Objective Bayesian Variable Selection

George CASELLA and Elías MORENO

A novel fully automatic Bayesian procedure for variable selection in normal regression models is proposed. The procedure uses the posterior probabilities of the models to drive a stochastic search. The posterior probabilities are computed using intrinsic priors, which can be considered default priors for model selection problems; that is, they are derived from the model structure and are free from tuning parameters. Thus they can be seen as objective priors for variable selection. The stochastic search is based on a Metropolis–Hastings algorithm with a stationary distribution proportional to the model posterior probabilities. The procedure is illustrated on both simulated and real examples.

KEY WORDS: Intrinsic prior; Metropolis–Hastings algorithm; Monte Carlo Markov chain methods; Normal linear regression.

1. INTRODUCTION

In this article the variable selection problem in normal regression models is analyzed from an objective Bayesian model choice perspective. A dependent random variable Y and a set $\{X_1, \ldots, X_k\}$ of k potential explanatory regressors are considered. It is assumed that every regression model with regressors $\{X_{i_1}, \ldots, X_{i_q}\}$, where $q = 0, 1, \ldots, k$ and $\{i_1, \ldots, i_q\}$ is a combination of the set of indices $\{1, \ldots, k\}$, is a priori a plausible model to explain the variable Y. The problem consists of choosing one of the foregoing alternative models based on the information provided by a sample $(\mathbf{y}, \mathbf{X}_1, \ldots, \mathbf{X}_k)$. Typically, if interest is in prediction rather than model choice, then the prediction can be taken to be a convex combination of the predictions under every model, where the weights are the model posterior probabilities. Hence in such a case there seems to be no selection problem. But when it is impossible to compute every model (as in Example 4), then before doing any model averaging, we must first select a number of good models to average. Thus prediction will be preceded by model selection in some cases.

We focus on two distinct aspects of the model selection problem. First, we want the selection mechanism to be *criterion-based* and *fully automatic*. A criterion-based selection mechanism allows us to clearly understand the properties of the selected models; that is, we are selecting models that perform well on the criterion. A fully automatic algorithm eliminates the need for specification of tuning parameters, hyperparameters, and other aspect, which makes it easy to implement and eliminates the need for a sensitivity analysis.

Second, we note that the model selection problem is fundamentally a problem of multiple-hypothesis testing, making it important to exactly specify the hypotheses to be tested at each model evaluation. Because we are typically looking for a reduced model that adequately explains the data, the evaluation of model M_R should be

$$\begin{aligned} H_0 &: M_R = \text{a reduced model} \qquad \text{versus} \\ H_A &: M = \text{the full model}, \end{aligned} \tag{1}$$

where the full model is the model with all predictor variables. Thus the full model is taken to be the overall reference model. This is the correct approach to take because the full model is the one given to us by the subject matter. We should assume that all of the predictors have some importance, and our task is to examine whether a smaller subset is adequate. Also, because we are going to use a Bayesian method to evaluate (1), it is important that the prior distribution be centered at H_0 and be specific to each null model M_R under consideration.

Two difficulties arise in computing the model posterior probabilities. First, with respect to the prior distribution on the parameters in each model, because we are not confident about any given set of regressors as explanatory variables, little prior information on their regression coefficients can be expected. (If we were confident about a particular model, then there would be no model selection problem!) This argument alone justifies the need for an objective model choice approach in which vague prior information is assumed. Hence within each model we would like to consider default prior distributions on the regression coefficients and the error variance. Unfortunately, default priors for the normal regression parameters are improper and thus cannot be used for either model choice or prediction in the presence of alternative models.

Nonobjective Bayesian variable selection has a long history, having been considered by Atkinson (1978), Smith and Spiegelhalter (1980), Pericchi (1984), Poirier (1985), Box and Meyer (1986), George and McCulloch (1993, 1995, 1997), Clyde, DeSimone, and Parmigiani (1996), Geweke (1996), and Smith and Kohn (1996), among others. The proposed prior distributions on the regression coefficients and the error variance within each model are typically either conjugate priors or some closely related distributions. That is, multivariate normal distributions are usually considered for the regression coefficients and inverse gammas are usually considered for the error variances. It is also typical to center the normal at 0, so that in the underlying testing problem the role of the null hypothesis is played by the model with no regressors. The covariance matrices and the hyperparameters in the inverse gamma often are fixed with the help of some subjective/empirical criteria. More recently, Brown, Vanucci, and Fearn (2002) considered a Bayesian model selection formulation implemented with a variation of a random-walk Metropolis–Hastings algorithm. They used the results of their stochastic search to minimize a decision-theoretic criterion for future prediction.

Some attempts at solving the problem in a form as "objective as possible" were made by Mitchell and Beauchamp (1988) and Spiegelhalter and Smith (1982). Mitchell and Beauchamp

George Casella is Distinguished Professor and Chair, Department of Statistics, University of Florida, Gainesville, FL 32611 (E-mail: *casella@stat.ufl.edu*). Elías Moreno is Professor, Department of Statistics, University of Granada, 18071 Granada, Spain (E-mail: *emoreno@ugr.es*). Supported by National Science Foundation grant DMS-99-71586 and supported by Ministerio de Ciencia y Tecnología grant BEC2001-2982. This work was done while Casella was on sabbatical at the University of Granada. The authors thank the editor, associate editor, and three referees for their careful reading of earlier versions of this article. Their thoughtful, and thought-provoking, comment and suggestions greatly improved the presentation.

Journal of the American Statistical Association
March 2006, Vol. 101, No. 473, Theory and Methods
DOI 10.1198/016214505000000646

(1988) assumed that regression coefficients were a priori independent and identically distributed with a prior distribution that concentrates some probability mass on 0 and distributes the rest uniformly on a compact set. Conventional improper priors were used for the error variance. Their variable selection problem is essentially an estimation problem that avoids the difficulties with improper priors but needs some criteria to specify the point masses. Spiegelhalter and Smith (1982) used conventional improper priors for the regression coefficients, but with the error variance and the arbitrary constants determined using subjective information on the value of the ratio of marginal densities at a sample point. We note that the Bayes factor involved in their analysis does not satisfy the strict definition of a Bayes factor.

A fully automatic analysis for model comparison in regression was given by Berger and Pericchi (1996). They used an encompassing approach and an empirical measure for model comparison—the intrinsic Bayes factor—which does not depend on any subjective information. For moderate or large sample sizes, this empirical measure closely approximates a Bayes factor for the so-called *intrinsic priors*.

In this article we derive a model choice solution based on intrinsic priors. The procedure that we use is as follows. For each reduced model M_R, we consider the pair $\{M_R, M\}$ and calculate the model posterior probability $\Pr(M_R|data)$. Then we order all reduced models with respect to their posterior probability. The interpretation is that the model with highest posterior probability represents the most plausible reduction in complexity of the full model, the second highest represents the second-most plausible model, and so on.

In computing $\Pr(M_R|data)$, we assume a priori that $\Pr(M_R) = \Pr(M) = 1/2$, and use the intrinsic priors for the parameters of the null model M_R and the full model M. Justifications for using the intrinsic priors have been given by Berger and Pericchi (1996, 1997a,b, 1998) and Moreno (1997). The method provides sensible priors for a wide variety of model selection problems involving nested models (see Berger and Pericchi 1996, 1998; Girón, Martínez, Moreno, and Torres 2003; Moreno and Liseo 2003; Moreno, Bertolino, and Racugno 1998, 1999, 2000; Moreno, Torres, and Casella 2005). We use the intrinsic priors because (a) the intrinsic prior distribution for the parameters of the full model takes into account the null hypothesis M_R, (b) they are automatically derived from the models, (c) there are no hyperparameters to be adjusted, (d) for any sample size they provide either Bayes factors or statistics that are as close as we want to Bayes factors, and (e) model posterior probabilities are easily computed.

But when k is not small, the number of competing models, say 2^{k-1} (the intercept is always included), is huge and precludes the calculation of all model posterior probabilities. Therefore, to avoid all of these calculations, a search for models having "high" posterior probability is needed. This is a second difficulty that we deal with here.

Modern search algorithms for variable selection were first developed by George and McCulloch (1993), using an approach based on the Gibbs sampler (see Chipman, George, and McCulloch 2001 for more recent developments). The novel idea of this approach was to use a stochastic search algorithm to visit models having high probability, and the advantages are that rather than returning the "best" model, a ranking of models is obtained, and there is some assurance that the search will not get stuck in local modes. But models are not ranked according to any criterion, and the relationship between the frequently selected models and models that are optimal against some criterion is not clear.

Here, because the models are ranked according to their posterior probabilities, we require that a stochastic search based on a Markov chain should have a stationary distribution that is proportional to the model posterior probabilities. This can be accomplished by using a Metropolis–Hastings algorithm. We develop such an algorithm to search through the model space.

The article is organized as follows. In Section 2 we formulate the models, and in Section 3 we derive intrinsic priors for variable selection and formulas for computing model posterior probabilities. We detail the stochastic search in Section 4, where we develop an independent Metropolis–Hastings algorithm with stationary distribution proportional to the posterior probabilities. We provide illustrations of the method using both simulated and real data in Section 5, and give concluding remarks in Section 6. We provide some technical details in two appendixes.

