution centred on a value of $\boldsymbol{\theta}$ for which the tested component of $\boldsymbol{\theta}$ is 0 and the covariance matrix is $n\tau\sigma^2(\mathbf{X}_c'\mathbf{X}_c)^{-1}$. The TBF between the full and reduced models can then be expressed as

$$(1+n\tau)^{-1/2}\left\{\frac{n-r+f}{n-r+f/(1+n\tau)}\right\}^{(n-r+1)/2},$$

where r is the dimension of $\boldsymbol{\theta}$ under the larger model and f is the observed value of the F -statistic (Johnson, 2005).

To fix τ and δ, we performed another prior simulation experiment to study the variation of the FDR and false positive rates with these hyperparameter values. Values of $\tau = 2$ and $\delta = 0.5$ produced FDRs and false positive rates that were similar to the results that were obtained from the LRTBF when $c = 2$ and $\delta = 0.5$, so we used those values to perform model selection with the F -statistic TBF.

3.1.1. Numerical results—We compared outputs from our method with those from several other Bayesian model selection procedures, including the Bayesian model averaging algorithm implemented with Occam's window and MC^3 (Raftery *et al.*, 1997) and SSVS (George and McCulloch, 1993). We implemented competing models using the hyperparameter values that were recommended in the original references, with one exception: for SVSS, we assumed that $\mathbf{R} \propto (\mathbf{X}'\mathbf{X})^{-1}$, where $\mathbf{X}$ denotes the relevant model design matrix. Our choice for $\mathbf{R}$ was one of two that were recommended by George and McCulloch (1993) (they used the other choice in their examples). We used that value of $\mathbf{R}$ to make the prior assumptions of the four selection methods more comparable. We also examined three common frequentist variable selection criteria: Efroymson's stepwise regression (e.g. Miller (1990)), Mallows's C_p (Mallows, 1973) and the adjusted R^2 (e.g. Weisberg (1985)).

A comparison of the four Bayesian model selection procedures appears in Table 2. The column to the right indicates the proportion of times that a model configuration was sampled. For convenience, the values of the model hyperparameters are displayed below each procedure.

The high probability models that were selected by using the four Bayesian procedures were quite similar. Most models that were selected by our method appear among the top six models that were chosen by Occam's window and MC^3, albeit in somewhat different order. For example, the top model based on the LRTBF matches the top model chosen by Occam's window and MC^3, and is also the model that was chosen by the classical stepwise procedure (Table 3).

Table 2 also displays estimated median probability models (i.e. models containing covariates with greater than 50% inclusion probabilities; see Barbieri and Berger (2004) for a theoretical justification of such models). The LRTBF model selection procedure and SSVS produced identical median probability models, whereas the median probability models that were selected by Occam's window and by procedures that were based on the F -statistic were slightly smaller. Interestingly, MC^3 selected variables 4 and 5 in the median probability model even though these variables were highly correlated.

The four Bayesian model selection procedures exhibited similar posterior-predictive properties. For instance, we performed a random 50% split of the observations into training and test samples. Under each model selection procedure, the training sample was then used to produce a 90% posterior-predictive coverage interval for each observation in the test sample. The observed coverage rates of these intervals fell within 4% of each other for all four procedures.

Finally, we estimated the posterior probability that each predictor was in the model. Table 3 lists these marginal posterior probabilities for each of the Bayesian model selection techniques that were described above.

Several trends are apparent from Table 3. First, there are several differences between the marginal probabilities implied by the various selection criteria, with the greatest differences occurring between Occam's window and the remaining models. For instance, Occam's window yields marginal probabilities of 0 for several parameters that were assigned non-negligible probability by the other selection criteria. Despite such differences, it is interesting to compare the results by ranking each predictor's marginal probability according to each of the models. These ranks are provided in parentheses in Table 3 and show that ranks that are based on the LRTBF were similar to all three fully parametric approaches, i.e. all the methods had relatively high agreement over the variables that were assigned highest posterior probability. These variables included income inequality (variable 13), mean years of schooling (variable 3), probability of imprisonment (variable 14), police expenditure in 1960 (variable 4) and number of non-whites per 1000 people (variable 9).

