A TUTORIAL IN BAYESIAN REGRESSION AND VARIABLE SELECTION

DAVID ROSSELL

1. Introduction

The purpose of this tutorial is to give a gentle introduction to the basic ideas in Bayesian regression models, assuming almost no background knowledge. There are many references for further reading, for instance Gelman et al. (2013) provide one of the most comprehensive descriptions of the Bayesian framework and its numerous applications. For a classical review on Bayesian model choice, which includes variable selection, see Kass and Wasserman (1995).

The main idea behind the Bayesian framework is to complement our usual probability model that describes the behaviour of a random entity of interest $\mathbf{Y}$ given some parameter $\boldsymbol{\theta} \in \Theta$ by setting a probability distribution on $\boldsymbol{\theta}$. Specifically, let $\mathbf{Y} \in \mathcal{Y}$ where $\mathcal{Y}$ is the sample space and suppose we pose a probability model for $\mathbf{Y}$ with corresponding density $P(\mathbf{Y} \mid \boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \Theta$ and Θ is the parameter space. In the Bayesian framework one poses a probability distribution with density $P(\boldsymbol{\theta})$ called the *prior distribution*, which encapsulates what is known about $\boldsymbol{\theta}$ before observing a particular realization $\mathbf{Y} = \mathbf{y}$. Depending on the application, this knowledge can be very vague, very precise or somewhere in between. In practice it is often unclear how one should set $P(\boldsymbol{\theta})$, and this is indeed an important consideration, but for now we shall assume that we are happy with some given $P(\boldsymbol{\theta})$.

Bayesian inference provides a systematic way (in the sense that it follows probability theory) to update our knowledge about $\boldsymbol{\theta}$ after observing $\mathbf{Y} = \mathbf{y}$. Given that we have a joint probability model on $(\mathbf{Y}, \boldsymbol{\theta})$, it seems natural to consider the distribution $P(\boldsymbol{\theta} \mid \mathbf{y})$ of $\boldsymbol{\theta}$ conditional on the fact that $\mathbf{Y} = \mathbf{y}$. $P(\boldsymbol{\theta} \mid \mathbf{y})$ is called the *posterior distribution*, and expresses our updated knowledge about $\boldsymbol{\theta}$ after observing the data. A direct application of Bayes theorem gives that

$$P(\boldsymbol{\theta} \mid \mathbf{y}) = \frac{P(\mathbf{y} \mid \boldsymbol{\theta})P(\boldsymbol{\theta})}{P(\mathbf{y})}, \tag{1}$$

where $P(\mathbf{y}) = \int P(\mathbf{y}, \boldsymbol{\theta}) d\boldsymbol{\theta} = \int P(\mathbf{y} \mid \boldsymbol{\theta}) P(\boldsymbol{\theta}) d\boldsymbol{\theta}$ is the marginal density of $\mathbf{Y} = \mathbf{y}$. Depending on the probability model we are considering, Expression (1) may either have a closed-form expression that facilitates

interpretation and calculations or may require some type of approximation. For convenience here we shall focus on cases where a closed-form expression exists.

Our discussion up to this point was fairly general to emphasize that the framework can be applied to essentially any phenomenon that can be described with a probability model. For simplicity, from now on we consider the particular case where $\mathbf{y} = (y_1, \ldots, y_n) \in \mathbb{R}^n$ is a vector of n independent observations given $\boldsymbol{\theta}$, so that their joint density can be written as $P(\mathbf{y} \mid \boldsymbol{\theta}) = \prod_{i=1}^n P(\mathbf{y}_i \mid \boldsymbol{\theta})$, and that $\boldsymbol{\theta}$ has finite dimension as is typically the case in parametric regression models. However, we point out that more generally either $\mathbf{Y}$ or $\boldsymbol{\theta}$ could be infinite-dimensional (*e.g.* to consider non-parametric regression).

Section 1 describes the important particular case where $P(y_i \mid \boldsymbol{\theta})$ is a linear regression model with Normal errors and a *fixed set of predictors* that one wishes to include in the model. Section 3 discusses the case where one wishes to consider various linear regression models, where each model differs in terms of which predictors are included.

2. Bayesian inference for one regression model

2.1. **Conjugate prior and associated posterior.** Let $\mathbf{y} = (y_1, \ldots, y_n)' \in \mathbb{R}^n$ be the n observed values for the response $\mathbf{Y}$, X a given $n \times p$ matrix with the observed predictor values and $\mathbf{x}_i \in \mathbb{R}^p$ the i^{th} row in X. The linear regression model assumes that $Y_i = \mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i$ with $\epsilon_i \sim N(0, \sigma^2)$ independent across $i = 1, \ldots, n$, or equivalently that $\mathbf{Y}$ follows the multivariate Normal distribution

$$\mathbf{Y} \sim N(X\boldsymbol{\beta}, \sigma^2 I),$$

with mean vector $X\boldsymbol{\beta}$ and diagonal covariance $\sigma^2 I$, where I is the $n \times n$ identity matrix. To match the notation in the Introduction here $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$. That is, the likelihood for the observed data is

$$P(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - X\boldsymbol{\beta})'(\mathbf{y} - X\boldsymbol{\beta})\right\} \tag{2}$$

and which we denote by $N(\mathbf{y}; X\boldsymbol{\beta}, \sigma^2 I)$.

To complete the Bayesian probability model we need to set a prior distribution on the unknown parameters $(\boldsymbol{\beta}, \sigma^2)$. While there are many possible choices, here we focus on an analytically convenient choice called the *conjugate prior*. In a nutshell, a conjugate prior is such that the posterior $P(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{y})$ belongs to the same family of distributions as the prior $P(\boldsymbol{\beta}, \sigma^2)$. Conjugate priors do not always exists, but they do for the linear model. Specifically, consider the prior

$$\begin{aligned} \boldsymbol{\beta} \mid \sigma^2 &\sim N(\mathbf{0}, \sigma^2 S_0^{-1}) \\ \sigma^{-2} &\sim \text{Gamma}(v_0/2, b_0/2), \end{aligned} \tag{3}$$

where S_0 is any known $p \times p$ positive definite matrix and $v_0, b_0 \in \mathbb{R}^+$ are known constants. Equivalently, we may say that a priori σ^2 follows an inverse Gamma distribution. The fact that the prior (3) for $\boldsymbol{\beta}$ has zero mean is generally accepted as a reasonable default choice. In the absence of any subject-of-matter information, it is often reasonable to expect that $\boldsymbol{\beta}$ is centred around $\mathbf{0}$. Setting S_0, v_0, b_0 can be more controversial, although as we shall see below there are some default values. Before we discuss further, let us take a look at the posterior distribution.