2. EVALUATING THE MODELS

Consider the standard normal regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \boldsymbol{\varepsilon},$$

where $\mathbf{y} = (y_1, \ldots, y_n)^t$ is the vector of observations, $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_k]$ is the $n \times k$ design matrix, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)^t$ is the $k \times 1$ column vector of the regression coefficients, and $\boldsymbol{\varepsilon}$ is an error vector distributed as $\boldsymbol{\varepsilon} \sim \mathrm{N}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$, where the error variance σ^2 is a nuisance parameter. This is the full model for $\mathbf{y}$ and is denoted by $\mathrm{N}_n(\mathbf{y}|\mathbf{X}\boldsymbol{\alpha}, \sigma^2\mathbf{I}_n)$.

2.1 Hypothesis Tests

Let $\boldsymbol{\gamma}$ denote a vector of length k with components equal to either 0 or 1, and let $\mathbf{Q}_{\boldsymbol{\gamma}}$ denote a $k \times k$ diagonal matrix with the elements of $\boldsymbol{\gamma}$ on the leading diagonal and 0 elsewhere. Because we want to include the intercept in every model, the first component of each $\boldsymbol{\gamma}$ is equal to 1. We let Γ denote the set of 2^{k-1} different configurations of $\boldsymbol{\gamma}$.

When some of the components of $\boldsymbol{\alpha}$ are 0, the meanings of the remaining components, and of the error variance, change, and we will change the notation accordingly (see Berger and Pericchi 1996; Clyde 2001). Therefore, when $\mathbf{y}$ is assumed to be explained by a given subset of regressors, the sampling model is written as $\mathrm{N}_n(\mathbf{y}|\mathbf{X}\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2_{\boldsymbol{\gamma}}\mathbf{I}_n)$, where $\boldsymbol{\beta}_{\boldsymbol{\gamma}} = \mathbf{Q}_{\boldsymbol{\gamma}}\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ is a configuration to be interpreted as

$$\gamma_i = \begin{cases} 0 & \text{if } \alpha_i = 0 \\ 1 & \text{otherwise} \end{cases}$$

for $i = 1, \ldots, k$. Hence the set of sampling models to be considered is $\{\mathrm{N}_n(\mathbf{y}|\mathbf{X}\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2_{\boldsymbol{\gamma}}\mathbf{I}_n), \boldsymbol{\gamma} \in \Gamma\}$.

To complete the specification of each model, we take a prior on the parameters $(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma_{\boldsymbol{\gamma}})$, so we have the Bayesian model

$$M_{\boldsymbol{\gamma}} : \left\{\mathrm{N}_n(\mathbf{y}|\mathbf{X}\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2_{\boldsymbol{\gamma}}\mathbf{I}_n), \pi(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma_{\boldsymbol{\gamma}})\right\}, \qquad \boldsymbol{\gamma} \in \Gamma.$$

For a set of data $(\mathbf{y}, \mathbf{X})$, the Bayes factor of a generic model $M_{\boldsymbol{\gamma}}$, when compared with the full model $M_{\mathbf{1}}$, is given by the ratio of marginal distributions

$$B_{\boldsymbol{\gamma}\mathbf{1}}(\mathbf{y}, \mathbf{X}) = \frac{m_{\boldsymbol{\gamma}}(\mathbf{y}, \mathbf{X})}{m_{\mathbf{1}}(\mathbf{y}, \mathbf{X})} = \frac{\int \mathrm{N}_n(\mathbf{y}|\mathbf{X}\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma_{\boldsymbol{\gamma}}^2\mathbf{I}_n)\pi(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma_{\boldsymbol{\gamma}})\, d\boldsymbol{\beta}_{\boldsymbol{\gamma}}\, d\sigma_{\boldsymbol{\gamma}}}{\int \mathrm{N}_n(\mathbf{y}|\mathbf{X}\boldsymbol{\alpha}, \sigma^2\mathbf{I}_n)\pi(\boldsymbol{\alpha}, \sigma)\, d\boldsymbol{\alpha}\, d\sigma}, \quad (2)$$

where $\boldsymbol{\gamma} = \mathbf{1}$ corresponds to the full model. For $\Pr(M_{\boldsymbol{\gamma}}) = \Pr(M_{\mathbf{1}}) = 1/2$, the posterior probability of $M_{\boldsymbol{\gamma}}$ is

$$\Pr(M_{\boldsymbol{\gamma}}|\mathbf{y}, \mathbf{X}) = \frac{B_{\boldsymbol{\gamma}\mathbf{1}}(\mathbf{y}, \mathbf{X})}{1 + B_{\boldsymbol{\gamma}\mathbf{1}}(\mathbf{y}, \mathbf{X})}, \qquad \boldsymbol{\gamma} \in \Gamma.$$

We note that $\Pr(M_{\boldsymbol{\gamma}}|\mathbf{y}, \mathbf{X})$ is an increasing function of $B_{\boldsymbol{\gamma}\mathbf{1}}(\mathbf{y}, \mathbf{X})$. Therefore, ranking the models according to their posterior probabilities $\{\Pr(M_{\boldsymbol{\gamma}}|\mathbf{y}, \mathbf{X}),\ \boldsymbol{\gamma} \in \Gamma\}$ is the same as ranking them according to their Bayes factors $\{B_{\boldsymbol{\gamma}\mathbf{1}}(\mathbf{y}, \mathbf{X}),\ \boldsymbol{\gamma} \in \Gamma\}$. Furthermore, the ordering is not altered by normalizing the Bayes factor against all models. This yields the set of probabilities

$$\Pr_c(M_{\boldsymbol{\gamma}}|\mathbf{y}, \mathbf{X}) = \frac{B_{\boldsymbol{\gamma}\mathbf{1}}(\mathbf{y}, \mathbf{X})}{1 + \sum_{\boldsymbol{\gamma}\in\Gamma, \boldsymbol{\gamma}\neq\mathbf{1}} B_{\boldsymbol{\gamma}\mathbf{1}}(\mathbf{y}, \mathbf{X})}, \qquad \boldsymbol{\gamma} \in \Gamma. \quad (3)$$

This forms the basis of our evaluation of model $M_{\boldsymbol{\gamma}}$ through the hypothesis test

$$H_0: \quad M = M_{\boldsymbol{\gamma}} \qquad \text{versus} \qquad M = M_{\mathbf{1}},$$

with (3) measuring the relative support for H_0. The advantage of this construction is that for the multiple tests carried out, each of the individual tests can be centered at its null (see Lemma 1 in Sec. 3.1), and the full model is taken to be the overall reference model.

2.2 Default Priors

To remove subjectivity from the choice of $\pi(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma_{\boldsymbol{\gamma}})$, and to make our procedure automatic, we want to use some type of default or "automatic" prior. We first consider the standard default prior on parameters $(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma_{\boldsymbol{\gamma}})$, giving the Bayesian model

$$M_{\boldsymbol{\gamma}} : \left\{\mathrm{N}_n(\mathbf{y}|\mathbf{X}\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma_{\boldsymbol{\gamma}}^2\mathbf{I}_n), \pi^N(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma_{\boldsymbol{\gamma}}) = c_{\boldsymbol{\gamma}}/\sigma_{\boldsymbol{\gamma}}^2\right\}, \qquad \boldsymbol{\gamma} \in \Gamma,$$

where $c_{\boldsymbol{\gamma}}$ is an arbitrary positive constant that cannot be determined because the integral of $\pi^N(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma_{\boldsymbol{\gamma}})$ is infinite.

Using this default prior leads to the expression (2) with marginal distribution

$$m_{\boldsymbol{\gamma}}^N(\mathbf{y}, \mathbf{X}) = c_{\boldsymbol{\gamma}} \int \frac{\mathrm{N}_n(\mathbf{y}|\mathbf{X}\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma_{\boldsymbol{\gamma}}^2\mathbf{I}_n)}{\sigma_{\boldsymbol{\gamma}}^2}\, d\boldsymbol{\beta}_{\boldsymbol{\gamma}}\, d\sigma_{\boldsymbol{\gamma}}$$

and Bayes factor $B_{\boldsymbol{\gamma},\mathbf{1}}^N(\mathbf{y}, \mathbf{X})$ defined up to the multiplicative constant $c_{\boldsymbol{\gamma}}/c_{\mathbf{1}}$. Hence the posterior probability of model $M_{\boldsymbol{\gamma}}$ is not uniquely defined when the default prior is used. A solution to this problem, that will still yield an automatic procedure is given in the next section.

3. INTRINSIC PRIORS

The intrinsic methodology was introduced by Berger and Pericchi (1996) to overcome the difficulty arising with conventional default priors in model selection and hypothesis testing. Here, for completeness and readability, we give a brief summary of these developments.

Suppose that we have two Bayesian models, $M_i : \{f_i(x|\theta_i), \pi_i^N(\theta_i)\}$, $i = 1, 2$, where $f_1(x|\theta_1)$ is nested in $f_2(x|\theta_2)$ and $\pi_i^N(\theta_i)$ are the conventional priors. Because $\pi_i^N(\theta_i)$ is typically improper, we can write $\pi_i^N(\theta_i) = c_i h_i(\theta_i)$, where $h_i(\theta_i)$ is a nonintegrable function and c_i is an arbitrary constant that cannot be determined. Berger and Pericchi (1996) proposed replacing $B_{21}^N(\mathbf{x})$ with the arithmetic intrinsic Bayes factor $B_{21}^{AI}(\mathbf{x})$, a partial Bayes factor that is justified as follows. The sample $\mathbf{x}$ is split into two parts, $(x(\ell), x(n-\ell))$. The part $x(\ell)$, called a *training sample*, is designed to convert the improper prior $\pi_i^N(\theta_i)$ into a proper posterior, that is,

$$\pi_i^N\big(\theta_i|x(\ell)\big) = \frac{f_i(x(\ell)|\theta_i)\pi_i^N(\theta_i)}{m_i^N(x(\ell))}, \qquad i = 1, 2,$$

where $x(\ell)$ is such that $0 < m_i^N(x(\ell)) < \infty$. With the remainder of the data, $x(n-\ell)$, the Bayes factor is computed using $\pi_i^N(\theta_i|x(\ell))$ as the prior. This gives the partial Bayes factor,

$$B_{21}^P(\mathbf{x}) = \frac{\int f_2(x(n-\ell)|\theta_2)\pi_2^N(\theta_2|x(\ell))\, d\theta_2}{\int f_1(x(n-\ell)|\theta_1)\pi_1^N(\theta_1|x(\ell))\, d\theta_1} = B_{21}^N(\mathbf{x})B_{12}^N(x(\ell)).$$

Note that $B_{21}^P(\mathbf{x})$ corrects $B_{21}^N(\mathbf{x})$ with the term $B_{12}^N(x(\ell))$, and that the arbitrary constants c_1 and c_2 cancel out.