Computationally, our method is an order of magnitude faster than the existing Bayesian model selection techniques. In a naive R implementation, our method was 10 times faster than professionally written MC^3 code and was over 20 times faster than SSVS. More generally, computational savings can be expected over competing model selection algorithms whenever these algorithms involve numerical simulation to compute what are essentially marginal densities of data. The function maximization that is required to obtain the LRS will generally be faster than the integration that is necessary to obtain the marginal densities of data.

3.2. Simulated data

To evaluate our model selection procedure further, we performed an experiment in which we simulated $n = 50$ observations with $p = 49$ potential covariates. Each of the potential predictors $\mathbf{X}_j$, $j = 1, \ldots, 49$, was simulated as an independent vector of $N(0, 1)$ deviates, and the dependent variable was generated according to

$$\mathbf{y}=\mathbf{X}\theta+\varepsilon,$$

where $\mathbf{X} = \{\mathbf{X}_j\}$. Regression coefficients were assigned values $(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6) = (0.5, 1, 1.5, 2, 2.5, 3)$, and $(\theta_7, \ldots, \theta_{49}) = (0, \ldots, 0)$. Observational errors were assumed to be generated independently from an $N(0, 1)$ distribution.

To fix hyperparameter values in the LRTBF procedure, we again performed prior model simulations using a variety of choices for c, δ and ρ. We took $c = 2$, the same value as used for the crime data, and chose $\delta = 0.001$. For these values of (c, δ), the average FDR was 0.10, and the number of false positive results was less than 1 with a probability of 0.98 for $\rho = 6/49$.

On the basis of 100000 updates of the Gibbs sampling algorithm that was described in Section 2.1 (after a burn-in period of 1000 updates), the estimates of the marginal posterior probabilities that were obtained from the LRTBF algorithm for the inclusion of the first six covariates were 0.16, 1, 1, 1, 1 and 1. The median model included variables 2–6, which correctly identified five non-zero coefficients and no false positive values. The default implementation of the SSVS algorithm yielded estimates of 0.03, 0.006, 0.62, 0.45, 0.98 and 0.97 for the marginal posterior probabilities of the first six covariates; the median probability model included variables 3, 5 and 6. The Bayesian model averaging package implementation (Raftery *et al.*, 2006) of MC^3 produced estimated marginal posterior inclusion probabilities of 0.33, 1, 1, 1, 1 and 1 for the

first six covariates; the median probability model included variables 2–6 and 8. Thus, the performance of MC^3 was similar to that of the LRTBF algorithm, whereas SSVS performed slightly worse. The poor performance of SSVS in this setting probably stemmed from its dependence on a variance estimate from the saturated model, which had only 1 degree of freedom.

To assess the sensitivity of the LRTBF method, we also ran the algorithm for $\delta = 0.01$ and $\delta = 0.1$. For $\delta = 0.01$, the median probability again contained only variables 2–6, whereas the median probability model for $\delta = 0.1$ contained variables 2–6 and 8. The posterior inclusion probabilities for the latter case were 0.39, 1, 1, 1, 1 and 1, which agreed closely with the results that were obtained by using MC^3.

Neither Bayesian model averaging nor SSVS extended to the case of $p > n$. However, to test our algorithm in this setting, we expanded the simulation to include an additional 251 spurious covariates (i.e. $p = 300$) and used the same hyperparameter values as cited above. With these additional covariates, the estimated marginal posterior probabilities for inclusion of the first six covariates were 0.013, 0.996, 1, 1, 1 and 1, and the median probability model included variables 2–6 and three other predictors.