Result 1. Posterior distribution under the conjugate prior.
Consider the linear model with likelihood (2) and prior (3). Then the posterior distribution is

$$\boldsymbol{\beta} \mid \sigma^2, \mathbf{y} \sim N\left(\tilde{\boldsymbol{\beta}}, \sigma^2 (X'X + S_0)^{-1}\right)$$
$$\sigma^{-2} \mid \mathbf{y} \sim \mathit{Gamma}\left(\frac{n + v_0}{2}, \frac{b_0 + \widetilde{SSR}}{2}\right),$$

where $\tilde{\boldsymbol{\beta}} = (X'X + S_0)^{-1} X'\mathbf{y}$ *and* $\widetilde{SSR} = \mathbf{y}'\mathbf{y} - \mathbf{y}'X\tilde{\boldsymbol{\beta}}$. *In particular, the posterior means are*

$$E(\boldsymbol{\beta} \mid \mathbf{y}) = \tilde{\boldsymbol{\beta}} = (X'X + S_0)^{-1} X'\mathbf{y}$$
$$E(\sigma^2 \mid \mathbf{y}) = \frac{b_0 + \widetilde{SSR}}{n + v_0 - 1}$$

Proof. We just provide a sketch of the proof. The joint density of $(\mathbf{y}, \boldsymbol{\beta}, \sigma^2)$ is given by

$$P(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2) P(\boldsymbol{\beta} \mid \sigma^2) P(\sigma^2) =$$
$$\left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - X\boldsymbol{\beta})'(\mathbf{y} - X\boldsymbol{\beta})\right\} \times$$
$$\frac{1}{(2\pi\sigma^2)^{p/2}|S_0|^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}\boldsymbol{\beta}' S_0 \boldsymbol{\beta}\right\} \times \frac{(b_0/2)^{v_0/2}}{\Gamma(v_0/2)} \left(\frac{1}{\sigma^2}\right)^{\frac{v_0}{2}+1} e^{-\frac{b_0}{2\sigma^2}}$$

Rearranging the expression and dropping terms that do not depend on $\boldsymbol{\beta}$ or σ^2 gives

$$\left(\frac{1}{\sigma^2}\right)^{\frac{n+v_0}{2}} e^{-\frac{b_0 + \mathbf{y}'\mathbf{y}}{2\sigma^2}} \frac{1}{(2\pi\sigma^2)^{p/2}} \exp\left\{-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}'(X'X + S_0)\boldsymbol{\beta} - 2\mathbf{y}'X\boldsymbol{\beta}\right)\right\}$$

Now, denoting $V = X'X + S_0$ and $\mathbf{m} = V^{-1}X\mathbf{y}'$ the expression in the last exponent is equal to $\boldsymbol{\beta}'V\boldsymbol{\beta} - 2\mathbf{y}'XV^{-1}V\boldsymbol{\beta} = \boldsymbol{\beta}'V\boldsymbol{\beta} - 2\mathbf{m}'V\boldsymbol{\beta} +$

$\mathbf{m}'V\mathbf{m} - \mathbf{m}'V\mathbf{m}$, giving

(4)
$$\left(\frac{1}{\sigma^2}\right)^{\frac{n+v_0}{2}} e^{-\frac{b_0+\mathbf{y}'\mathbf{y}-\mathbf{m}'V\mathbf{m}}{2\sigma^2}} \frac{1}{(2\pi\sigma^2)^{p/2}} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\mathbf{m})'V(\boldsymbol{\beta}-\mathbf{m})\right\}.$$

To complete the proof, simply note that $P(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{y}) = P(\boldsymbol{\beta}, \sigma^2, \mathbf{y})/P(\mathbf{y})$, *i.e.* the posterior of $(\boldsymbol{\beta}, \sigma^2)$ is given by (4) up to a proportionality constant that does not depend on $(\boldsymbol{\beta}, \sigma^2)$. The first two terms in (4) do not incorporate $\boldsymbol{\beta}$ and the latter two define a multivariate normal pdf with mean $\mathbf{m}$ and covariance $\sigma^2 V$ (*i.e.* they integrate to 1 with respect to $\boldsymbol{\beta}$). Therefore, the posterior $P(\boldsymbol{\beta} \mid \sigma^2, \mathbf{y}) = N(\boldsymbol{\beta}; \mathbf{m}, \sigma^2 V)$. This also implies that the marginal posterior $P(\sigma^2 \mid \mathbf{y})$ is given by the first two terms, which are equal to an inverse gamma density with parameters $(n + v_0)/2$ and $(b_0 + \mathbf{y}'\mathbf{y} - \mathbf{m}'V\mathbf{m})/(2\sigma^2)$ (up to a normalization constant that does not depend on σ^2).

The posterior means follow from the fact that $E(\boldsymbol{\beta} \mid \mathbf{y}) = E(E(\boldsymbol{\beta} \mid \sigma^2, \mathbf{y})) = \tilde{\boldsymbol{\beta}} = (X'X + S_0)^{-1}X'\mathbf{y}$. For σ^2 the expression of the mean for an inverse gamma gives that $E(\sigma^2 \mid \mathbf{y}) = \frac{b_0 + \widetilde{SSR}}{n+v_0-1}$. □

The probability distribution given by Result 1 characterizes all our knowledge about $(\boldsymbol{\beta}, \sigma^2)$. For instance, to obtain point estimates we may report the posterior expectation $E(\boldsymbol{\beta} \mid \mathbf{y})$. Intuitively, usually $X'X$ grows linearly with n (its (j, l) entry is $\sum_{i=1}^n x_{ij}x_{il}$), hence for any fixed S_0 as $n \to \infty$ the matrix $X'X + S_0$ is dominated by $X'X$ and $\tilde{\boldsymbol{\beta}}$ becomes arbitrarily close to the MLE $\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\mathbf{y}$. In fact, when the diagonal entries in S_0 converge to 0 it is not hard to show that $\tilde{\boldsymbol{\beta}}$ converges to $\hat{\boldsymbol{\beta}}$, for any fixed n. Regarding σ^2, for fixed S_0, v_0, b_0 as $n \to \infty$ we obtain $\widetilde{SSR} \xrightarrow{P} \mathbf{y}'\mathbf{y} - \mathbf{y}'X\hat{\boldsymbol{\beta}}$, the sum of squared residuals (SSR) under $\hat{\boldsymbol{\beta}}$ and hence $E(\sigma^2 \mid \mathbf{y}) \xrightarrow{P} \hat{\sigma}^2$, where $\hat{\sigma}^2$ denotes the MLE. Hence b_0 can be loosely interpreted as a prior guess for the SSR and v_0 the number of prior observations the guess is based upon.

We now discuss how to set default values for v_0, b_0, S_0. Based on these observations, small v_0, b_0 and any S_0 with small diagonal entries can be regarded as being fairly uninformative, in the sense that point estimates are close to the MLE. Reasonable default values are $v_0 = b_0 = 1$, so that the prior on σ^2 carries as much information as a single observation, although some authors suggest even smaller v_0, b_0. Regarding S_0, a common choice is the so-called Zellner's g-prior $S_0^{-1} = g(X'X)^{-1}$ (Zellner, 1986), where $g > 0$ can be any constant but often set to $g = n$ by default. $g = n$ gives the so-called Unit Information Prior (Schwarz, 1978), which can be interpreted as containing as much information as a single observation (this is because $\sigma^2(X'X)$ is the Fisher information for n observations, or heuristically because $(X'X)^{-1}$ is of order n^{-1}). As we shall see later on, the g-prior has asymptotic connections

with the BIC. An alternative to Zellner's g-prior is the independence prior $S_0^{-1} = gI$, in which case $\tilde{\boldsymbol{\beta}}$ is the Ridge regression estimate. Here setting $S_0^{-1} = I$ (*i.e.* $g = 1$) is a fairly uninformative default value.