It should be noted that for a given sample $\mathbf{x}$, we can consider different training samples $x(\ell)$, and hence there exists a multiplicity of partial Bayes factors, one for each training sample. To avoid dependence on a particular training sample, Berger and Pericchi first suggested considering all possible subsamples $x(\ell)$ for which there is no proper subsample satisfying the inequalities $0 < m_i^N(x(\ell)) < \infty$ for any c_i. They termed this subsample a *minimal training sample*. Second, they considered the arithmetic mean of $B_{21}^P(\mathbf{x})$ for all minimal training samples. This produces the so-called "arithmetic intrinsic Bayes factor," defined as

$$B_{21}^{AI}(\mathbf{x}) = B_{21}^N(\mathbf{x})\frac{1}{L}\sum_{\ell=1}^{L} B_{12}^N(x(\ell)),$$

where L is the number of minimal training samples contained in the sample. Other ways of "averaging" $B_{21}^P(\mathbf{x})$ are possible, but whereas the arithmetic mean produces priors for model selection, other methods may not necessarily do the same.

Note that $B_{21}^{AI}(\mathbf{x})$ is not a Bayes factor. Furthermore, stability of $B_{21}^{AI}(\mathbf{x})$ is also a matter of concern. Conceivably, for a given sample $\mathbf{x}$, the number of minimal training samples might be small, and minor changes in the data could cause this number to vary substantially. Moreover, the equality $B_{21}^{AI}(\mathbf{x}) = 1/B_{12}^{AI}(\mathbf{x})$ is not necessarily satisfied, so the coherent equality $\Pr(M_1|\mathbf{x}) = 1 - \Pr(M_2|\mathbf{x})$ does not hold. To be coherent, it is important to know whether $B_{21}^{AI}(\mathbf{x})$ corresponds to a Bayes factor for sensible priors. If so, then consistency of the $B_{21}^{AI}(\mathbf{x})$ is ensured. With the so-called *intrinsic* priors, the foregoing question has been

asymptotically answered. There are priors $\pi_1^I(\theta_1)$ and $\pi_2^I(\theta_2)$ for which the corresponding Bayes factor,

$$B_{21}(\mathbf{x}) = \frac{\int_{\Theta_2} f_2(\mathbf{x}|\theta_2)\pi_2^I(\theta_2)\,d\theta_2}{\int_{\Theta_1} f_1(\mathbf{x}|\theta_1)\pi_1^I(\theta_1)\,d\theta_1}$$

and $B_{21}^{AI}(\mathbf{x})$ are asymptotically equivalent under the two models M_1 and M_2. Note that if we use intrinsic priors for computing the Bayes factor instead of the improper priors that we started from, then coherency is ensured. By equating the limit of $B_{21}^{AI}(\mathbf{x})$ and $B_{21}(\mathbf{x})$ as $n \to \infty$ under the two models, Berger and Pericchi showed that intrinsic priors satisfy the functional equations

$$E^{M_1}_{x(\ell)|\theta_1} B_{12}^N(x(\ell)) = \frac{\pi_2^I(\psi_2(\theta_1))}{\pi_2^N(\psi_2(\theta_1))}\frac{\pi_1^N(\theta_1)}{\pi_1^I(\theta_1)}$$

and

$$E^{M_2}_{x(\ell)|\theta_2} B_{12}^N(x(\ell)) = \frac{\pi_2^I(\theta_2)}{\pi_2^N(\theta_2)}\frac{\pi_1^N(\psi_1(\theta_2))}{\pi_1^I(\psi_1(\theta_2))}.$$

The expectations in these equations are taken with respect to $f_1(x(\ell)|\theta_1)$ and $f_2(x(\ell)|\theta_2)$; $\psi_2(\theta_1)$ denotes the limit of the maximum likelihood estimator $\hat{\theta}_2(\mathbf{x})$ under model M_1 at point θ_1, and $\psi_1(\theta_2)$ denotes the limit of $\hat{\theta}_1(\mathbf{x})$ under model M_2 at point θ_2.

For nested models, as is our case in variable selection, the foregoing functional equations collapse into a single equation. Although the solution (π_1^I, π_2^I) to this single equation is not unique, and the resulting class is not robust (Moreno 1997), a sensible selection is to take the conditional intrinsic prior for θ_2 as

$$\pi_2^I(\theta_2|\theta_1) = \pi_2^N(\theta_2) E^{M_2}_{x(\ell)|\theta_2} \frac{f_1(x(\ell)|\theta_1)}{\int_{\Theta_2} f_2(x(\ell)|\theta_2)\pi_2^N(\theta_2)\,d\theta_2}$$

and the intrinsic prior for θ_1 as $\pi_1^I(\theta_1) = \pi_1^N(\theta_1)$. (For a justification for choosing this pair, see Moreno et al. 1998.)

3.1 Intrinsic Priors for Variable Selection

For each configuration $\boldsymbol{\gamma}$, the Bayes factor $B_{\boldsymbol{\gamma}\mathbf{1}}^N(\mathbf{y}, \mathbf{X})$ compares the model $\{\mathrm{N}_n(\mathbf{y}|\mathbf{X}\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma_{\boldsymbol{\gamma}}^2\mathbf{I}_n), \pi^N(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma_{\boldsymbol{\gamma}})\}$ with the full model $\{\mathrm{N}_n(\mathbf{y}|\mathbf{X}\boldsymbol{\alpha}, \sigma^2\mathbf{I}_n), \pi^N(\boldsymbol{\alpha}, \sigma)\}$. Because the sampling model $\mathrm{N}_n(\mathbf{y}|\mathbf{X}\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma_{\boldsymbol{\gamma}}^2\mathbf{I}_n)$ is nested in $\mathrm{N}_n(\mathbf{y}|\mathbf{X}\boldsymbol{\alpha}, \sigma^2\mathbf{I}_n)$ we can apply the intrinsic method to derive intrinsic priors for comparing model $M_{\boldsymbol{\gamma}}$ and $M_{\mathbf{1}}$, for any $\boldsymbol{\gamma} \in \Gamma$.

We first take an arbitrary but fixed point $(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma_{\boldsymbol{\gamma}})$ in the null space, and then find the intrinsic prior for $(\boldsymbol{\alpha}, \sigma)$ conditional on $(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma_{\boldsymbol{\gamma}})$. To do this, we note that a theoretical minimal training sample for this problem is a random vector $\mathbf{y}^{ts}$ of dimension $k+1$ such that it is $\mathrm{N}_{k+1}(\mathbf{y}^{ts}|\mathbf{Z}^{ts}\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma_{\boldsymbol{\gamma}}^2\mathbf{I}_n)$ distributed under the null model and is $\mathrm{N}_{k+1}(\mathbf{y}^{ts}|\mathbf{Z}^{ts}\boldsymbol{\alpha}, \sigma^2\mathbf{I}_n)$ distributed under the full model. Here $\mathbf{Z}^{ts}$ represents a $(k+1)\times k$ unknown design matrix associated with $\mathbf{y}^{ts}$, to which we return in Section 3.2.

Therefore, application of the standard intrinsic method yields the formal expression for the conditional intrinsic prior,

$$\pi^I(\boldsymbol{\alpha}, \sigma|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma_{\boldsymbol{\gamma}}) = \pi^N(\boldsymbol{\alpha}, \sigma) \times E_{\mathbf{y}^{ts}|\boldsymbol{\alpha},\sigma} \frac{\mathrm{N}_{k+1}(\mathbf{y}^{ts}|\mathbf{Z}^{ts}\beta_{\boldsymbol{\gamma}}, \sigma_{\boldsymbol{\gamma}}^2\mathbf{I}_n)}{\int \mathrm{N}_{k+1}(\mathbf{y}^{ts}|\mathbf{Z}^{ts}\alpha, \sigma^2\mathbf{I}_n)\pi^N(\boldsymbol{\alpha}, \sigma)\,d\boldsymbol{\alpha}\,d\sigma}.$$

The resulting distribution is given in the following lemma.