4. Simulation examples with binary outcomes

We also considered model selection in binary regression models. Although our method extends directly to generalized linear models with arbitrary link functions, we focused on probit regression models to facilitate comparisons with existing, fully parametric Bayesian model selection procedures as discussed in, for example, Lee *et al.* (2003).

The standard probit regression model may be written as

$$y_i \sim \text{Bernoulli}(\pi_i), \quad i=1,\ldots,n, \quad y_i \perp y_j, \quad i \neq j,$$

where

$$\pi_i=\Phi(\mathbf{x}_i'\theta). \tag{14}$$

Here, $\Phi(\cdot)$ denotes the standard normal distribution function and $\mathbf{x}_i$ denotes a $p \times 1$ vector of covariates that are relevant for predicting π_i. As before, we focused on selecting subsets of the p covariates for use in the regression equation.

To assess the performance of our LRTBF model selection procedure and to compare its performance with a fully parametric procedure that was described in Lee *et al.* (2003), we simulated $n = 30$ observations and $p = 15$ predictors. All components of each predictor $\{\mathbf{x}_i\}$ were again generated from independent standard normal distributions. On the basis of these covariates, success probabilities were determined according to equation (14), where the regression coefficients were assigned values $(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6) = (0.5, 1, 1.5, 2, 2.5, 3)$ and $(\theta_7, \ldots, \theta_{15}) = (0, \ldots, 0)$. Outcome variables y_i were generated from independent Bernoulli distributions with probabilities π_i.

The LRTBF model selection procedure was implemented by using the hyperparameter values $c = 2$ and $\delta = 0.5$. In prior simulations using the resulting selection algorithm, these parameter settings controlled the FDR at 22% and limited the number of false positive results to fewer

than or equal to 4 in 92% of the sampled models when ρ was equal to 0.4. We also implemented the procedure of Lee *et al.* (2003) by assigning independent prior probabilities of 0.4 (the true probability) to the inclusion indicators γ_i and setting $\kappa = 60$. The value of κ in that algorithm plays a role similar to $cn = 60$ in our algorithm. We obtained 100000 MCMC updates of all the regression parameters for each method. Results from this simulation are displayed in Table 4, which lists the estimated marginal posterior probabilities that variables 1–6 were included in a model sampled, along with the corresponding ranks of their marginal inclusion probabilities. In both approaches, variables 3–6 were among the top six highest posterior probability models. Variable 2 was also among the top six variables by using the procedure of Lee *et al.* (2003). Also displayed in Table 4 are the median probability models that were obtained from both procedures. In this case, the LRTBF model selection procedure generated a median probability model that contained fewer false positive results than the method of Lee *et al.* (2003), which selected all the covariates. (We note that smaller median models can be obtained with the method of Lee *et al.* (2003) by assigning a value that was less than 0.4 to the prior probability that a variable appears in the model.) Under our model, the posterior distribution of π had an estimated mean of 0.43, which was slightly larger than the true value of 0.4.

We then repeated the experiment by using more covariates. Specifically, we increased the number of covariates to $p = 50$. The values of the first six regression coefficients were left unchanged, and the remaining coefficients were assigned values of 0. We used the same hyperparameters in the LRTBF method that were used in the previous simulation study (when $p = 15$); these values led to an FDR of 0.5 when $\rho = 6/50$ and led to fewer than nine false positive results in 92% of the models sampled. We implemented the approach of Lee *et al.* (2003) using the same parameter values as before, except that we set $p_i = 6{=}50$ (the true value). Table 4 shows that both procedures identified variables 2, 5 and 6 among the six highest probability models, whereas the method of Lee *et al.* (2003) also included variable 1. The median model that was obtained by using our method contained only variables 5 and 6. The method of Lee *et al.* (2003) tended to assign a higher marginal posterior probability to all models and resulted in a median probability model that contained variables 5, 6 and 24.