At this point we should also mention that some authors take this "uninformativeness" argument to the limit by proposing improper priors, *i.e.* positive measures on $(\boldsymbol{\beta}, \sigma^2)$ that integrate to ∞. In our prior (3) this intuitively corresponds to $S_0^{-1} = gI$ where $g \to \infty$, $v_0 \to 0$ and $b_0 \to 0$. Such priors need special care to ensure that the posterior distribution is indeed a probability distribution and can lead to very undesirable consequences for variable selection, therefore they fall out of the scope of this introductory material.

2.2. **Credibility intervals.** As usual, when performing statistical inference we are not only interested in point estimates but also in portraying their inherent uncertainty. The posterior distribution 1 provides a natural way to do this using the so-called *credibility intervals* or *credibility region*. A $(1-\alpha)$ credibility interval for β_j is any set $C = (a, b)$ such that $P(\beta_j \in C \mid \mathbf{y}) = 1 - \alpha$, usually taken to be centred at $\tilde{\beta}_j$. A credibility region is the direct multivariate extensions, *i.e.* $C \in \mathbb{R}^p$ such that $P(\boldsymbol{\beta} \in C \mid \mathbf{y}) = 1 - \alpha$.

Result 2. Credibility intervals for β_j

(1) *If σ^2 is known $\tilde{\beta}_j \pm 1.96\sigma s_j$ gives a 95% credibility interval, where s_j is the (j, j) element in $(X'X + S_0)^{-1}$.*

(2) *When σ^2 is not known the marginal distribution*

$$\beta_j \mid \mathbf{y} \sim T_{n+v_0}\left(\beta_j; \tilde{\beta}_j, s_j \frac{b_0 + \widetilde{SSR}}{n + v_0}\right),$$

where $T_\nu(m, s)$ denotes the t distribution with ν degrees of freedom, location m and scale s.

Proof. We only provide a sketch of the proof. Part (1) stems directly from the posterior distribution $\beta_j \mid \sigma^2, \mathbf{y} \sim N(\tilde{\beta}_j, \sigma^2 s_j)$. Part (2) follows from integrating $P(\beta_j \mid \mathbf{y}) = \int N(\beta_j; \tilde{\beta}_j, \sigma^2 s_j)\mathrm{IG}(\sigma^2; (v_0 + n)/2, (b_0 + \mathbf{y}'\mathbf{y} - \tilde{\boldsymbol{\beta}}' V \tilde{\boldsymbol{\beta}})/2) d\sigma^2$, where $\mathrm{IG}(\sigma^2; a, b)$ denotes an inverse gamma pdf with parameters (a, b) evaluated at σ^2. After doing the integral and rearranging terms we obtain the pdf of a t distribution. □

Again we note that as $n \to \infty$ the t-distribution in the unknown σ^2 case converges to the Normal for the known σ^2 case. The expressions bears strong resemblance with the 95% *confidence interval* for the MLE $\hat{\beta}_j$, but there is a very important fundamental difference. A $(1-\alpha)$ credibility interval can be interpreted by saying that there is $1 - \alpha$ probability to contain the parameter value, but this is not the case for

frequentist confidence intervals. Instead, a $(1-\alpha)$ confidence interval is defined to be an interval that, if we were to repeat our experiment many times over, it would contain $\hat{\beta}_j$ a proportion $1-\alpha$ of the times. That is, the confidence interval is based on the probability that a data-based estimate $\hat{\beta}_j$ falls in a certain region, whereas the Bayesian credibility interval is the probability that β_j falls in a region. The difference may be subtle, especially given that as we just saw the actual intervals are actually not very different, but it lies at the very heart of the essential difference between the two paradigms. In fact, in more complex situations credibility and confidence intervals may present strong differences, *e.g.* when $\boldsymbol{\beta}$ is highly or even infinitely dimensional, but these are beyond the scope of this introduction.

2.3. **Connections to penalized likelihood.** In Section 2.1 we proposed using the posterior mean of the parameters $E(\boldsymbol{\beta} \mid \mathbf{y})$ and $E(\sigma^2 \mid \mathbf{y})$ as point estimates. This choice can be formally justified as that minimizing the posterior expected squared error, but obviously there are other reasonable ways of summarizing the posterior distribution to get parameter estimates. An alternative is to report the *posterior mode*

$$(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2) = \text{argmax}_{\boldsymbol{\beta},\sigma^2} P(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{y}) = \text{argmax}_{\boldsymbol{\beta},\sigma^2} P(\boldsymbol{\beta}, \sigma^2, \mathbf{y}), \tag{5}$$

where the last equality follows from the fact that $P(\mathbf{y})$ does not depend on $(\boldsymbol{\beta}, \sigma^2)$. For the conjugate prior (3) gives rise to the posterior on $\boldsymbol{\beta}$ given in Result 1, which is symmetric and unimodal, and hence the posterior mean is equal to the posterior mode. For other prior choices however both quantities are in general different, and in fact gives some interesting connections with penalized likelihood. For a general prior $P(\boldsymbol{\beta}, \sigma^2)$ the posterior mode is obtained by maximizing

$$\begin{aligned} &\log\left(N(\mathbf{y}; \boldsymbol{\beta}, \sigma^2 I)\right) + \log\left(P(\boldsymbol{\beta}, \sigma^2)\right) = \\ &c - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 + \log\left(P(\boldsymbol{\beta}, \sigma^2)\right), \end{aligned} \tag{6}$$

where c is a constant not depending on $(\boldsymbol{\beta}, \sigma^2)$. Expression (6) fits neatly in the penalized likelihood approach, where the penalty is given by the negative log-prior density $-P(\boldsymbol{\beta}, \sigma^2)$.

In the particular case of the conjugate prior (3) with $S_0 = \lambda I$, where $\lambda \in \mathbb{R}^+$ is a given constant, $P(\boldsymbol{\beta}, \sigma^2) = -\log(\sigma^2)^{p/2} + \frac{1}{2\sigma^2}\lambda\sum_{j=1}^{p}\beta_j^2$. That is, the penalty on $\boldsymbol{\beta}$ is identical to that in Ridge regression and the posterior mode $\tilde{\boldsymbol{\beta}} = (X'X + \lambda I)^{-1}X'\mathbf{y}$ coincides with the Ridge regression estimate.