Lemma 1. The intrinsic prior for parameters $\boldsymbol{\alpha}$ and σ conditional on a fixed point $(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma_{\boldsymbol{\gamma}})$ is given by

$$\pi^I(\boldsymbol{\alpha}, \sigma|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma_{\boldsymbol{\gamma}}) = \mathrm{N}_k\big(\boldsymbol{\alpha}|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, (\sigma_{\boldsymbol{\gamma}}^2 + \sigma^2)\mathbf{W}^{-1}\big)\frac{1}{\sigma_{\boldsymbol{\gamma}}}\left(1 + \frac{\sigma^2}{\sigma_{\boldsymbol{\gamma}}^2}\right)^{-3/2}, \quad (4)$$

where $\mathbf{W} = \mathbf{Z}^t\mathbf{Z}$ and $\mathbf{Z}$ is a theoretical design matrix of dimensions $(k+1)\times k$. The unconditional intrinsic prior for $\boldsymbol{\alpha}$ and σ, obtained by integrating out $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ and $\sigma_{\boldsymbol{\gamma}}$ against $\pi^N(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma_{\boldsymbol{\gamma}})$, is

$$\pi^I(\boldsymbol{\alpha}, \sigma) = c_{\boldsymbol{\gamma}} \int \mathrm{N}_k\big(\boldsymbol{\alpha}|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, (\sigma_{\boldsymbol{\gamma}}^2 + \sigma^2)\mathbf{W}^{-1}\big) \times \frac{1}{\sigma_{\boldsymbol{\gamma}}^3}\left(1 + \frac{\sigma^2}{\sigma_{\boldsymbol{\gamma}}^2}\right)^{-3/2} d\boldsymbol{\beta}_{\boldsymbol{\gamma}}\,d\sigma_{\boldsymbol{\gamma}}.$$

For the proof see Appendix A.

We note that $\pi^I(\boldsymbol{\alpha}, \sigma|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma_{\boldsymbol{\gamma}})$ is, by construction, a probability density. In (4) it is factored as

$$\pi^I(\boldsymbol{\alpha}, \sigma|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma_{\boldsymbol{\gamma}}) = \pi^I(\boldsymbol{\alpha}|\sigma, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma_{\boldsymbol{\gamma}})\pi^I(\sigma|\sigma_{\boldsymbol{\gamma}}),$$

so that the intrinsic prior for $\boldsymbol{\alpha}$, conditional on the null $(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma_{\boldsymbol{\gamma}})$, depends on the nuisance parameter σ. The distribution of σ depends only on $\sigma_{\boldsymbol{\gamma}}$. The marginal of $\boldsymbol{\alpha}$,

$$\pi^I(\boldsymbol{\alpha}|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma_{\boldsymbol{\gamma}}) = \int \mathrm{N}_k\big(\boldsymbol{\alpha}|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, (\sigma_{\boldsymbol{\gamma}}^2 + \sigma^2)\mathbf{W}^{-1}\big)\frac{1}{\sigma_{\boldsymbol{\gamma}}}\left(1 + \frac{\sigma^2}{\sigma_{\boldsymbol{\gamma}}^2}\right)^{-3/2} d\sigma,$$

is an elliptical multivariate distribution with mean vector $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$. Therefore, the intrinsic prior for $\boldsymbol{\alpha}$ is centered at the null, which seems a natural requirement for a sharp null hypothesis (Morris 1987). This requirement is not fulfilled for many of the variable selection priors proposed in the literature.

Second-order or higher-order moments of $\boldsymbol{\alpha}$ do not exist, because the second-order moment of the mixing distribution is infinite. This implies that the intrinsic prior for $\boldsymbol{\alpha}$, conditional on the null $(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma_{\boldsymbol{\gamma}})$, has a very heavy tail as expected for a default prior.

The pair $\{\pi^N(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma_{\boldsymbol{\gamma}}), \pi^I(\boldsymbol{\alpha}, \sigma)\}$ is called the intrinsic prior for comparing $M_{\boldsymbol{\gamma}}$ and $M_{\mathbf{1}}$, and although they are improper, they are well calibrated because both depend on the same arbitrary constant $c_{\boldsymbol{\gamma}}$. Further, they are a well-established limit of proper priors (Moreno et al. 1998).

3.2 Evaluating the Intrinsic Prior and the Model Posterior Probabilities

The matrix $\mathbf{W}^{-1}$ in (4) is defined by the regressors of a theoretical training sample of minimal size for the full model $k+1$. A way of assessing $\mathbf{W}^{-1}$ is to use the original idea of the arithmetic intrinsic Bayes factor (Berger and Pericchi 1996). This entails averaging over all possible training samples of minimal size contained in the sample. This would give the matrix

$$\mathbf{W}^{-1} = \frac{1}{L}\sum_{\ell=1}^{L}\big(\mathbf{Z}^t(\ell)\mathbf{Z}(\ell)\big)^{-1},$$

where $\{\mathbf{Z}(\ell), \ell = 1, \ldots, L\}$ is the set of all submatrices of $\mathbf{X}$ of order $(k+1) \times k$ of rank k.

For the data $(\mathbf{y}, \mathbf{X})$, the Bayes factor for comparing models $M_{\boldsymbol{\gamma}}$ and $M_{\mathbf{1}}$ with the intrinsic priors $\{\pi^N(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma_{\boldsymbol{\gamma}}), \pi^I(\boldsymbol{\alpha}, \sigma)\}$ has the formal expression

$$B_{\boldsymbol{\gamma}\mathbf{1}}(\mathbf{y}, \mathbf{X}) = \int \mathrm{N}_n(\mathbf{y}|\mathbf{X}\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma_{\boldsymbol{\gamma}}^2\mathbf{I}_n)\pi^N(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma_{\boldsymbol{\gamma}})\, d\boldsymbol{\beta}_{\boldsymbol{\gamma}}\, d\sigma_{\boldsymbol{\gamma}} \times \left(\int \mathrm{N}_n(\mathbf{y}|\mathbf{X}\boldsymbol{\alpha}, \sigma^2\mathbf{I}_n)\pi^I(\boldsymbol{\alpha}, \sigma|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma_{\boldsymbol{\gamma}}) \times \pi^N(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma_{\boldsymbol{\gamma}})\, d\boldsymbol{\beta}_{\boldsymbol{\gamma}}\, d\sigma_{\boldsymbol{\gamma}}\, d\boldsymbol{\alpha}\, d\sigma \right)^{-1},$$

and from this expression, it clearly does not depend on any arbitrary constant. Computation of this Bayes factor is quite simple, as we show in the next lemma, whose proof is straightforward (and hence is omitted).

In what follows we partition the design matrix $\mathbf{X}$ as $\mathbf{X} = (\mathbf{X}_{0\boldsymbol{\gamma}}|\mathbf{X}_{1\boldsymbol{\gamma}})$, where $\mathbf{X}_{1\boldsymbol{\gamma}}$ contains the column j of $\mathbf{X}$ if the configuration $\boldsymbol{\gamma}$ is such that $\gamma_j = 1$. Therefore, the dimension of $\mathbf{X}_{1\boldsymbol{\gamma}}$ is $n \times k_{\boldsymbol{\gamma}}$, where $k_{\boldsymbol{\gamma}} = \sum_{i=1}^{k} \gamma_i$.

Lemma 2. The Bayes factor for comparing models

$$M_{\boldsymbol{\gamma}} : \left\{\mathrm{N}_n(\mathbf{y}|\mathbf{X}\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma_{\boldsymbol{\gamma}}^2\mathbf{I}_n), \pi^N(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma_{\boldsymbol{\gamma}})\right\}$$

and

$$M_{\mathbf{1}} : \left\{\mathrm{N}_n(\mathbf{y}|\mathbf{X}\boldsymbol{\alpha}, \sigma^2\mathbf{I}_n), \pi^I(\boldsymbol{\alpha}, \sigma)\right\}$$

is given by

$$B_{\boldsymbol{\gamma}\mathbf{1}}(\mathbf{y}, \mathbf{X}) = \left(|\mathbf{X}_{1\boldsymbol{\gamma}}^t\mathbf{X}_{1\boldsymbol{\gamma}}|^{1/2}\left(\mathbf{y}^t(\mathbf{I}_n - \mathbf{H}_{\boldsymbol{\gamma}})\mathbf{y}\right)^{(n-k_{\boldsymbol{\gamma}}+1)/2} I_{\boldsymbol{\gamma}}\right)^{-1}, \quad (5)$$

where $\mathbf{H}_{\boldsymbol{\gamma}} = \mathbf{X}_{1\boldsymbol{\gamma}}(\mathbf{X}_{1\boldsymbol{\gamma}}^t\mathbf{X}_{1\boldsymbol{\gamma}})^{-1}\mathbf{X}_{1\boldsymbol{\gamma}}^t$,

$$I_{\boldsymbol{\gamma}} = \int_0^{\pi/2} \frac{d\varphi}{|\mathbf{A}_{\boldsymbol{\gamma}}(\varphi)|^{1/2}|\mathbf{B}(\varphi)|^{1/2}E_{\boldsymbol{\gamma}}(\varphi)^{(n-k_{\gamma}+1)/2}},$$

$$\mathbf{B}(\varphi) = (\sin^2\varphi)\mathbf{I}_n + \mathbf{X}\mathbf{W}^{-1}\mathbf{X}^t,$$

$$\mathbf{A}_{\boldsymbol{\gamma}}(\varphi) = \mathbf{X}_{1\boldsymbol{\gamma}}^t\mathbf{B}^{-1}(\varphi)\mathbf{X}_{1\boldsymbol{\gamma}},$$

and

$$E_{\boldsymbol{\gamma}}(\varphi) = \mathbf{y}^t\left(\mathbf{B}^{-1}(\varphi) - \mathbf{B}^{-1}(\varphi)\mathbf{X}_{1\boldsymbol{\gamma}}\mathbf{A}_{\boldsymbol{\gamma}}^{-1}(\varphi)\mathbf{X}_{1\boldsymbol{\gamma}}^t\mathbf{B}^{-1}(\varphi)\right)\mathbf{y}.$$

Although here we consider only models with an intercept, so that $\gamma_1 = 1$, the comparisons can be extended to consider all models. In this case we need a special calculation for the model with no regressors, which comes as a corollary to Lemma 2.