5. Convergence diagnostics

In this section, we examine the convergence rate of the Gibbs sampler that was defined in Section 2.1. For brevity, we restrict attention to the sampler that was used to explore the space of models applied to the crime data that were introduced in Section 3.1. The diagnostics that were chosen for this application are based on the coupling methodology that was proposed by Johnson (1996, 1998). The essential idea of this methodology is to examine the number of updates that are required for multiple chains started at distinct values to merge. Although each chain is updated according to the Gibbs sampler that was specified in Section 2.1, correlations are introduced between the random deviates used in the updates across chains to encourage the chains to merge quickly.

The conclusions based on these convergence diagnostics are as follows. If parameter values were thinned so that only every 96th update in a chain was saved, then the total variation distance between two thinned updates and a random sample of size 2 from the target distribution was less than 0.0008. Similarly, the total variation distance between the distribution of updates after 100 complete Gibbs updates and an exact draw from the posterior distribution was less than 0.0002, i.e. burn-in almost certainly required fewer than 100 iterations. These results seem to contradict common thought regarding the number of draws that are required to explore such model spaces, which in this setting had a dimension of $2^{15} = 32768$. Published accounts of MCMC-based model selection procedures appear to imply that several hundred thousand

updates are required to overcome burn-in or to obtain independent draws from the posterior distribution.

6. Conclusions

We have proposed a model selection procedure that can be applied to linear and generalized linear models. The innovation of our method lies in its use of TBFs to select between nested models. The resulting procedure substantially reduces the methodological burden that is associated with the use of traditional Bayesian model selection methods by eliminating the requirement to specify proper prior distributions on regression and variance parameters, which offers substantial gains in computational efficiency. Our method also facilitates the selection of unknown model hyperparameters through prior simulation of model operating characteristics. In the future, we plan to apply this methodology to other generalized linear models (e.g. ordinal, gamma and Poisson models), as well as to nominal regression models, semiparametric linear models, generalized linear models and generalized additive models. More generally, the methodology would seem to have application in settings involving nested models for which the asymptotic distribution of the LRS can be defined under both the null hypothesis and for a suitable class of alternative models.

Acknowledgments

We thank Jennifer Hoeting for her assistance in fitting the MC^3 and Occam's razor procedures and an Associate Editor and three referees for their numerous helpful comments. This research work was partially supported by the US National Science Foundation grants DMS-0706818 and NIH R01 RGM080503A.

References

Atkinson AC. Posterior probabilities for choosing a regression model. Biometrika 1978;65:39–48.

Bahadur, RR. An optimal property of the likelihood ratio statistic. In: LeCam, LM.; Neyman, J., editors. Proc. 5th Berkeley Symp. Mathematical Statistics and Probability; Berkeley: University of California Press; 1965. p. 13-26.

Banerjee M. Likelihood ratio tests under local alternatives in regular semiparametric models. Statist Sin 2005;15:635–644.

Barbieri M, Berger JO. Optimal predictive model selection. Ann Statist 2004;32:870–897.

Cordeiro GM, Botter DA, Ferrari S. Nonnull asymptotic distributions of three classic criteria in generalised linear models. Biometrika 1994;81:709–720.

Davidson RR, Lever WE. The limiting distribution of the likelihood ratio statistic under a class of local alternative. Sankhya A 1970;32:209–224.

Draper D. Assessment and propagation of model uncertainty (with discussion). J R Statist Soc B 1995;57:45–97.

George EI, McCulloch RE. Variable selection via Gibbs sampling. J Am Statist Ass 1993;88:881–889.

George EI, McCulloch RE. Approaches for Bayesian variable selction. Statist Sin 1997;7:339–374.

Hodges JS. Uncertainty, policy analysis and statistics (with discussion). Statist Sci 1987;2:259–291.

Johnson VE. Studying convergence of Markov chain Monte Carlo algorithms using coupled sampling paths. J Am Statist Ass 1996;91:154–166.