We now consider a different (non-conjugate) prior that gives a connection with the LASSO. Suppose that we assume that a priori β_j are independent realizations from the Double exponential (DE) distribution with parameter λ/σ^2, which we denote $\beta_j \sim \text{DE}(\lambda/\sigma^2)$. The

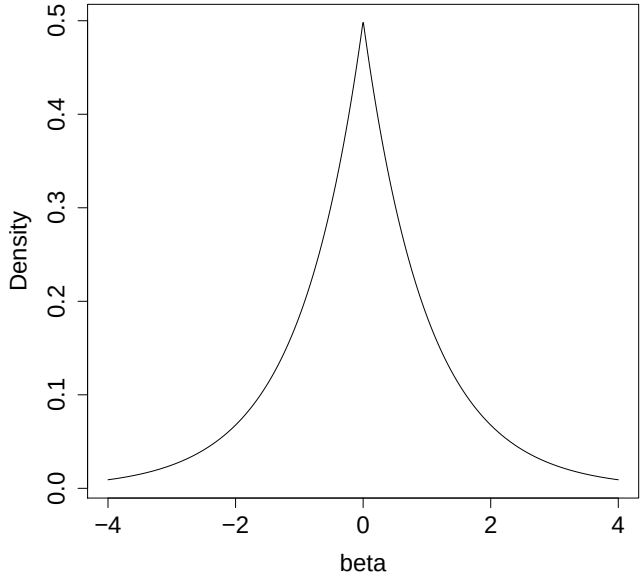

FIGURE 1. Probability density function of the Double exponential distribution

probability density function for the double exponential is

$$P(\beta_j \mid \sigma^2) = \frac{\lambda}{2\sigma^2}\exp\left\{-\frac{\lambda}{\sigma^2}|\beta_j|\right\}, \tag{7}$$

i.e. the usual exponential distribution extended to the negative numbers. Figure 1 shows a plot of this prior density. In this case $P(\boldsymbol{\beta} \mid \sigma^2) = -\frac{p}{2}\log(\sigma^2) - \frac{1}{\sigma^2}\lambda\sum_{j=1}^{p}|\beta_j|$, so that minimizing (6) with respect to $\boldsymbol{\beta}$ becomes equivalent to the LASSO. For this reason the Double exponential prior (7) is often referred to as the *Bayesian LASSO*. Relative to the LASSO, an interesting property of its Bayesian counterpart is that credibility intervals can be simply obtained by reporting the quantiles of the posterior $P(\beta_j \mid \mathbf{y})$. Unlike in the conjugate prior case, the Bayesian LASSO posterior does not have a closed-form expression, hence in practice one needs to resort to some type of numerical approximation. A common approach is to use Markov Chain Monte Carlo, which we shall not discuss here but is described in Hans (2009).

There are many other possible prior choices which also lead to interesting likelihood penalties and lead to posterior modes where some of the $\tilde{\beta}_j = 0$ exactly. Although these contributions are doubtlessly valuable in that one may end up dropping variables even when considering a single model, here we shall focus on the standard Bayesian variable selection framework where one considers several models. One computes the posterior probability for each of the 2^p possible models

computes posterior model probabilities and uses these probabilities to decide which variables should be included. We discuss this in Section 3.

3. Bayesian variable selection

The traditional approach to Bayesian model selection (of which variable selection is a particular case) is to consider a set of K models, which we denote as $M_1, \ldots, M_K$, and evaluate their posterior probabilities $P(M_k \mid \mathbf{y})$ conditional on the observed data $\mathbf{y}$. In variable selection where X is an $n \times p$ matrix with predictors we have $K = 2^p$ and each M_k is defined by the subset of predictors X_k included into the model, where X_k is $n \times p_k$ and p_k is the number of variables in M_k. We denote by $\boldsymbol{\beta}_k$ the regression coefficients for X_k and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ the vector containing the coefficients for all variables. Following Bayes theorem, posterior model probabilities are given by

$$P(M_k \mid \mathbf{y}) = \frac{P(\mathbf{y} \mid M_k)P(M_k)}{P(\mathbf{y})} \propto P(\mathbf{y} \mid M_k)P(M_k), \tag{8}$$

where $P(\mathbf{y} \mid M_k)$ is called the *marginal likelihood* or *integrated likelihood* under M_k (for reasons to become apparent shortly), $P(M_k)$ is the *model prior probability* and the marginal density $P(\mathbf{y}) = \sum_{k=1}^{K} P(\mathbf{y} \mid M_k)P(M_k)$ does not depend on M_k. Now, as usual $P(\mathbf{y} \mid M_k)$ can be obtained by integrating the joint distribution $P(\mathbf{y}, \boldsymbol{\beta}, \sigma^2 \mid M_k)$, *i.e.*

$$\begin{aligned} P(\mathbf{y} \mid M_k) &= \int\int P(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2, M_k)P(\boldsymbol{\beta}, \sigma^2 \mid M_k)d\boldsymbol{\beta}d\sigma^2 = \\ &\int\int N(\mathbf{y}; X_k\boldsymbol{\beta}_k, \sigma^2 I)P(\boldsymbol{\beta}_k, \sigma^2 \mid M_k)d\boldsymbol{\beta}d\sigma^2. \end{aligned} \tag{9}$$

The last line in (9) is obtained from the linear model assumptions, the fact that under M_k then $\mathbf{y}$ depends on X, $\boldsymbol{\beta}$ only through X_k, $\boldsymbol{\beta}_k$, and that under M_k all coefficients out of $\boldsymbol{\beta}_k$ are 0 with probability 1 and hence can be dropped from $P(\boldsymbol{\beta}, \sigma^2 \mid M_k)$.

Example 1. *Consider predicting the rental price of a flat (Y) based on its square feet ($x_1 \in \mathbb{R}^+$) and whether its located or not in a multistory building ($x_2 \in \{0, 1\}$) using the linear relationship $E(Y_i \mid \mathbf{x}_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$. We consider the four models below.*

$$\begin{aligned} M_1 &: \beta_1 = 0, \beta_2 = 0 \\ M_2 &: \beta_1 = 0, \beta_2 \neq 0 \\ M_3 &: \beta_1 \neq 0, \beta_2 = 0 \\ M_4 &: \beta_1 \neq 0, \beta_2 \neq 0 \end{aligned}$$

To highlight the flexibility of the framework, nothing is to stop us from considering a transformation of x_1, say by defining $x_3 = log(x_1)$

which was suggested by some exploratory data analysis. One option would be to then consider the 8 models

$$
\begin{aligned}
M_1 &: \beta_1 = 0, \beta_2 = 0, \beta_3 = 0 \\
M_2 &: \beta_1 = 0, \beta_2 \neq 0, \beta_3 = 0 \\
M_3 &: \beta_1 \neq 0, \beta_2 = 0, \beta_3 = 0 \\
M_4 &: \beta_1 \neq 0, \beta_2 \neq 0, \beta_3 = 0 \\
M_5 &: \beta_1 = 0, \beta_2 = 0, \beta_3 \neq 0 \\
M_6 &: \beta_1 = 0, \beta_2 \neq 0, \beta_3 \neq 0 \\
M_7 &: \beta_1 \neq 0, \beta_2 = 0, \beta_3 \neq 0 \\
M_8 &: \beta_1 \neq 0, \beta_2 \neq 0, \beta_3 \neq 0
\end{aligned}
$$

Another possibility would be to consider that we want to include either x_1 or x_3 but not both, by removing M_7, M_8 from the models under consideration. Along similar lines we could also consider a linear model for transformations of Y, which has formal connections with Box-Cox transforms. In fact any other model adjustments can be incorporated into the framework, e.g. non-normal or dependent residuals, but in this case we would need to adjust (9) as the likelihood would no longer be that of a linear model.