Corollary 1. The Bayes factor for comparing the model

$$M_{\mathbf{0}} : \left\{\mathrm{N}_n(\mathbf{y}|\mathbf{0}, \sigma_{\boldsymbol{\gamma}}^2\mathbf{I}_n), \pi^N(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma_{\boldsymbol{\gamma}})\right\}$$

and

$$M_{\mathbf{1}} : \left\{\mathrm{N}_n(\mathbf{y}|\mathbf{X}\boldsymbol{\alpha}, \sigma^2\mathbf{I}_n), \pi^I(\boldsymbol{\alpha}, \sigma)\right\},$$

where $M_{\mathbf{0}}$ corresponds to the simplest model with no regressor, is given by

$$B_{\mathbf{01}}(\mathbf{y}, \mathbf{X}) = \left((\mathbf{y}^t\mathbf{y})^{(n+1)/2}\int_0^{\pi/2} \frac{d\varphi}{|\mathbf{B}(\varphi)|^{1/2}E_{\mathbf{0}}(\varphi)^{(n+1)/2}}\right)^{-1}, \quad (6)$$

where

$$E_{\mathbf{0}}(\varphi) = \mathbf{y}^t\mathbf{B}^{-1}(\varphi)\mathbf{y}.$$

For any $\boldsymbol{\gamma} \in \Gamma$, the probability of $M_{\boldsymbol{\gamma}}$ for the intrinsic priors is

$$\Pr_c(M_{\boldsymbol{\gamma}}|\mathbf{y}, \mathbf{X}) = \frac{B_{\boldsymbol{\gamma}\mathbf{1}}(\mathbf{y}, \mathbf{X})}{1 + \sum_{\boldsymbol{\gamma}\in\Gamma, \boldsymbol{\gamma}\neq\mathbf{1}} B_{\boldsymbol{\gamma}\mathbf{1}}(\mathbf{y}, \mathbf{X})}, \quad (7)$$

where $B_{\boldsymbol{\gamma}\mathbf{1}}(\mathbf{y}, \mathbf{X})$ is as given in (5) for any $\boldsymbol{\gamma} \neq \mathbf{0}$ and $B_{\mathbf{01}}(\mathbf{D})$ is as given in (6). Simulations presented in Section 5 show that the intrinsic priors behave extremely well.

4. STOCHASTIC SEARCH

We have now developed a mechanism for ranking the candidate models based on their intrinsic posterior probabilities. To now choose the "best" model, or to examine a range of good models, we would like to rank the models by their posterior probabilities. However, except in small problems, this is not possible, because the number of models can be prohibitively large. (For example, for a regression with three independent variables x_1, x_2, and x_3, if the squares and all interactions were included in the models, then there would be 18 predictor variables, not counting the intercept, which would result in $2^{18} = 262{,}144$ models.) Thus a search algorithm is needed.

Because we cannot calculate all of the posterior probabilities, the next best action would be to draw independent samples from a distribution $\Pr(M_{\boldsymbol{\gamma}}|\mathbf{y}, \mathbf{X})$. But this may also be impossible, because it would entail the exhaustive calculation of all of the posterior probabilities.

What is possible is to construct an MCMC algorithm with $\Pr(M_{\boldsymbol{\gamma}}|\mathbf{y}, \mathbf{X})$ as the stationary distribution. Such an algorithm, if properly constructed, would not only visit every model, but also would visit the better models more often. Thus a frequency count of visits to the models is directly proportional to the posterior probabilities.

In theory, constructing of such an algorithm is easy. At iteration t, if the chain is in model $M_{\boldsymbol{\gamma}_t}$, then we draw a candidate model from a candidate distribution $\mathcal{G}$, say $M_{\boldsymbol{\gamma}_t'}$, and move to this new model with probability

$$\min\left\{1, \frac{\Pr(M_{\boldsymbol{\gamma}_t'}|\mathbf{y}, \mathbf{X})\mathcal{G}(M_{\boldsymbol{\gamma}_t})}{\Pr(M_{\boldsymbol{\gamma}_t}|\mathbf{y}, \mathbf{X})\mathcal{G}(M_{\boldsymbol{\gamma}_t'})}\right\}. \quad (8)$$

If the draws from $\mathcal{G}$ are independent, then this is a reversible ergodic Markov chain with stationary distribution $\Pr(M_{\boldsymbol{\gamma}}|\mathbf{y}, \mathbf{X})$. [Note that the denominator of the probability in (7) cancels out in (8) and hence does not have to be calculated. This is good, because in large problems this sum may not be calculable.]

The one difficulty is specifying a good candidate distribution $\mathcal{G}$. Although $\mathcal{G}$ can be completely arbitrary (e.g., can be taken to be uniform on the model space), a completely arbitrary $\mathcal{G}$ will almost certainly will not be a good choice. We want our candidate distribution to be able to adequately explore the entire space, so as to not get trapped in local modes, and to often visit models that have a high posterior probability. To ensure these properties, we construct our candidate distribution in two parts. First, we do the following:

1. Write the set of models as $\mathcal{B} = \bigcup_i \mathcal{B}_i$, where

$$\mathcal{B}_i = \{M_{\boldsymbol{\gamma}} : \boldsymbol{\gamma} = \{1, \boldsymbol{\gamma}'\}, \quad \boldsymbol{\gamma}' \text{ has exactly } i \text{ components equal to } 1\}.$$

2. For fixed $\nu < 1$, select a random sample of size $\nu \times 2^{k-1}$ from $\mathcal{B}$ in such a way that $100 \times \nu$ percent is sampled from each $\mathcal{B}_i$. (The constant ν is selected so that $\nu \times 2^{k-1}$ is a reasonable number.)
3. For the selected models, calculate the posterior probabilities, p_{ij}, of model j in $\mathcal{B}_i$. The probability $\sum_{j\in\mathcal{B}_i} p_{ij} / \sum_{ij} p_{ij}$ is an estimate of the posterior probability of $\mathcal{B}_i$.

The estimated (and true) posterior probabilities can be highly variable, with some close to 0. To ensure that we adequately explore the model space, we add a second term to our candidate distribution to keep the chain "hot" at the beginning of the search [similar in spirit to the simulated tempering of Geyer and Thompson (1995)]. Thus at iteration t, the distribution $\mathcal{G} = (\hat{P}_0, \ldots, \hat{P}_k)$ on the subsets $\mathcal{B}_i$, $i = 0, \ldots, k$, given by

$$\hat{P}_i \propto \frac{1}{k+1}\frac{1}{\log(t+1)} + \sum_{j\in\mathcal{B}_i} p_{ij} \Big/ \sum_{ij} p_{ij}$$

is our candidate distribution. Moreover, at each iteration when we calculate a new Bayes factor, we update the second term. The stochastic search is then done as follows:

1. At iteration t, choose a candidate model $M_{\boldsymbol{\gamma}_{t'}}$ by first selecting $\mathcal{B}_i$ according to the distribution $\mathcal{G}$ and then selecting $\boldsymbol{\gamma}_{t'}$ at random from $\mathcal{B}_i$.
2. With probability (8), move to $M_{\boldsymbol{\gamma}_{t'}}$.

The candidate distribution estimates the distribution that is proportional to the posterior probabilities and focuses the search on subsets that have higher posterior probabilities. However, it moves around randomly in the subsets, allowing for good mixing. As we show in the next section, the search algorithm performs quite well.

5. EXAMPLES

In this section we look at a number of examples to examine the behavior of the intrinsic posterior probabilities and the search algorithm. Example 1 verifies that the true model typically will have the maximum posterior probability. Examples 1 and 2 show that the posterior probabilities are a reasonable tool for finding the true model, and that the stochastic search finds models with high posterior probabilities. Examples 3 and 4 apply our method to some real data, illustrating its usefulness.

Example 1. In this example we consider the full model to be

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \varepsilon,$$

where $\varepsilon \sim \mathrm{N}(0, \sigma^2)$. The x_i values are generated uniformly in the interval (0, 10), and we included the squares to make it a bit more difficult to find the true model.

We simulated 1,000 datasets, each with $n = 10$ observations, using two different true models. In the first case (Table 1), we generated data from the model

$$y = \beta_0 + \beta_1 x_1 + \varepsilon, \tag{9}$$

with $\beta_0 = \beta_1 = 1$, and calculated the posterior probabilities of all $2^4 = 16$ models.

The performance of the intrinsic posterior probability is quite good, as shown in Table 1. The true model had an average posterior probability of .41 and was selected 57.9% of the time. When the true model was not chosen, the alternates seemed quite reasonable. In contrast, the performance of Mallows' C_p was not that good. It chose the correct model only 22.6% of the time and when it did not choose the correct model, it tended to choose more complex models. For example, it chose the full model 9% of the time, and chose models with three regressors almost 25% of the time.

In the second case (Table 2) we generated data from the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon, \tag{10}$$

with $\beta_0 = \beta_1 = 1$ and $\beta_2 = 2$. The performance, in terms of correctness of selection, is similar. We want to point out here that the posterior probability selections tend to favor the smaller models, in contrast to C_p, which tends toward larger models in its incorrect choice.