Johnson VE. A coupling-regeneration scheme for diagnosing convergence in Markov chain Monte Carlo algorithms. J Am Statist Ass 1998;93:238–248.

Johnson VE. Bayes factors based on test statistics. J R Statist Soc B 2005;67:689–701.

Johnson VE. Properties of Bayes factors based on test statistics. Scand J Statist 2008;35:354–368.

Leamer, EE. Specification Searches. New York: Wiley; 1978.

Lee KE, Sha N, Dougherty ER, Vannucci M, Mallick BK. Gene selection: a Bayesian variable selection approach. Bioinformatics 2003;19:90–97. [PubMed: 12499298]

Lempers, FB. Posterior Probabilities of Alternative Linear Models. Rotterdam: Rotterdam University Press; 1971.

Liang, F.; Paulo, R.; Molina, G.; Clyde, MA.; Berger, JO. Discussion Paper 05-12. Department of Statistical Science, Duke University; Durham: 2005. Mixtures of g-priors for Bayesian variable selection. (Available from http://ftp.stat.duke.edu/WorkingPapers/05-12.html

Madigan D, Raftery AE. Model selection and accounting for model uncertainty in graphical models using Occam's window. J Am Statist Ass 1994;89:1535–1546.

Mallows CL. Some comments on C_p. Technometrics 1973;15:661–676.

Miller, AJ. Subset Selection in Regression. New York: Chapman and Hall; 1990.

Mitchell TJ, Beauchamp JJ. Bayesian variable selection in linear regression (with discussion). J Am Statist Ass 1988;83:1023–1036.

Nishii R. Asymptotic properties of criteria for selection of variables in multiple regression. Ann Statist 1984;12:758–765.

Pericchi LR. An alternative to the standard Bayesian procedure for discrimination between normal linear models. Biometrika 1984;71:575–586.

Raftery AE. Approximate Bayes factors and accounting for model uncertainty in generalised linear models. Biometrika 1996;83:251–266.

Raftery, AE.; Hoeting, JA.; Volinsky, C.; Painter, I.; Yeung, KY. Bayesian Model Averaging Version 3.03. Seattle: University of Washington; 2006. (Available from: http://cran.r-project.org.)

Raftery AE, Madigan D, Hoeting JA. Bayesian model averaging for linear regression models. J Am Statist Ass 1997;92:179–191.

Schwarz G. Estimating the dimension of a model. Ann Statist 1978;6:461–464.

Smith AFM, Spiegelhalter DJ. Bayes factors and choice criteria for linear models. J R Statist Soc B 1980;42:213–220.

Spiegelhalter DJ, Smith AFM. Bayes factors for linear and log-linear models with vague prior information. J R Statist Soc B 1982;44:377–387.

Stewart L. Hierarchical Bayesian analysis using Monte Carlo integration: computing posterior distributions when there are many possible models. Statistician 1987;36:211–219.

Taniguchi M. Third-order asymptotic properties of a class of test statistics under a local alternative. J Multiv Anal 1991;37:223–238.

Vandaele, W. Partipation in illegitimate activities; Ehrlich revisited. In: Blumstein, A.; Cohen, J.; Nagin, D., editors. Deterrence and Incapacitation. Washington DC: National Academy of Sciences; 1978. p. 270-335.

Weisberg, S. Applied Linear Regression. Vol. 2. New York: Wiley; 1985.

Zellner, A. Posterior odds ratios for regression hypotheses: general considerations and some specific results. In: Zellner, A., editor. Basic Issues in Econometrics. Chicago: University of Chicago Press; 1984. p. 275-305.

Appendix A: Regularity conditions

The following regularity assumptions are assumed in Davidson and Lever (1970), for almost all $\mathbf{x} \in S$ and all $\boldsymbol{\theta} \in \Theta$ and $r, s, t = 1, \ldots, p$.