Expressions (8)-(9) tell us that, in principle, to obtain posterior model probabilities we just need to set $P(M_k)$ and $P(\boldsymbol{\beta}_k, \sigma^2 \mid M_k)$ for $k = 1, \ldots, K$ and then find some way to evaluate the integral in (9). Ideally, the priors should reflect subject-of-matter knowledge related to the problem at hand, but this is often either not available or impractical (specially when p is large). Section 3.1 discusses default strategies to set $P(M_k)$, whereas Section 3.2 focuses on conjugate $P(\boldsymbol{\beta}_k, \sigma^2 \mid M_k)$ that give closed-form expressions for (9). Section 3.3 discusses how after obtaining $P(M_k \mid \mathbf{y})$ we can use them to decide which variables to include in our final model. Note that when p is large it is unfeasible to obtain $P(M_k \mid \mathbf{y})$ exhaustively for all $k = 1, \ldots, 2^p$. Section 3.4 introduces a simple Markov Chain Monte Carlo algorithm that can be used to explore an interesting subset of models, namely those with high $P(M_k \mid \mathbf{y})$.

3.1. **Prior probabilities on model space.** Default choices for $P(M_k)$ are usually guided by not being too informative and, in complex situations where p is large, a desire to encourage parsimonious answers. When p is small it is often reasonable to use the so-called *uniform prior*, which sets equal prior probabilities $P(M_k) = 2^{-p}$. Another possible choice is to use the *binomial prior*, which poses that *a priori* all variables are independent and have an inclusion probability π. That is, let $\gamma_j = I(\beta_j \neq 0)$ be a indicator for variable $j = 1, \ldots, p$ being

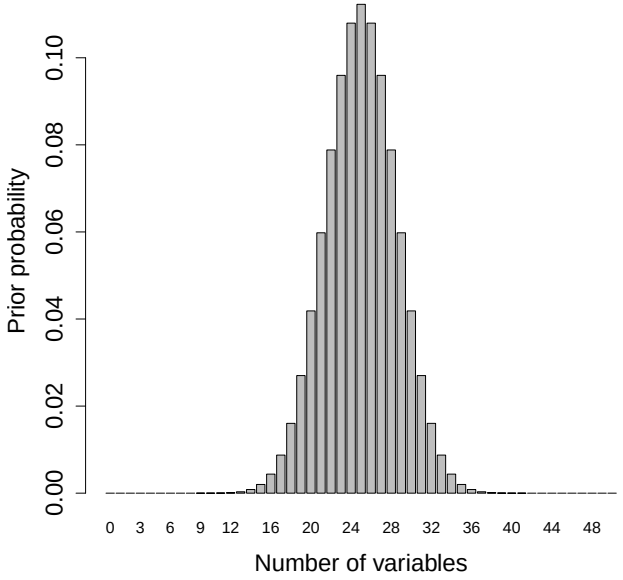

FIGURE 2. Prior probabilities of model size under the uniform prior, $p = 50$

included in the model, then $\sum_{j=1}^{p} \gamma_j \sim \text{Bin}(p, \pi)$ where π is some user-specified value and probability is split equally across all models with the same number of variables.

Definition 1. *The binomial prior assigns prior model probabilities*

$$P(M_k) = \frac{\binom{p}{p_k}\pi^{p_k}(1-\pi)^{p-p_k}}{\binom{p}{p_k}} = \pi^{p_k}(1-\pi)^{p-p_k} \tag{10}$$

Note that setting $\pi = 0.5$, a common default value, gives $P(M_k) = 1/2^p$ and hence the binomial prior becomes equivalent to the uniform prior.

Example 2. *Consider the flat rental prices in Example 1. A uniform prior on the model space assigns $P(M_1) = P(M_2) = P(M_3) = P(M_4) = 0.25$, which may be reasonable in the absence of any prior knowledge. Of course, in this example we might want to reflect our knowledge that the square feet do have an effect on the rental price, for instance by setting $P(M_1) = P(M_2) = 0.1$, $P(M_3) = P(M_4) = 0.9$ or even $P(M_1) = P(M_2) = 0$, $P(M_3) = P(M_4) = 0.5$ (which is equivalent to dropping M_1, M_2 from the list of models).*

Although the uniform and binomial priors can be useful, when p is large they have the undesirable property that they focus most probability on mid-size models and almost no probability to either the full

model or the null model with no variables. According to the uniform prior the probability of having q variables in the model is $P(\sum_{j=1}^{p}\gamma_j = q) = \binom{p}{q}/2^p$, which is largest for $q \in [\lfloor p/2 \rfloor, \lceil p/2 \rceil]$. For the binomial prior as p grows $P\left(\sum_{j=1}^{p}\gamma_j \in \pi p \pm 1.96\sqrt{p\pi(1-\pi)}\right) \approx 0.95$, so that most prior mass falls on models with roughly πp variables. As an example consider $p = 50$ and the uniform prior (equivalently, $\text{Bin}(50, 0.5)$), which give the probabilities shown in Figure 2. The implication of such a prior is that it strongly discourages models with small size, which is the opposite of what we would like in most applications. The *beta-binomial prior* provides an interesting extension of the binomial prior. The idea set a uniform distribution on the model size $\sum_{j=1}^{p}\gamma_j$ (rather than a binomial) and then split the probability equally amongst all models of that given size.

Definition 2. *The Beta-binomial(1,1) prior assigns prior model probabilities*

$$P(M_k) = P\left(\sum_{j=1}^{p}\gamma_i = p_k\right) P\left(M_k \mid \sum_{j=1}^{p}\gamma_i = p_k\right) = \frac{1}{p}\binom{p}{p_k}^{-1} \tag{11}$$

By definition the Beta-binomial(1,1) prior assigns equal prior probability to all model sizes, avoiding unduly favouring mid-size models as was the case for the uniform or Binomial priors in Figure 2. The beta-binomial name arises from the fact that if $\sum_{j=1}^{p}\gamma_i \mid \pi \sim \text{Bin}(p, \pi)$ and we consider π to be an unknown quantity for which we place a prior distribution $\pi \sim U(0, 1)$, then the prior distribution of the model size is $\sum_{j=1}^{p}\gamma_i \sim \text{BetaBin}(1, 1)$. More generally, one could set $\pi \sim \text{Beta}(a, b)$ (which reduces to $U(0, 1)$ for $a = b = 1$) to obtain a Beta-binomial(a, b) prior. This distribution is quite flexible, *e.g.* setting $a < b$ encourages models of small size, but by default we shall stick to the Beta-Binomial(1,1).

3.2. **Posterior probabilities under conjugate prior.** There are many possible strategies to choose reasonable default priors $P(\boldsymbol{\beta}_k, \sigma^2 \mid M_k)$. Here we focus on the conjugate prior (3) with $S_0 = g^{-1}(X_k'X_k)$, as this leads to closed-form expressions that will help us gain some intuition as to how integrated likelihoods (and therefore posterior probabilities) behave. Without loss of generality, let M_1 be the null model that includes no variables and consider $\text{BF}_{k1} = \frac{P(\mathbf{y}|M_k)}{P(\mathbf{y}|M_1)}$, which is called the *Bayes factor* between M_k and M_1. Plugging the conjugate prior density into (9) and carrying out some algebra delivers

$$\text{BF}_{k1} = (1+ng)^{\frac{-p}{2}}\left(1 + ng\left(1 + \frac{\tilde{\boldsymbol{\beta}}_k' X_k' X_k \tilde{\boldsymbol{\beta}}_k}{b_0 + \widetilde{SSR}}\right)^{-1}\right)^{-\frac{v_0+n}{2}}, \tag{12}$$

where $S\tilde{S}R = \mathbf{y}'\mathbf{y} - \mathbf{y}'X\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}_k$ are as in Result 1. This is interesting, as the term $\tilde{\boldsymbol{\beta}}_k' X_k' X_k \tilde{\boldsymbol{\beta}}_k$ can be roughly interpreted as the sum of squares explained by M_k and $b_0 + \widetilde{SSR}$ as the residual sum of squares, so that their ratio is related to the usual F-test statistic to compare M_k and M_1 (the null model). In the particular case of the Unit Information Prior Schwarz (1978) showed the following equivalence with the Bayesian Information Criterion.