Table 1. For Example 1, Average Posterior Probabilities, Percentage of Maximum Bayes Factor, and Percentage of Maximum Mallows' C_p, Corresponding to the Case n= 10, $\sigma = 2$ With True Model (9), 1,000 Simulations

Variables in model	Average posterior probability	Standard deviation	Percent maximum Bayes factor	Percent maximum C_p
x_1	.41	.213	.579	.226
x_1^2	.259	.182	.250	.106
Intercept only	.0969	.168	.104	<.001
x_1, x_1^2	.048	.0904	.016	.087
x_1, x_2^2	.0398	.06	.015	.077
x_1, x_2	.0387	.0489	.008	.062
x_1, x_2, x_2^2	.0264	.0407	<.001	.045
x_2, x_1^2	.0257	.0389	.003	.036
x_2	.0123	.0313	.004	.004
x_2^2	.0121	.0258	.002	<.001

NOTE: All models contain an intercept term.

Table 2. For Example 1, Average Posterior Probabilities, Percentage of Maximum Bayes Factor, and Percentage of Maximum Mallows' C_p, Corresponding to the Case $n = 10$, $\sigma = 2$ With True Model (10), 1,000 Simulations

Variables in model	Average posterior probability	Standard deviation	Percent maximum Bayes factor	Percent maximum C_p
x_1, x_2	.309	.208	.419	.259
x_2, x_1^2	.215	.182	.220	.171
x_1, x_2^2	.11	.143	.097	.068
x_2	.0961	.155	.104	.004
x_1^2, x_2^2	.0787	.109	.054	.045
x_2^2	.0594	.112	.053	.003
x_1, x_2, x_1^2	.0355	.0741	.014	.146
x_1, x_2, x_2^2	.0292	.0594	.008	.114
x_2, x_1^2, x_2^2	.0205	.041	.008	.080
x_2, x_2^2	.0157	.0554	.007	.005

NOTE: All models contain an intercept term.

Example 2. Our next example illustrates the effectiveness of the stochastic search and the performance of the candidate distribution construction. We consider the 10-predictor model

$$y = \beta_0 + \sum_{i=1}^{3} \beta_i x_i + \sum_{i=1}^{3} \tau_i x_i^2 + \sum_{i>j} \eta_{ij} x_i x_j + \eta_{123} x_1 x_2 x_3 + \varepsilon, \quad (11)$$

where $\varepsilon \sim \mathrm{N}(0, \sigma^2)$. The x_i values are generated uniformly in the interval $(0, 10)$, and the true model is (10). There are $2^{10} = 1{,}024$ candidate models, and to check our search algorithm, we calculated all of the Bayes factors.

Two cases are considered, corresponding to $\sigma = 2$ and $\sigma = 5$. In the case of $\sigma = 2$, the true model, which has two predictors, has posterior probability .449, almost equal to the total posterior probability of all of the models with two predictors. In fact, other than the true model, there are no other models with a posterior probability $>.080$. Table 3 gives the results of a 10,000-iteration search; it can be seen that the algorithm performs extremely well, spending almost all of its time in the true model.

In contrast, in the case of $\sigma = 5$, the posterior probabilities of the models are more equal. For the same model as (11) but with increased variance, the true model has a posterior probability of only .136, with competing models having posterior probabilities somewhat closer than the case of $\sigma = 2$. Table 3 gives the posterior probabilities for the top five models and the performance of the stochastic search. The true model is the one that the search visits most, although the differences are not as dramatic as in the case of $\sigma = 2$.

Table 3. For Example 2, True Posterior Probabilities and Percentage of Visits to the Top Five Models, Corresponding to the Cases $n = 10$, $\sigma = 2$ or 5 With True Model (10), 1,000 Simulations

Model	Posterior probability	Proportion of visits
$\sigma = 2$		
x_1, x_2	.449	.860
x_1, x_2, x_1^2	.080	.035
x_1, x_2, x_1x_2	.070	.008
x_1, x_2, x_2^2	.040	.011
x_1, x_2, x_3^2	.035	.013
$\sigma = 5$		
x_1, x_2	.136	.188
x_1, x_2^2	.075	.112
x_2	.063	.075
x_1^2, x_2	.051	.076
x_2^2	.045	.065

NOTE: All models include the intercept term and are sorted by the true posterior probabilities.

Example 3. For our next example, we look at an ancient and oft-analyzed dataset, the Hald data. Although it is painful to once again consider these data, doing so provides a comparison of our procedure to other related procedures.

The Hald dataset comes from an experiment of the effect of heat on the composition of cement (Wood, Steinour, and Starke 1932) and contains 13 observations on the dependent variable (heat) and 4 predictor variables (which relate to the composition of the cement). Because there are only $2^4 = 16$ possible models, there is no need for a stochastic search, because we can calculate all of the posterior probabilities. These are given in Table 4.

We compare Table 4 with the findings of Berger and Pericchi (1996), who compared these models using intrinsic Bayes factors and encompassing models. Based on pairwise comparisons, they concluded that "$\{x_1, x_2\}$ is moderately preferred to $\{x_1, x_4\}$ and quite strongly preferred to $\{x_3, x_4\}$." Any conclusion derived from Table 4 is certainly in concert with these findings; in fact, we would declare a strong preference for $\{x_1, x_2\}$ over both of those other models. Table 4 also agrees with the conclusions of Draper and Smith (1981, sec. 6.1), who concluded, based on R^2, that the favored models are $\{x_1, x_2\}$ and

Table 4. Posterior Probabilities for the Models of the Hald Data

Variables	Posterior probability
x_1, x_2	.5224
x_1, x_4	.1295
x_1, x_2, x_3	.1225
x_1, x_2, x_4	.1098
x_1, x_3, x_4	.0925
x_2, x_3, x_4	.0120
x_1, x_2, x_3, x_4	.0095
x_3, x_4	.0013

NOTE: All other models had posterior probability $<.00001$.

$\{x_1, x_4\}$, with their preference being $\{x_1, x_4\}$ (based on the fact that x_4 is the single best predictor). Although $\{x_1, x_2, x_4\}$ also had a high R^2, Draper and Smith argued that because of the high correlation between X_2 and X_4 ($r_{24} = -.973$), the variable X_4 should not be added to the equation. A similar high correlation exists between X_1 and X_3.

George and McCulloch (1993) also analyzed the Hald dataset. Their method is based on the posterior distribution of the vector of configurations $\boldsymbol{\gamma}$, and their results depend on the chosen values for the hyperparameters of their prior distributions on the regression coefficients, the variance errors, and the prior on $\boldsymbol{\gamma}$. Some of their chosen hyperparameter values favor the model with no regressors (intercept only) followed by the model with only one regressor, and visited the model $\{x_1, x_2\} < 7\%$ of the time. This possibly illustrates a shortcoming of the George and McCulloch algorithm, in that it takes the model with no regressors (not even an intercept) as the null model in all comparisons. With this model as the reference, the intercept-only model (which fits the mean) can appear very significant. However, for other hyperparameter configurations, their stochastic search procedure identified the model $\{x_1, x_2\}$ as one of the best models.

Example 4. Our last example uses the ozone data analyzed by Breiman and Friedman (1985). (See App. B for a description of this dataset.) Using the ACE algorithm, Breiman and Friedman identified a set of four predictors $\{x_7, x_8, x_9, x_{10}\}$ that yielded (for their data; our dataset is somewhat different) an R^2 of .78 when the maximum R^2 with all predictors is .79. Most recently, Breiman (2001) remarked that in the 1980s, large linear regressions were run, using squares and interaction terms, with the goal of selecting a good prediction model. However, the project was not successful, because the false-alarm rate was too high. We revisit this model selection problem to assess how the intrinsic prior stochastic search will perform.

Preliminary examination of the data reveals a high degree of multicollinearity, causing ill-conditioned design matrices. The variables [temperature (°F) measured at El Monte, CA and inversion base temperature (°F) at LAX] are highly correlated with a number of other predictors, and are deleted from the predictor set.

Thus we have 10 predictor variables, and we first perform a variable selection on the set of models made up of only the 10 predictors. In this case there are $2^{10} = 1{,}024$ models, so an exhaustive calculation of posterior probabilities is possible—we do not need to do a stochastic search. Of the $n = 203$ observations, we randomly selected 25 to hold out of the fitting set to used to assess prediction, so we used 178 observations to fit the models. The results are reported in Table 5.

The intrinsic-prior posterior probabilities identified a number of good models and stayed more with simple models (fewer predictors) than with complex models. Also, the top five models all contain variable x_7, which Breiman and Friedman identified as "the most influential predictor variable." We also note that all of the models picked out by the intrinsic-prior posterior probabilities had better prediction mean squared error than that of Breiman and Friedman, as well as better R^2 [and close to the full model (linear predictors only) R^2 of .708].

Next, we try to improve prediction by taking the full model to be all linear, quadratic, and two-way interactions, giving us

Table 5. Posterior Probabilities for the Models for the Ozone Data Using Only Linear Predictors

Variables	Posterior probability	R^2	Average prediction error
x_6, x_7, x_8	.491	.686	.992
x_1, x_6, x_7, x_8, x_{10}	.156	.699	.974
x_1, x_6, x_7, x_8, x_9	.041	.696	.972
x_1, x_6, x_7, x_8	.028	.691	.964
x_1, x_4, x_6, x_7, x_8	.027	.694	.968
x_7, x_8, x_9, x_{10}	<.00001	.669	1.056

NOTE: All other models had posterior probability <.00001. The final column is the square root of the mean of the squared prediction errors of the 25 observations held out of the fitting set. The last model (x_7, x_8, x_9, x_{10}) is the one chosen by the Breiman–Friedman ACE algorithm.