- **a.** $\partial \ln(f)/\partial\theta_r$, $\partial^2 \ln(f)/\partial\theta_r\partial\theta_s$ and $\partial^3 \ln(f)/\partial\theta_r\,\partial\theta_s\partial\theta_t$ exist.
- **b.** $|\partial \ln(f)/\partial\theta_r| < F_r(\mathbf{x})$ and $|\partial^2 \ln(f)/\partial\theta_r\partial\theta_s| < F_{rs}(\mathbf{x})$ where $F_r(\mathbf{x})$ and $F_{rs}(\mathbf{x})$ are integrable over S.
- **c.** The matrix $\mathbf{C} = \{C_{rs}(\boldsymbol{\theta})\}$ with elements

$$C_{rs}(\theta) = E_\theta[\,\partial \ln(f)/\partial\theta_r|_\theta \partial \ln(f)/\partial\theta_s|_\theta]$$

is positive definite with a finite determinant.

d.

$$\left|\frac{\partial^3 \ln(f)}{\partial\theta_r \partial\theta_s \partial\theta_t}\Big|_{\theta}\right| < H_{rst}(\mathbf{x})$$

where there is an $M > 0$ such that $E_{\boldsymbol{\theta}}[H_{rst}(\mathbf{x})] < \mathrm{M} < \infty$, and $\kappa, L > 0$ such that $E_{\boldsymbol{\theta}}[|H_{rst}(\mathbf{x})] - E[H_{rst}(\mathbf{x})]|^{1+\kappa}] < L < \infty$.

e. There are $v, T > 0$ such that, whenever $\|\theta'' - \theta'\| \equiv \sum_1^k |\theta_r'' - \theta_r'| < v, \theta'', \theta' \in \Theta$, $\boldsymbol{\theta''}, \boldsymbol{\theta'} \in \Theta$,

$$E_{\theta'}\left[\left\{\frac{\partial^3 \ln(f)}{\partial\theta_r \partial\theta_s \partial\theta_t}\Big|_{\theta'}\right\}^2\right] < T < \infty.$$

f. There are $\eta, K > 0$ such that

$$E_{\theta}\left[\left|\frac{\partial \ln(f)}{\partial\theta_r}\Big|_{\theta}\right|^{2+\eta}\right] < K < \infty.$$