Result 3. *Consider the Unit Information Prior as a particular case of the conjugate prior (3) with* $S_0 = n^{-1}(X'X)$. *Then as* $n \to \infty$

$$-2log(P(\mathbf{y} \mid M_k) \xrightarrow{P} BIC_k + c$$

for some constant c *that does not depend on* n, *where* $BIC_k = log(P(\mathbf{y} \mid \hat{\boldsymbol{\beta}}_k, \hat{\sigma}^2)) + plog(n)$ *is the BIC and* $(\hat{\boldsymbol{\beta}}_k, \hat{\sigma}^2)$ *is the MLE under* M_k.

The implication is that choosing the model that minimizes the BIC is asymptotically equivalent to maximizing $P(\mathbf{y} \mid M_k)$, *i.e.* choosing the model with largest posterior probability under uniform $P(M_k) = 1/K$.

From (8) posterior model probabilities can be computed from the Bayes factors as

(13)

$$P(M_k \mid \mathbf{y}) = \frac{P(\mathbf{y} \mid M_k)P(M_k)}{\sum_{k'=1}^{K} P(\mathbf{y} \mid M_{k'})P(M_{k'})} = \frac{\mathrm{BF}_{k1}\frac{P(M_k)}{P(M_1)}}{1 + \sum_{k'=2}^{K} \mathrm{BF}_{k'1}\frac{P(M_{k'})}{P(M_1)}}.$$

That is, posterior probabilities combine the evidence in favour of M_k provided by the Bayes factor times the ratio of prior model probabilities.

We do remark that the literature in prior choice is extensive and not devoid of controversy, and that various non-conjugate formulations for $P(\boldsymbol{\beta}_k \mid \sigma^2, M_k)$ have been shown to possess better mathematical properties than the conjugate formulation outlined here. For instance, one could choose a functional form with thick tails Jeffreys (1961); Bayarri and Garcia-Donato (2007), set $S_0 = g^{-1}(X'X)$ and treat g as a hyper-parameter so be learnt from the data Liang et al. (2008), set prior densities that induce a probabilistic separation between the considered models Johnson and Rossell (2010, 2012) or that satisfy certain desirable properties Bayarri et al. (2012).

3.3. **Using $P(M_k \mid \mathbf{y})$ to select variables.** We now discuss how one can use posterior model probabilities $P(M_k \mid \mathbf{y})$ to select variables. Intuitively, large $P(M_k \mid \mathbf{y})$ indicates that the model is supported by the data (combined with our model prior probabilities), hence one option is to choose the posterior mode, *i.e.* M_{k^*} where $k^* = \mathrm{argmax}_{k \in \{1,\ldots,K\}} P(M_k \mid \mathbf{y})$. This often works well, specially when p (and hence $K = 2^p$) is small as in Example 1, and can be formally

Model	$P(\mathbf{y} \mid M_k)$	$P(M_k)$	$P(M_k \mid \mathbf{y})$	BIC
$M_1 : \beta_1 = 0, \beta_2 = 0$	$e^{-262.0}$	1/4	0.00	363.2
$M_2 : \beta_1 \neq 0, \beta_2 = 0$	$e^{-261.2}$	1/4	0.00	360.1
$M_3 : \beta_1 = 0, \beta_2 \neq 0$	$e^{-233.5}$	1/4	0.70	303.2
$M_4 : \beta_1 \neq 0, \beta_2 \neq 0$	$e^{-234.4}$	1/4	0.30	306.3

TABLE 1. Hypothetical posterior probabilities for flat rental example under uniform $P(M_k)$

justified as the decision-theoretic optimal choice under a 0/1 loss function. One limitation is that when p is large often $P(M_k \mid \mathbf{y})$ is quite small for any given k, so that there is not an overwhelming amount of evidence in favour of any given model. A reasonable alternative is to include variables with large marginal *posterior inclusion probability* (PIP) $P(\gamma_j = 1 \mid \mathbf{y}) = \sum_{k:\gamma_j=1} P(M_k \mid \mathbf{y})$. A common option is to include variables with $P(\gamma_j \mid \mathbf{y}) > 0.5$, which is called the *median probability model* and can be shown to be reasonable when our goal is to predict $\mathbf{y}$ (Barbieri and Berger, 2004). If the goal is to explain $\mathbf{y}$, then one often sets a higher threshold such as $P(\gamma_j \mid \mathbf{y}) > 0.95$, which guarantees that the posterior expected proportion of false discoveries (variables wronly included into the model) is ≤ 0.05 (Müller et al., 2007).

Example 3. *Consider again four models for flat rental prices in Example 1. Suppose that upon observing* $\mathbf{y}$ *for* $n = 100$ *flats we obtain the posterior probabilities in Table 1. The largest posterior probability is allocated to* M_3 *and the marginal inclusion probabilities are* $P(\gamma_1 = 1 \mid \mathbf{y}) = 0 + 0.7 = 0.7$ *and* $P(\gamma_2 = 1 \mid \mathbf{y}) = 0.7 + 0.3 = 1$ *(up to rounding). Either the posterior mode rule or requiring* $P(\gamma_j \mid \mathbf{y}) > 0.95$ *would choose* M_3 *and hence include only* X_2 *into the model, whereas the median probability model would also include* X_1*. This example illustrates a common situation, namely that rules that target high predictive ability often end up choosing more complex models than rules seeking an explanatory model.*

Example 4. *Consider now a more challenging scenario with* $n = 100$ *observations and* $p = 50$ *predictors. For this example a data set was simulated by setting* $\beta_1 = \beta_{40} = 0$, $\beta_{41} = \ldots = \beta_{45} = 0.5$, $\beta_{46} = \ldots = \beta_{50} = 1$ *and* $\sigma^2 = 1$*. The predictor values were generated* $\mathbf{x}_i \sim N(\mathbf{0}, \Sigma)$ *independently for* $i = 1, \ldots, n$*, where* Σ *has diagonal elements equal to 1 and all off-diagonal elements equal to 0.5 (which is meant to emulate a situations where predictors are moderately correlated). Because* p *is moderately large, we set* $P(M_k)$ *using the Beta-Binomial(1,1) (Section 3.1). The number of models in this example is* 2^{50}*, which is too large for it to be practical to enumerate exhaustively. Therefore we resorted*