$10 + 10 + 45 = 65$ predictors and $2^{65} = 36, 893, 488, 147, 419, 103, 232$ models. Enumeration of the posterior probabilities is no longer possible, and we now do a stochastic search as outlined in Section 4. We ran our search for 50,000 iterations, and found three models that were visited most often by the search. The performance of these model is summarized in Table 6.

A number of other models were visited with frequencies in the .02–.09 range, but the three given in Table 6 are the best performers. It is interesting that the search found such a simple model as $\{x_1x_9, x_1x_{10}, x_4x_6, x_5x_8, x_6x_7\}$ to be one of the best performers. Moreover, we see that the models tend to use variables x_7–x_{10} more often than the other variables. Both the R^2 and prediction errors of the three models in Table 6 are very good, with the third model being an exceptionally good predictor. Figure 1 shows a scatterplot of the actual ozone and the predicted ozone, using the prediction data, for both the original model of Breiman and Friedman and the third model in Table 6. It turns out that the third model alone does a slightly better job of prediction than the average of the top three models. From Figure 1, we see that the third model is quite an improvement over the original model (especially for the lower ozone concentrations) and somewhat (but not totally) alleviates the problem of overprediction. Overall, the prediction error is reduced by 22%.

6. DISCUSSION

We have presented two distinct parts of a method to select a linear model in regression, where each part can function independently. First, we wanted to construct an objective Bayesian criterion for model selection, which was accomplished using an intrinsic prior to calculate posterior probabilities. Second, the model selection criterion was used to direct a search algorithm that was based on a Markov chain with stationary distribution proportional to the criterion.

It should be clear that either part of our method can be used in other settings. For example, we can use other priors to calculate the posterior probabilities for model selection, and can use other criteria (e.g., R^2) to direct the stochastic search.

We also note that the approach that we take here with respect to model comparisons is fundamentally different from that of George and coworkers. When evaluating a model $M_{\boldsymbol{\gamma}}$ against $M_{\mathbf{1}}$, we are effectively testing the null hypothesis H_0: the true model is $M_{\boldsymbol{\gamma}}$ against the alternative H_A: the true model is $M_{\mathbf{1}}$, using a prior centered on the null hypothesis. Thus each candidate model is the null model for its comparison, and the comparison is against the full model in which all models are

Table 6. Posterior Probabilities for the Top Three Models for the Ozone Data Using All Linear, Quadratic, and Two-Way Interactions

Variables	Proportion of visits	R^2	Average prediction error
$\{x_2, x_1^2, x_7^2, x_9^2, x_1x_5, x_2x_6, x_3x_7, x_4x_6, x_6x_8, x_6x_{10}\}$	.214	.758	.873
$\{x_1x_9, x_1x_{10}, x_4x_6, x_5x_8, x_6x_7\}$	.122	.718	.908
$\{x_6, x_5^2, x_7^2, x_9^2, x_1x_{10}, x_4x_7, x_4x_8, x_5x_{10}, x_6x_8\}$	.114	.748	.818

NOTE: The final column is the square root of the mean of the squared prediction errors of the 25 observations held out of the fitting set.

nested. In comparison, George and coworkers tested the hypotheses H_0: the true model has $\boldsymbol{\alpha} = \mathbf{0}$ against H_A: the true model is $M_{\boldsymbol{\gamma}}$. Although this testing setup has worked well in practical problems, we believe that testing against a model with all zero regressors is not appropriate, and rather each candidate model should be considered as its own null.

We reiterate our point from Section 1. Model selection is a multiple testing problem in which we test whether any possible reduction in complexity of the full model is plausible. We can do this with the method presented here and obtain a complete ranking of the models (in smaller problems), or identify a suite of acceptable models (in larger problems). Unless the comparisons are set up to allow this, simultaneous model comparisons may not be possible (as with the encompassing model approach; see Moreno 2005).

Using the Metropolis–Hastings algorithm to drive a stochastic search, especially one that is based on intrinsic posterior probabilities, is a strategy that has not had much investigation. Some reviewers of this work expressed concerns about the properties of the procedure, and we now address those concerns.

Intrinsic Priors and Variable Selection. Are the intrinsic priors good tools for variable selection? Intrinsic priors are currently the unique available objective priors for variable selection. In fact, default priors typically used for estimating model parameters are improper, and thus they are not suitable for computing model posterior probabilities. The commonly used "vague priors," a limit of a sequence of conjugate priors, is typically an ill-defined notion.

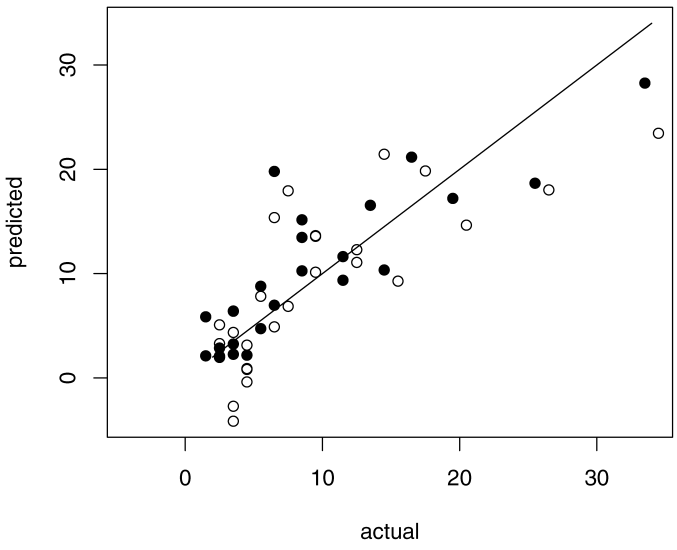

Figure 1. Scatterplot of Observed versus Predicted, Using the Prediction Dataset, for the Breiman–Friedman Model (∘) and the Third Model of Table 6 (•).

In contrast, intrinsic priors are well defined, depend on the sampling model structure, and do not contain any tuning parameters to be adjusted. This last point is important because variable selection is quite sensitive to the values assigned to the hyperparameters. (See, e.g., the analysis in George and McCullogh 1993 of the Hald data; their results depend heavily on the four sets of hyperparameter values chosen.)

The essence of the variable selection problem is that of reducing the complexity of the full model; hence one wants to do multiple comparisons with different reduced models. The intrinsic prior for the full model parameters is "centered" at the reduced model under consideration (a widely accepted desirable property) and has the expected heavy tails of a default prior.

Stochastic Search With Many Models. When there are very many models and the evaluation of all models becomes unfeasible, is the notion of stochastic search valuable? It is important to realize that the goal of a search is somewhat different from the goal of estimating a posterior distribution (the objective function in our case). We are interested in finding good models from the (sometimes) almost infinite number of candidates, and are less interested in estimating all of the modes. For example, in doing multiple runs we do not get the same models that appear in Table 6. However, all of the good models that we get are superior to the ones that have been found previously. As the goal is to find good predictors, with much less emphasis on the actual form of the model, the goal is satisfied.

This may not be a particularly pleasing answer, but it is a practical one. Brown et al. (2002), in running a stochastic search on a large model space, noted that "it is also possible to find promising $\boldsymbol{\gamma}$-vectors (the vectors that define the submodel) even though one has explored a very small fraction of the space of 2^p possibilities."

Algorithm Efficiency. Is the proposed algorithm efficient? The stochastic search is a straightforward application of the independent Metropolis–Hastings algorithm, an algorithm with excellent convergence properties. In implementation, one difficulty arises in choosing an appropriate candidate distribution. In large-scale searches the candidate has to be able to both find states with large values of the criterion, and to escape from modes to be able to adequately explore the space.

The two-part construction proposed here seems to be able to do this. Moreover, an independent Metropolis algorithm does not suffer from some of the disadvantages of a random-walk algorithm, in that it does not have a tendency to get stuck in local modes. We are also using a changing candidate distribution, one that is adapting to look more like the target. This also

increases efficiency while preserving ergodicity. (Although the candidate distribution changes at each iteration, each candidate results in a kernel that satisfies the detailed balance condition with the same stationary distribution, resulting in an ergodic Markov chain.)

The fact remains, however, that we are searching large complex spaces, and cannot hope to find every "good" model. Moreover, as a reviewer noted, the complexity of the space is a factor in determining how many iterations to run. Rigorous evaluations of these algorithms, in terms of convergence and mixing, are quite difficult to do and, for the most part, have not been done (Jerrum and Sinclair 1996). These same authors also make the point that careful application of the Metropolis–Hastings algorithm is among the better strategies available.