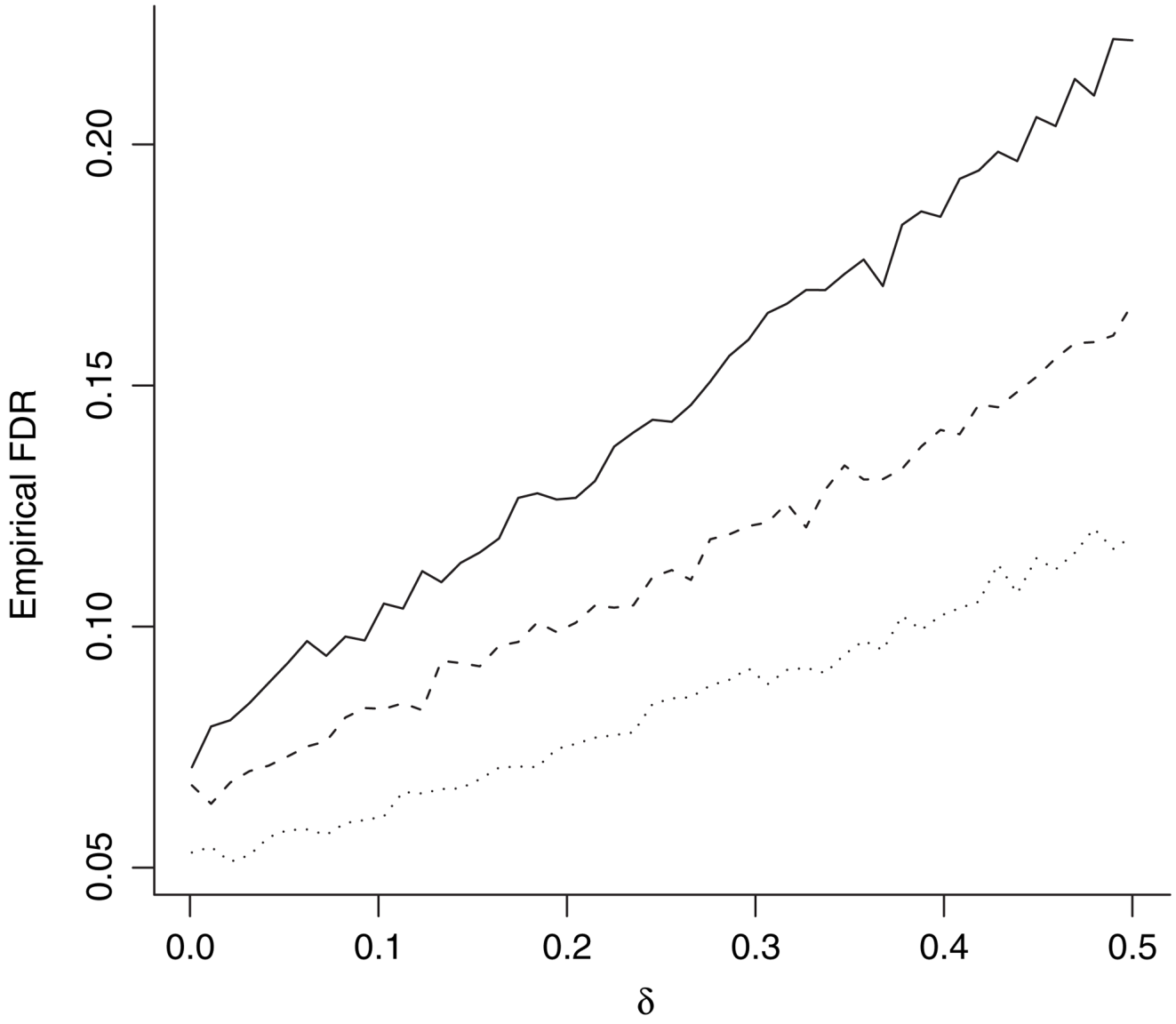

Fig. 1.
FDR *versus* δ:——, $\rho = 1/3$; - - - - -, $\rho = 7/15$; ·····, $\rho = 3/5$

Table 1

Crime rate variables

Predictor	Description of variable
1	Percentage of males 14–24 years
2	Indicator variable for southern state
3	Mean years of schooling
4	Police expenditure in 1960
5	Police expenditure in 1959
6	Labour force participation rate
7	Number of males per 1000 females
8	State population
9	Number of non-whites per 1000 people
10	Unemployment rate of urban males 14–24 years
11	Unemployment rate of urban males 35–39 years
12	Wealth
13	Income inequality
14	Probability of imprisonment
15	Average time served in state prisons