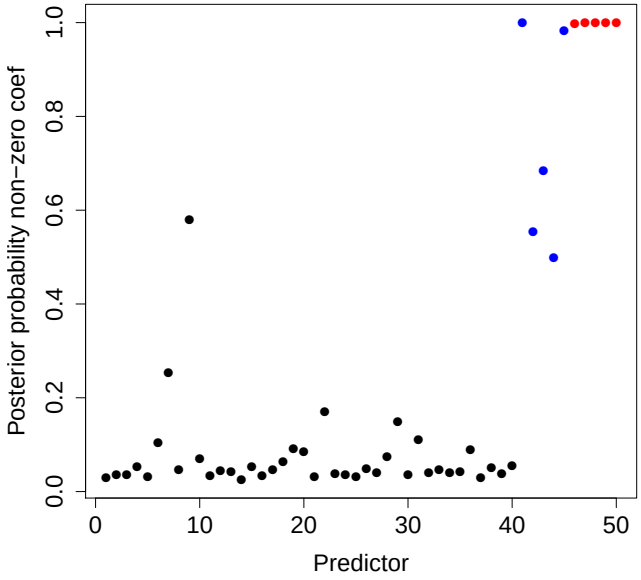

FIGURE 3. Marginal posterior inclusion probabilities $P(\gamma_j = 1 \mid \mathbf{y})$ for Example 4

to function `bms` *in R package* `BMS`*, which estimates* $\hat{P}(M_k \mid \mathbf{y})$ *by conducting a random search on the model space (Section 3.4). For now let us not worry about the details of this algorithm but simply inspect the posterior inclusion probabilities (PIP)* $P(\gamma_j \mid \mathbf{y})$*, shown in Figure 3. The PIPs range in* $(0, 0.2)$ *for most of the 40 variables with* $\beta_j = 0$*, in* $(0.5, 1)$ *for those with* $\beta_j = 0.5$ *and are essentially 1 for* $\beta_j = 1$*. Here the median probability model would include 9 out of the 10 variables where truly* $\beta_j \neq 0$*, plus a false positive that has PIP*≈ 0.6*. With the more conservative threshold* $P(\gamma_j \mid \mathbf{y}) > 0.95$ *we would not incur any false positives, at the cost of missing 3 variables with true* $\beta_j = 0.5$*. Again, we appreciate a tension between rules that focus on predictive and explanatory models.*

3.4. **Stochastic model search.** Given that when p is large we cannot enumerate all $K = 2^p$ models we need to resort to some type of model exploration. To achieve this there are certainly many possibilities that one could consider such as deterministic greedy searches or integer programming algorithms, a popular option is to use random searches based on Markov Chain Monte Carlo (MCMC) methods. The basic idea is quite simple. Let $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$ be the variable inclusion indicators, then we are interested in determining $P(\boldsymbol{\gamma} = \mathbf{g}_k \mid \mathbf{y})$ where $\mathbf{g}_k \in \{0, 1\}^p$ is a binary vector indicating which variables are included in M_k. Given that $\boldsymbol{\gamma}$ is a categorical random variable (it has 2^p possible

values), suppose that we were able to obtain T samples $\boldsymbol{\gamma}^{(1)}, \ldots, \boldsymbol{\gamma}^{(T)}$ from its posterior distribution $P(\boldsymbol{\gamma} \mid \mathbf{y})$ then we could use the Monte Carlo estimate

$$\hat{P}(M_k \mid \mathbf{y}) = \frac{1}{T} \sum_{t=1}^{T} \mathrm{I}(\boldsymbol{\gamma}^{(t)} = \mathbf{g}_k) \tag{14}$$

That is, (14) estimates the probability of M_k with the proportion of samples for which $\boldsymbol{\gamma} = \mathbf{g}_k$, which by the Law of large numbers $\hat{P}(M_k \mid \mathbf{y}) \xrightarrow{a.s.} P(M_k \mid \mathbf{y})$ as $T \to \infty$.

Of course, the approach above hinges on being able to sample from $P(\boldsymbol{\gamma} \mid \mathbf{y})$. While obtaining *independent* samples can be quite challenging, fortunately obtaining *dependent* samples is much easier (note that for (14) to be a consistent estimator we do not need the $\boldsymbol{\gamma}^{(t)}$'s to be independent). This is precisely what MCMC does for us, namely by defining a Markov Chain on $\boldsymbol{\gamma}^{(t)}$ that is guaranteed to have $P(\boldsymbol{\gamma} \mid \mathbf{y})$ as its stationary distribution. In its usual form MCMC starts with an arbitrary initial $\boldsymbol{\gamma}^{(0)}$ and then runs a Markov Chain for a large number of iterations T, and this chain is constructed in such a way that eventually it is guaranteed to converge in distribution to $P(\boldsymbol{\gamma} \mid \mathbf{y})$. Algorithm 1 gives a sketch of how one would implement this on a computer.

Algorithm 1. MCMC on the model space

Set $\boldsymbol{\gamma}^{(0)}$ *at an arbitrary value,* $t = 0$. *Repeat Steps (1)-(2) until* $t = T$.

(1) *Given the current state* $\boldsymbol{\gamma}^{(t)} = \mathbf{g}_k$, *update to* $\boldsymbol{\gamma}^{(t+1)} = \mathbf{g}_{k'}$ *with probability* $p_{kk'}$.
(2) *Set* $t = t + 1$.

Let $\mathbf{P}$ be the $K \times K$ matrix with transition probabilities $p_{kk'}$, the question is how to set $\mathbf{P}$ such its stationary distribution is $\mathbf{q} = (P(M_1 \mid \mathbf{y}), \ldots, P(M_K \mid \mathbf{y}))'$, which formally requires that $\mathbf{Pq} = \mathbf{q}$. As it turns out the choice of $\mathbf{P}$ is not unique and some choices are preferrable to others, *e.g.* converging faster to the stationary distribution or giving samples that have a lower degree of dependence. Here we shall introduce a simple but often effective choice called *Gibbs sampling*, which in our particular application can be seen as a stochastic version of stepwise methods. See Casella and George (1992) for a gentle introduction to Gibbs sampling and Madigan et al. (1995) for its specific application to Bayesian model selection that we outline below. Denote by $\boldsymbol{\gamma}_{-j} = (\gamma_1, \ldots, \gamma_{j-1}, \gamma_{j+1}, \ldots, \gamma_K)$ the vector obtained by removing γ_j from $\boldsymbol{\gamma}$, Gibbs sampling proceeds by sequentially updating each γ_j given $\boldsymbol{\gamma}_{-j}$, as outlined below. In the context of model selection, the algorithm is often referred to as MC^3 (Model Composition MCMC).