APPENDIX A: PROOF OF LEMMA 1

Consider a theoretical training sample for the full model, say $\{y_j, x_{1j}, \ldots, x_{kj}, j = 1, \ldots, k+1\}$, that is, the vector $\mathbf{y} \sim \mathrm{N}_{k+1}(\mathbf{y}|\mathbf{Z}\boldsymbol{\alpha}, \sigma^2\mathbf{I}_n)$. The intrinsic prior for $\boldsymbol{\alpha}, \sigma$ conditional on $\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma_{\boldsymbol{\gamma}}$ is given by

$$\pi^I(\boldsymbol{\alpha}, \sigma|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma_{\boldsymbol{\gamma}}) = \pi^N(\boldsymbol{\alpha}, \sigma) E_{\mathbf{y}|\boldsymbol{\alpha},\sigma} B_{\mathbf{1}\boldsymbol{\gamma}}(\mathbf{y}),$$

where the expectation is taken with respect to the density $\mathrm{N}_{k+1}(\mathbf{y}|\mathbf{Z}\boldsymbol{\alpha}, \sigma^2\mathbf{I}_n)$. This expectation is easily found with the help of the equality

$$\begin{aligned}\int_{\mathbb{R}^n} \left(\mathbf{y}^t\mathbf{K}\mathbf{y}\prod_{i=1}^{2}\mathrm{N}_n(\mathbf{y}|\mathbf{X}\boldsymbol{\theta}_i, \sigma_i^2\mathbf{I}_n)\right) d\mathbf{y}\\ = \frac{\sigma_2^2\,\mathrm{tr}(\mathbf{K})|\mathbf{X}^t\mathbf{X}|^{-1/2}}{(2\pi\sigma_1^2)^{(n-k)/2}(1+\sigma_2^2/\sigma_1^2)^{(n-k+2)/2}}\\ \times \mathrm{N}_k\big(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1, (\sigma_1^2+\sigma_2^2)(\mathbf{X}^t\mathbf{X})^{-1}\big),\end{aligned}$$

where $\mathbf{K}$ is an $n \times n$ symmetric matrix, $\mathbf{X}$ is a $n \times k$ matrix of rank k such that $\mathbf{K}\mathbf{X} = \mathbf{0}$, and $\boldsymbol{\theta}_i$ are k-dimensional vectors for $i = 1, 2$.

APPENDIX B: DESCRIPTION OF OZONE DATA

This is a description of the ozone data taken from the BLSS data library. The dataset comprises Los Angeles ozone pollution data from 1976, where each observation is 1 day. There were 366 observations on 13 variables. Because of missing data, only 203 cases were used here. The full data are available at *http://www.fmrp.usp.br/augusto/ps/breiman/breiman.html*.

Two predictor variables that were in the original dataset—temperature (degrees F) measured at El Monte, CA and inversion base temperature (degrees F) at LAX—were not used here because of multicollinearity considerations.

[Received October 2002. Revised March 2005.]

REFERENCES

Atkinson, A. C. (1978), "Posterior Probabilities for Choosing a Regression Model," *Biometrika*, 65, 39–48.

Berger, J. O., and Pericchi, L. R. (1996), "The Intrinsic Bayes Factor for Model Selection and Prediction," *Journal of the American Statistical Association*, 91, 109–122.

——— (1997a), "On the Justification of Default and Intrinsic Bayes Factor," in *Modeling and Prediction*, eds. J. C. Lee et al., New York: Springer-Verlag, pp. 276–293.

——— (1997b), "Accurate and Stable Bayesian Model Selection: The Median Intrinsic Bayes Factor," *Sankhyā*, Ser. B, 60, 1–18.

——— (1998), "On Criticism and Comparison of Default Bayes Factors for Model Selection and Hypothesis Testing," in *Proceedings of the Workshop on Model Selection*, ed. W. Racugno, Bologna: Pitagora, pp. 1–50.

Box, G. E. P., and Meyer, R. D. (1986), "An Analysis for Unreplicated Fractional Factorial," *Technometrics*, 28, 11–18.

Breiman, L. (2001), "Statistical Modeling: The Two Cultures" (with discussion), *Statistical Science*, 16, 199–231.

Breiman, L., and Friedman, J. (1985), "Estimating Optimal Transformations for Multiple Regression and Correlation," *Journal of the American Statistical Association*, 80, 580–598.

Brown, P. J., Vanucci, M., and Fearn, T. (2002), "Bayes Model Averaging With Selection of Regressors," *Journal of the Royal Statistical Society*, Ser. B, 64, 519–536.

Chipman, H., George, E., and McCulloch, R. E. (2001), *The Practical Implementation of Bayesian Model Selection*, Hayward, CA: IMS, pp. 67–116.

Clyde, M. (2001), Discussion of *The Practical Implementation of Bayesian Model Selection*, by H. Chipman, E. George, and R. E. McCulloch, in *Model Selection*, Hayward, CA: IMS, pp. 117–124.

Clyde, M., DeSimone, H., and Parmigiani, G. (1996), "Prediction via Orthogonalized Model Mixing," *Journal of the American Statistical Association*, 91, 1197–1208.

Draper, N., and Smith, H. (1981), *Applied Regression Analysis* (2nd ed.), New York: Wiley.

George, E. I., and McCulloch, R. E. (1993), "Variable Selection via Gibbs Sampling," *Journal of the American Statistical Association*, 88, 881–889.

——— (1995), "Stochastic Search Variable Selection," in *Practical Markov Chain Monte Carlo in Practice*, eds. W. R. Gilks et al., London: Chapman & Hall, pp. 339–348.

——— (1997), "Approaches for Variable Selection," *Statistica Sinica*, 7, 339–373.

Geweke, J. (1996), "Variable Selection and Model Comparison in Regression," in *Bayesian Statistics 5*, eds. J. M. Bernardo et al., Oxford University Press, pp. 169–194.

Geyer, C. J., and Thompson, E. A. (1995), "Annealing Markov Chain Monte Carlo With Applications to Ancestral Inference," *Journal of the American Statistical Association*, 90, 909–920.

Girón, F. J., Martínez, M. L., Moreno, E., and Torres, F. (2003), "Bayesian Analysis of Matched Pairs in the Presence of Covariates," in *Bayesian Statistics 7*, eds. J. M. Bernardo et al., Oxford, U.K.: Oxford University Press, pp. 553–563.

Jerrum, M., and Sinclair, A. (1996), "The Markov Chain Monte Carlo Method: An Approach to Approximate Counting and Integration," in *Approximation Algorithms for NP-Hard Problems*, Boston: PWS Publishing, pp. 482–487.

Table B.1. Variables Used in Example 4

Variable	*Description*
y	Response = Daily maximum 1-hour-average ozone reading (ppm) at Upland, CA
x_1	Month: 1 = January, ..., 12 = December
x_2	Day of month
x_3	Day of week: 1 = Monday, ..., 7 = Sunday
x_4	500-millibar pressure height (m) measured at Vandenberg AFB
x_5	Wind speed (mph) at Los Angeles International Airport (LAX)
x_6	Humidity (%) at LAX
x_7	Temperature (°F) measured at Sandburg, CA
x_8	Inversion base height (feet) at LAX
x_9	Pressure gradient (mm Hg) from LAX to Daggett, CA
x_{10}	Visibility (miles) measured at LAX

Mitchell, T. J., and Beauchamp, J. J. (1988), "Bayesian Variable Selection in Linear Regression" (with discussion), *Journal of the American Statistical Association*, 83, 1023–1036.

Moreno, E. (1997), "Bayes Factor for Intrinsic and Fractional Priors in Nested Models: Bayesian Robustness," in L_1-*Statistical Procedures and Related Topics*, ed. D. Yadolah, Hayward, CA: Institute of Mathematical Statistics, pp. 257–270.

——— (2005), "Objective Bayesian Analysis for One-Sided Testing," *Test*, 14, 181–198.

Moreno, E., Bertolino, F., and Racugno, W. (1998), "An Intrinsic Limiting Procedure for Model Selection and Hypotheses Testing," *Journal of the American Statistical Association*, 93, 1451–1460.

——— (1999), "Default Bayesian Analysis of the Behrens–Fisher Problem," *Journal of Statistical Planning and Inference*, 81, 323–333.

——— (2000), "Bayesian Model Selection Approach to Analysis of Variance Under Heteroscedasticity," *Journal of the Royal Statistical Society*, Ser. D, 46, 1–15.

Moreno, E., and Liseo, B. (2003), "A Default Bayesian Test for the Number of Components in a Mixture," *Journal of Statistical Planning and Inference*, 111, 129–142.

Moreno, E., Torres, F., and Casella, G. (2005), "Testing Equality of Regression Coefficients in Heteroscedastic Normal Regression Models," *Journal of Statistical Planning and Inference*, 131, 117–134.

Morris, C. N. (1987), Comment on "Testing of a Point Null Hypothesis: The Irreconcilability of Significance Levels and Evidence," by J. O. Berger and T. Sellke; and "Reonciling Bayesian and Frequentist Evidence in the One-Side Testing Problem," by G. Cassella and R. L. Berger, *Journal of the American Statistical Association*, 82, 106–139.

Pericchi, L. R. (1984), "An Alternative to the Standard Bayesian Procedure for Discrimination Between Normal Linear Models," *Biometrika*, 71, 575–586.

Poirier, D. J. (1985), "Bayesian Hypothesis Testing in Linear Models With Continuously Induced Conjugate Prior Across Hypotheses," in *Bayesian Statistics 2*, eds. J. M. Bernardo et al., New York: Elsevier, pp. 711–722.

Smith, M., and Kohn, R. (1996), "Nonparametric Regression Using Bayesian Variable Selection," *Journal of Econometrics*, 75, 317–344.

Smith, A. F. M., and Spiegelhalter, D. J. (1980), "Bayes Factor and Choice Criteria for Linear Models," *Journal of the Royal Statistical Society*, Ser. B, 42, 213–220.

Spiegelhalter, D. J., and Smith, A. F. M. (1982), "Bayes Factor for Linear and Log-Linear Models With Vague Prior Information," *Journal of the Royal Statistical Society*, Ser. B, 44, 377–387.

Wood, H., Steinour, H. H., and Starke, H. R. (1932), "Effect of Composition of Portland Cement on Heat Evolved During Hardening," *Industrial and Engineering Chemistry*, 24, 1207–1214.