Table 2

Highest posterior probability models for Bayesian model selection procedures

Method	Model variables chosen									Posterior probability (%)
Occam's window	1	3	4		9	11	13	14		12.6
(OR = 20; maxCol = 30)	1	3	4			11	13	14		9.0
(OR.fix = 2; nbest = 150)	1	3	4		9		13	14		8.4
	1	3		5	9	11	13	14		8.0
		3	4	8	9		13	14		7.6
	1	3	4				13	14		6.3
Median model	1	3	4		9		13	14		
MC^3	1	3	4		9	11	13	14		2.6
($\nu = 2.58$; $\lambda = 0.28$)	1	3	4			11	13	14		1.8
($\varphi = 2.85$; $a = 0.05$)	1	3	4		9		13	14		1.7
	1	3	4	5	9		13	14		1.6
	1	3	4		9	11	13	14	15	1.6
	1	3	4		9		13	14	15	1.6
Median model	1	3	4	5	9		13	14		
LRTBF	1	3	4		9	11	13	14		3.1
($c = 2$)	1	3	4		9	11	13	14	15	2.6
	1	3	4			11	13	14		2.2
	1	3		5	9	11	13	14		1.9
	1	3	4		9		13	14	15	1.5
	1	3	4	8	9	11	13	14		1.5
Median model	1	3	4		9	11	13	14		
SSVS	1	3	4		9		13	14		2.0
$(\sigma_{\theta_i}/\tau_i, c_i) = (1, 8)$	1	3	4		9		13	14	15	1.6
($\nu = 0$, $\lambda = 1$)		3	4	8	9		13	14		1.2
($p_i = 0.5$)	1	3	4		9	11	13	14		1.2
($\mathbf{R} = (\mathbf{X}'\mathbf{X})^{-1}$)	1	3	4	8	9		13	14		1.1
	1	3	4				13	14		1.1
Median model	1	3	4		9	11	13	14		
F-statistic TBF	1	3	4			11	13	14		2.3
($\tau = 2$)	1	3	4		9	11	13	14		1.9
	1	3	4			11	13			1.8
	1	3	4				13	14		1.7
	1	3	4				13			1.7
	1	3		5		11	13	14		1.6
Median model	1	3	4				13	14		

Table 3

Marginal probabilities that variables are included in sampled models[†]

Predictor	Marginal probabilities (×100) for the following methods:						
	Occam's window	**MC^3**	**SSVS**	**LRTBF (c = 2)**	**Stepwise regression**	**Mallows's C_p**	**Adjusted R^2**
1	73 (4)	79 (4)	50 (7)	81 (4)	‡	‡	‡
2	2 (10)	17 (12)	42 (9)	21 (12)			
3	99 (2)	98 (2)	50 (6)	95 (2)	‡	‡	‡
4	64 (5)	72 (5)	66 (3)	69 (5)	‡	‡	‡
5	36 (8)	50 (7)	37 (10)	40 (8)			
6	0 (12)	6 (15)	21 (14)	15 (15)			
7	0 (12)	7 (14)	15 (15)	17 (14)			‡
8	12 (9)	23 (10)	47 (8)	32 (10)			‡
9	53 (6)	62 (6)	60 (4)	64 (6)	‡	‡	‡
10	0 (12)	11 (13)	24 (13)	20 (13)			
11	43 (7)	45 (8)	51 (5)	58 (7)	‡	‡	‡
12	1 (11)	30 (9)	30 (11)	31 (11)		‡	‡
13	100 (1)	100 (1)	90 (1)	100 (1)	‡	‡	‡
14	83 (3)	83 (3)	85 (2)	83 (3)	‡	‡	‡
15	0 (15)	22 (11)	27 (12)	33 (9)		‡	‡

[†]The rank of each inclusion probability within each model is listed in parentheses.

‡Variable selected by the frequentist criteria.

Table 4

$\Pr(\beta_i \neq 0)$ in binary data simulations, $i = 1, 2, \ldots, 6$

Method	Model	β_1	β_2	β_3	β_4	β_5	β_6
p = 15							
LRTBF	Marginal probability	0.16	0.22	0.44	0.78	0.82	0.99
	Rank	13	8	5	3	2	1
	Median			4,5,6			
Lee *et al.* (2003)	Marginal probability	0.58	0.83	0.88	0.93	0.99	1
	Rank	15	6	4	3	2	1
	Median			1–15			
p = 50							
LRTBF	Marginal probability	0.20	0.26	0.22	0.18	0.58	0.80
	Rank	9	5	7	14	2	1
	Median			5,6			
Lee *et al.* (2003)	Marginal probability	0.43	0.45	0.40	0.31	0.72	0.92
	Rank	6	5	7	9	2	1
	Median			5,6,24			