Algorithm 2. MC^3

Set $\boldsymbol{\gamma}^{(0)}$ *at an arbitrary value,* $t = 0$*. Repeat Steps (1)-(2) until* $t = T$*.*

(1) *Set* $\boldsymbol{\gamma}^{(t)} = \boldsymbol{\gamma}^{(t-1)}$*. For* $j = 1, \ldots, p$*, set* $\gamma_j^{(t)} = 1$ *with probability* $r = P(\gamma_j = 1 \mid \boldsymbol{\gamma}_{-j}^{(t)}, \mathbf{y})$ *and* $\gamma_j^{(t)} = 0$ *with probability* $1 - r$*.*

(2) *Set* $t = t + 1$

The key is that the update probability $r =$

$$\frac{P\left(\gamma_j = 1, \boldsymbol{\gamma}_{-j}^{(t)} \mid \mathbf{y}\right)}{P(\boldsymbol{\gamma}_{-j}^{(t)} \mid \mathbf{y})} = \left(1 + \frac{P(\mathbf{y} \mid \gamma_j = 0, \boldsymbol{\gamma}_{-j}^{(t)}) P(\gamma_j = 0, \boldsymbol{\gamma}_{-j}^{(t)})}{P(\mathbf{y} \mid \gamma_j = 1, \boldsymbol{\gamma}_{-j}^{(t)}) P(\gamma_j = 1, \boldsymbol{\gamma}_{-j}^{(t)})}\right)^{-1}$$

only requires comparing the two models $\gamma_j = 1$ and $\gamma_j = 0$ for the (fixed) currently included/excluded remaining variables. That is, instead of computing $P(M_k \mid \mathbf{y})$ for 2^p models it now suffices to consider pairs of models by sequentially adding/dropping variables. After running the chain for large enough iterations T we are obtaining (dependent) samples from $P(M_k \mid \mathbf{y})$, which means that M_k with high posterior probability will be sampled more often that those with lower probability. Therefore, MC^3 can be viewed as a stochastic stepwise algorithm that (upon convergence of the chain) tends to focus on promising models, in the sense of having large posterior probability.

Naturally, the immediate practical questions are how do we know when the chain has reached its stationary distribution and how long should we run the chain for. Answering these questions in a mathematically rigorous manner is extremely challenging and has been the subject of intensive research that is beyond the scope of this introduction. A pragmatic approach to assess convergence is to either run the chain multiple times and compare the estimated posterior model or variable inclusion probabilities $\hat{P}(M_k \mid \mathbf{y})$ and $\hat{P}(\gamma_j = 1 \mid \mathbf{y})$ (respectively), or run the chain for a long time and obtain some visual diagnostics. For instance one may inspect how some summary evolves as t grows, *e.g.* the model size $\sum_{j=1}^{t} \gamma_j^{(t)}$, decide at which time t_0 it has "stabilized" and then discard the iterations previous to t_0 (iterations $1, \ldots, t_0$ are called the *burn-in period*). The following example provides an illustration.

Example 5. *Consider the simulation in Example 4. R function* `bms` *implements a variation of MC*^{3} *(Algorithm 2) where, rather than sequentially adding/dropping* $j = 1, \ldots, p$*, a random* j *is uniformly sampled from* $\{1, \ldots, p\}$*. The idea is that then the order in which variables are added/dropped is no longer fixed and that this may help escale local modes, but for our purposes here let us just focus on inspecting the resulting chain.*

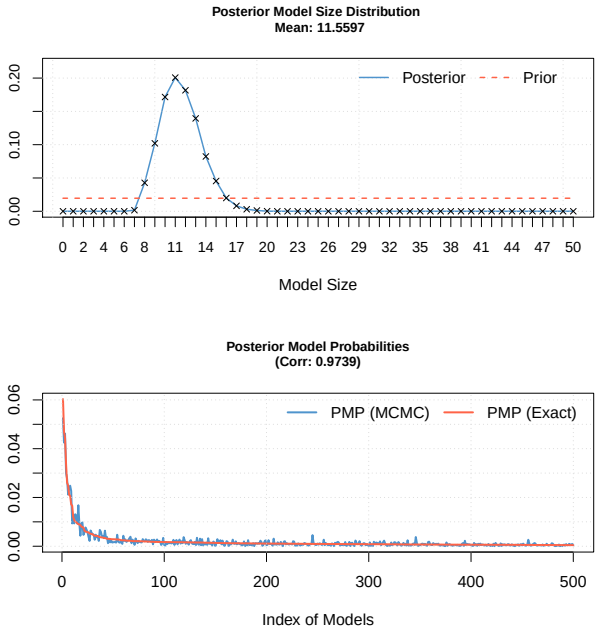

FIGURE 4. MCMC convergence checks for Example 4-5.

We run the chain for $T = 100,000$ *iterations. Let* $\mathcal{M}$ *be the set of models that were visited at some point during the* T *iterations. If the chain had converged, then the proportion of visits* $\hat{P}(M_k \mid \mathbf{y})$ *to each* M_k *should be roughly equal to* $P(M_k \mid \mathbf{y}, M_k \in \mathcal{M})$*. The bottom pannel of Figure 4 compares* $\hat{P}(M_k \mid \mathbf{y})$ *vs.* $P(M_k \mid \mathbf{y}, M_k \in \mathcal{M})$ *for the top 500 models, ordered from largest to smallest* $P(M_k \mid \mathbf{y}, M_k \in \mathcal{M})$*. In this example both quantities are fairly similar, supporting that the chain may indeed have converged to its stationary distribution. The upper panel in Figure 4 compares the prior and posterior distributions on the model size* $\sum_{j=1}^{p} \gamma_j$*. While the Beta-Binomial prior spread its mass equally across all model sizes, the posterior distribution suggests that the actual model size is probably between 8 and 16. This plot does not assess convergence per se, although observing a posterior that does not change smoothly with the model size often indicates that too few iterations were used, which leads to noisy estimates* $\hat{P}(\sum_{j=1}^{p} \gamma_j = k)$ *for* $k = 1, \ldots, p$*. As a further convergence check, we run the chain multiple times and found that the obtained* $\hat{P}(\gamma_j = 1 \mid \mathbf{y})$ *had fairly low variability.*

REFERENCES

M.M. Barbieri and J.O. Berger. Optimal predictive model selection. *The Annals of Statistics*, 32(3):870–897, 2004.

M.J. Bayarri and G. Garcia-Donato. Extending conventional priors for testing general hypotheses in linear models. *Biometrika*, 94:135–152, 2007.

M.J. Bayarri, J.O. Berger, A. Forte, and G. Garcia-Donato. Criteria for bayesian model choice with application to variable selection. *The Annals of statistics*, 40:1550–1577, 2012.

G. Casella and E.I. George. Explaining the Gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.

A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin. *Bayesian Data Analysis*. Chapman & Hall, 3nd edition, 2013.

C. Hans. Bayesian lasso regression. *Biometrika*, 96:835–845, 2009.

H. Jeffreys. *Theory of Probability*. Oxford University Press, Oxford, England, third edition, 1961.

V.E. Johnson and D. Rossell. Prior densities for default Bayesian hypothesis tests. *Journal of the Royal Statistical Society B*, 72:143–170, 2010.

V.E. Johnson and D. Rossell. Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association*, 24(498):649–660, 2012.

R.E. Kass and L. Wasserman. A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, 90:928–934, 1995.

F. Liang, R. Paulo, G. Molina, M.A. Clyde, and J.O. Berger. Mixtures of g-priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103:410–423, 2008.

D. Madigan, J. York, and D. Allard. Bayesian graphical models for discrete data. *International Statistical Review*, 63(2):215–232, 1995.

P. Müller, G. Parmigiani, and K. Rice. *FDR and Bayesian Multiple Comparisons Rules*. Oxford University Press, 2007.

G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

A. Zellner. On assessing prior distributions and bayesian regression analysis with g-prior distributions. In *Bayesian inference and decision techniques: essays in honor of Bruno de Finetti*, Amsterdam; New York, 1986. North-Holland/Elsevier.